I. R. IRAN

ISSN: 1728-144X
e-ISSN: 1735-9244

International Journal of Engineering

Journal Homepage: www.ije.ir

MERC

TRANSACTIONS B: APPLICATIONS

Volume 36, Number 08, August 2023

Materials and Energy Research Center

In the Name of God

INTERNATIONAL JOURNAL OF ENGINEERING
Transactions B: Applications

DISCLAIMER

The publication of papers in International Journal of Engineering does not imply that the editorial board, reviewers or publisher accept, approve or endorse the data and conclusions of authors.

International **J**ournal of **E**ngineering *Transactions A: Basics* (ISSN 1728-1431) (EISSN 1735-9244)
International **J**ournal of **E**ngineering *Transactions B: Applications* (ISSN 1728-144X) (EISSN 1735-9244)
International **J**ournal of **E**ngineering *Transactions C: Aspects* (ISSN 2423-7167) (EISSN 1735-9244)
Web Sites: www.ije.ir & www.ijeir.info E-mails: ije.editor8@gmail.com, Tel: (+9821) 88771578, Fax: (+9821) 88773352
Materials and Energy Research Center (MERC)

ISSN: 1728-144X, e-ISSN: 1735-9244

International Journal of Engineering, Volume 36, Number 08, August 2023

CONTENTS

Transactions B: Applications

ISSN: 1728-144X, e-ISSN: 1735-9244

International Journal of Engineering, Volume 36, Number 08, August 2023

IJE TRANSACTIONS B: Applications Vol. 36, No. 08, (August 2023) 1398-1408

International Journal of Engineering

Journal Homepage: www.ije.ir

Optimal Operation of Multi-microgrid System Considering Uncertainty of Electric Vehicles

M. Gholami*[a], M. J. Sanjari[b]

[a] *Department of Electrical Engineering, University of Science and Technology of Mazandaran, P.O. Box 48518-78195, Behshahr, Iran*
[b] *School of Engineering, Griffith University, QLD 4222, Australia*

PAPER INFO

Paper history:
Received 08 February 2023
Received in revised form 16 May 2023
Accepted 18 May 2023

Keywords:
Microgrid
Electrical Vehicle
Monte-Carlo Simulation
Uncertainty Management

ABSTRACT

Integration of electric vehicles (EVs) into the power systems has been a concern for distribution system operators due to their impacts on several aspects of power system operation, such as congestion management, power quality, voltage regulation, and peak time changing. In this paper uncertainty parameters such as charging time, traveled distance, and plug-in location of EVs are considered and their effects on the optimal daily operation of microgrids (MG) are discussed. A power system, including geographically-adjacent quasi-independently controlled MGs, each of which has a different operation objective function (OF) is modeled in this paper. A set of socioeconomic OFs i.e. minimum purchase power from the main grid, maximum usage of green power, and minimum Expected Energy Not Supplied (EENS) are considered for each MG which appear in the optimization process with different weights based on the MG policy. The effect of EV integration into the Multi Microgrid System (MMS) is also investigated in this paper and the performance effectiveness of different operation management policies against EV integration is discussed.

doi: *10.5829/ije.2023.36.08b.01*

NOMENCLATURE

Acronyms	
DG	Distributed generation
MG	Microgrid
MMS	Multi microgrid system
EV	Electric vehicles
PV	Photovoltaic
D-IPFC	Distribution Interline Power Flow Controller
GA	Genetic Algorithm
S3P	Small power generation
V2G	Vehicle to grid
MCS	Monte-Carlo simulation
SOC	state of charging
PDF	probabilistic distribution function
OF	Objective function
ESS	Energy Storage Systems

Variables	
n_m	Number of batteries in *m*-th microgrid
e_m	Number of electrical vehicles in *m*-th microgrid
h	Time horizon (hour)
C_h	Cost of energy at time *h* ($)
P_h	Purchased energy from the main grid at time *h* (kWh)
$E^h{}_j$	Greenhouse emission of *j*-th DG at time *h* (ton/h)
$P^h{}_j$	Generation power of *j*-th DG at time *h* (kW)
α_j	Greenhouse emission rate of *j*-th DG (ton/kWh)
τ_i	Interruption time of *i*-th bus (h)
$P_{i,h}{}^D$	Load on the *i*-th bus at time h (kWh)
SOC(h)	Battery state of charge at time *h* (kWh)
$SOC_{n_m}^{k_m}$	State of charge in n_m-th battery in k_m-th microgrid (kWh)
$SEV_{n_m}^{k_m}$	Start time of e_m-th EV charging in k_m-th microgrid (h)
η_{bat}	ESS charging/discharge efficiency (%)
$P_{PV,h}$	PV generated power at time *h* (kWh)
$P_{bat,h}$	Battery charge/discharge power at time *h* (kWh)
$P_{l,h}^{i}$	Power demand at time h in bus i (kWh)
$P_{ev,h}^{i}$	EVs charging power at time *h* in bus i (kWh)
P_{loss}	Active power loss (kWh)
ROC	Rate of Charge (kWh/h)
ROD	Rate of Discharge (kWh/h)
SOC_{min}	Minimum charge level of battery (kWh)
SOC_{max}	Maximum charge level of battery (kWh)
P_{move}	EV energy consumption rate (kWh/km)
ΔL	Distance traveled by the EV (km)
L	Distance between different zones (MGs) (km)

Subscripts	
i	Index of buses
l	Index of loads
m	Index of microgrid
j	Index of power source

*Corresponding Author Email: m.gholami@mazust.ac.ir (M. Gholami)

Please cite this article as: M. Gholami, M. J. Sanjari, Optimal Operation of Multi-microgrid System Considering Uncertainty of Electric Vehicles, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1398-1408

1. INTRODUCTION

High penetration of distributed generation (DG) units and storage systems along with emerging the concept of microgrid (MG) [1, 2] have enhanced the reliability of electrical energy supply [3]. Another technology significantly changed the paradigm of power system operation is EV, which has gained considerable attention in the last years [4-6] because they have negligible CO_2 emission, which has a significant influence in decreasing greenhouse gases. However, strategies for their charging and discharging cycles are the primary concern of distribution and MGs operators [7].

Management of energy in such systems has several concerns, such as optimal management of DGs [8, 9]. Wouters et al. [10] addressed the necessity of designing a local energy system via the integrity of DG and microgrids. A mixed-integer linear programming model was introduced to optimize local energy systems. Based on the proposed model by Wouters et al. [10], DGs, heating units, and storage systems can supply the electrical, and cooling/heating energy of a small residential neighborhood. A central controller was considered in the proposed model. The system's annual cost was considered as the OF and GAMS software was used for solving the problem. The concept of multiagent systems was used for the optimal operation of microgrids [11]. Genetic Algorithm (GA) [12] was used as a meta-heuristic algorithm for the optimal operation of microgrids. The concept of multi-microgrid has been introduced in this paper for clustering the available houses. This concept was introduced by Arefifar et al. [13] and it has been modified in this paper. A distribution interline power flow controller (D-IPFC) was introduced as a new device by Kargarian and Rahmani [14]. By presenting a model for injection power by D-IPFC, Kargarian and Rahmani [14] showed that it can improve the operational capability of the distribution system. The D-IPFC was used to connect several MGs to form a MMS. The nonlinear loads have been modeled in the optimal operation of the standalone microgrid [15]. Energy management based on contingency analysis for a MMS has been presented by Aghdam et al. [16]. The economic comparison between the microgrid development versus the conventional distribution system has been discussed by Parag and Ainspan [17].

Moreover, the uncertain parameters make the decision-making process more complicated. Niknam et al. [18] presented a stochastic model for optimal management of energy in a microgrid, in which the operating cost and greenhouse gas emission were minimized. Uncertain parameters such as load, wind turbine, and photovoltaic power output as well as the tariff of purchasing electricity from the main grid were considered. A scenario-based method, i.e. roulette wheel mechanism, was used for uncertainty modeling considering the normal distribution function for input parameters. Some new devices have been introduced for improving the distribution system capability. Khodaei et al. [19] investigated the MG planning problem and its economic viability deployment. The optimal generation mix of distributed energy resources (DERs) for installation were determined considering uncertainties. A robust optimization approach was adopted for considering uncertainty in load forecast error, variable renewable generation, market prices, and microgrid islanding. Xiang et al. [20] developed a scenario-based robust energy management method accounting for the worst-case amount of renewable generation and load. The economic and robust model was formulated to maximize the total exchange cost while getting the minimum social benefits cost at the same time. The uncertainty of renewable generation and load demand was described as an uncertain set produced by interval prediction. Then, Taguchi's orthogonal array testing method was used to provide possible testing scenarios.

The storage system has incrementally grown in the networks in the last several years [21, 22], due to the high penetration of distributed energy resources such as wind turbines and photovoltaics [23] and the intermittency nature of renewable energies [24]. ESSs have brought many benefits to the power system such as short-term power supply, improving the power quality, and ancillary services in microgrids. A cost-based formulation was presented for the optimum sizing of storage units in the microgrid [25, 26]. Xiao [27] proposed the hierarchical control of ESS, composed of both centralized and distributed control to enhance system reliability. Xu et al. [28] used ESSs to support the frequency control in microgrid systems, due to the intermittency of the renewable generation and constantly changing load demand. A distributed cooperative control strategy was proposed for coordinating the ESSs to maintain the supply-demand balance and minimize the total power loss associated with charging/discharging inefficiency. A review of hybrid energy storage system usage in standalone microgrids has been proposed by Jing et al. [29]. In fact, different control strategies have been compared by Jing et al. [29].

The electric vehicles can perform as a storage system in the distribution networks and at the same time act as a distributed load for the distribution operator [30]. The EV was defined as a small power generation (S3P) for improving the security and reliability of the power system [31]. Vehicle-to-grid (V2G) technology can significantly affect power grids, but there should be a smart program for electrical parking lots. Zhang et al. [32] redefined the unit commitment problem by considering demand response and EVs. These technologies can reform the demand curve of the grid and can be used as a reserve source as well. Lin et al. [33] presented the distribution system planning by considering charging stations of EVs. The costs regarding the investment, operation, and maintenance were considered as OFs. Rana et al. [34] introduced a modified droop control for frequency support of microgrids based on EVs. Derakhshandeh et al. [35] used EVs for the coordination of generation scheduling in an industrial microgrid manner.

In this paper, a new model for the optimal daily operation of geographically-adjacent MGs, including PV integration is presented. In the proposed model, the uncertain parameters related to EVs are considered for the hour-ahead scheduling process. The uncertainties of the daily traveled distance of EVs inside a MG and among MGs are considered in the modeling. There are different uncertain parameters such as charging time, traveled distance, and number of EVs in the network, which are considered in this paper with Monte-Carlo simulation (MCS) [1]. MMS consists of several quasi-independent MGs, each of which has different OFs, i.e. minimum energy cost, minimum greenhouse gas emission, and maximum reliability.

The contributions of this paper are as follows:

- Behavior of MMS with different OFs is analyzed and its optimal operation with the presence of daily travelled distance of EVs as an uncertain parameter is investigated.
- A new model for the daily optimal operation of geographically-adjacent MGs, including PV integration is proposed.
- A comparison is made to assess the effect of weighting factors on the optimal operation of MMS.

The rest of the paper is organized as follows. System modeling is introduced in section 2, in which the mathematical models of all elements of the microgrid are presented. In section 3, the objective function, constraints, decision variable, and pseudo code of the optimal operation optimization of the MMS system are proposed. This optimization is the main function of the central controller in each microgrid in connected mode. In section 4, the proposed method is implemented on a MMS system and the results are discussed. Finally, section 5 concludes the paper.

2. MULTI-MICROGRID POWER SYSTEM MODELING

2. 1. System Description and Modeling

In this section, a MMS, which is composed of some adjacent MGs is modeled, each of which includes several EVs whose charging time intervals are controlled by a central controller. As shown in Figure 1, the EVs' batteries can be charged by connecting to the local distribution system. As mentioned earlier, the starting time of charging the EVs is managed by the microgrid central controller while the EV's owners set the allowed timespan for this purpose.

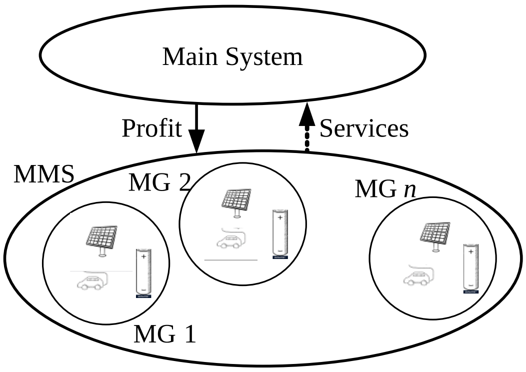


Figure 1. MMS layout

As shown in Figure 1, each MG has at least a PV source and a battery to store the excess energy of photovoltaic sources in the case that surplus energy generation exists. The daily charging and discharging plan of the batteries is controlled by the microgrid operator.

The MMS can have two types of controllers [13]. In one type a central controller is considered for all microgrids, as studied in this paper while in the other type, each microgrid has a dedicated controller. The former version has lower implementation costs.

The distribution lines are parameterized based on pi-section modeling. As the length of the lines in the distribution systems and microgrids is short, this model is reasonably accurate. The shunt admittance should be considered because there are underground cables in the distribution network; therefore, a simplified short-line model without shunt admittance modeling will lead to inaccurate results.

Power output by the PV system is considered in the power system analysis as a deterministic variable as the main objective of this paper is about the effect of uncertainty of EVs on MMS operation.

EVs are modeled based on their charging rate, which is assumed to be a constant value. This charging is added to the load profiles of each home in load flow studies.

$$\sum_i P_{i,h}^{D} = P_{l,h}^{i} + P_{ev,h}^{i} \tag{1}$$

The cost of energy is assumed to be based on the market price, which is variable over the day. The time horizon for each time interval for price change is considered one hour.

$$Energy\,Cost = \sum_h C_h P_h \tag{2}$$

The upstream network is modeled as an all-the-time available infinite bus. Although it can be considered that the upstream network has limited availability, but this assumption does not affect the presented method.

2. 2. Uncertainty Modeling– Monte Carlo Simulation The following steps are presented for scheduling the proposed multi-microgrid structure.

Step 1: Collect the customers' load data and PV power generation. Real data sets are used in this step. The customer loads and photovoltaic generations are based on the real data of residential loads in Tehran.

Step 2: Random number generation. As it is concerned before, MCS is used to model the uncertainty. So, random numbers are generated for every single EV in each MG, the distance moved in each day, and the charging hour of each EV. Normal probability distribution function (PDF) is used for the number of EVs and uniform PDF is used for the distance covered and charging hour. Each random number set is concerned with a scenario.

Step 3: Solving the optimization problem. The optimization problem of each microgrid based on its own OFs and different scenarios is solved. The number of optimization problems is equal to the number of microgrids. Each MG has its own OF. Different objectives and constraints are presented in the next section. The decision variables in these problems are the daily charging and discharging plan of batteries and the charging time of EVs.

Step 4: Data Analysis. The results of optimization problems are analyzed and the exchanging power between the microgrids and main grids is stored.

3. MMS OPTIMAL OPERATION

In this section, the OFs and constraints of the MMS optimal operation problem are explained. Three OFs are considered, i.e. the cost of energy, gas emission, and reliability. It is assumed that the MGs are connected to the main grid from which the required energy is supplied.

3. 1. Objective Functions MGs are running with different OFs, i.e. minimization of cost of energy, minimization of green gas emission, and minimization of expected not-supplied energy. The weighted sum of the mentioned OFs results in the proposed OF in this paper. The weight factors can be tuned based on the global and upstream rules and/or objectives of the microgrid operators, which can be changed from time to time, based on the nature of the grid and special events of the year. The OFs are stated in Equations (3), (4), and (6).

1) Cost of the Energy

$$OF_1 = \sum_h C_h P_h \tag{3}$$

In this paper, the time horizon for each time interval for price charging is considered one hour. Ph is considered negative if the generated power in the microgrid is more than the demand in each hour and positive when there is a power surplus in MG. It is assumed that the excess energy of the microgrid can deliver to the main grid. In other words, the connection between the main grid and microgrids is bi-direction.

2) Greenhouse Gas Emission

$$OF_2 = \sum_j \sum_h E_h^j \tag{4}$$

E_h^j can be considered as a coefficient of Phj based on Equation (5). The available source of energy in the microgrids is PV systems which are emission-free. However, the excess energy which is purchased from the main grid causes greenhouse gas emissions. This explains why the greenhouse gas emission rate shows a straight relation with the delivered energy from the upstream grid to the MG.

$$E_h^j = \alpha_j P_h^j \tag{5}$$

3) Expected Energy not Supplied

$$OF_3 = EENS = \sum_i \tau_i P_i^D \tag{6}$$

Values of τi, denoting the average yearly interruption time for each bus are calculated based on historical data, i.e. yearly unavailable periods.

The decision variables in this problem are the set of SOC of batteries in each microgrid and the time of EV charging. Based on the SOC the amount of charging/discharging of the battery is calculated as follows:

$$P_{bat}(h) = \eta_{bat}\left(SOC(h) - SOC(h-1)\right) \tag{7}$$

where Pbat is the amount of charging (positive value) and/or discharging (negative value) of the battery and SOC(h) is the SOC of the battery in hour, h.

3. 2. Constraints The constraints are as follows:

1) Power Balance

The generation and demand values of active power should be equal at all times to prevent frequency deviation in the system.

$$P_h + P_{PV,h} = \sum_L P_{l,h} + P_{ev,h} + P_{bat,h} + P_{loss} \tag{8}$$

2) Battery SOC limitation

The SOC of the battery has limitations to guarantee that it works in safe operating conditions.

$$SOC_{\min} \le SOC(h) \le SOC_{\max} \tag{9}$$

The upper and lower limits are designed based on the battery structure, type, and usage. In some cases, SOCmin can be zero, but in other ones always there should be some minimum charge.

3) Intraday energy transfer

The SOC of the battery on the first and the last time interval of a day are considered equal. This assumption makes the proposed algorithm comparable with the others without dependency on the initial condition of the battery.

$$SOC(h=1)=SOC(h=24) \tag{10}$$

4) EVs charging Power

The available charge of EVs is calculated by knowing the SOC at the last time interval and traveled distance by the EV, the latter is reflected in Pmove.

$$SOC(h)=SOC(h-1)-P_{move}\times\Delta L \tag{11}$$

Based on the distance which is covered by an EV, the discharging amount of the EV is calculated. The discharging rate of the EVs is considered constant and by multiplying the distance in the discharging rate, the amount of reduction in SOC is calculated.

5) EV Charging Time

In this paper, it is considered that the charging time of EVs is 2 hours continuously because the discontinuous charging of batteries will reduce their lifetime.

6) Battery charge/discharge rate limitation

In this paper, the charge and discharge rates have been limited as follows:

$$\begin{cases} SOC(h)-SOC(h-1)<ROC & if \quad SOC(h)>SOC(h-1) \\ SOC(h)-SOC(h-1)<ROD & if \quad SOC(h-1)>SOC(h) \end{cases} \tag{12}$$

3. 3. Decision Variables

The decision variables in this problem are the SOC of batteries in each hour and the time of EV charging. While the EV owners set the desired interval for charging, the starting time is optimally decided by the central control. In other words, the following set of decision variables is determined optimally by the central controller of MMS.

$$\begin{aligned} U=[&SOC_1^{k_1},...,SOC_{n_1}^{k_1},...,SOC_1^{k_m},...,SOC_{n_m}^{k_m},\\ &SEV_1^{k_1},...,SEV_{e_1}^{k_1},...,SEV_1^{k_m},...,SEV_{e_m}^{k_m}] \end{aligned} \tag{13}$$

3. 4. Optimization Procedure

As one of the best optimization algorithms in discrete variables, GA is adopted for MMS optimal operation in this paper. As mentioned in step 3 of section 2.B, MCS is used in this paper.

Two sets of input data are provided for the GA. The first set includes deterministic data such as the load profiles of customers and generated power by PV systems, while the second set consists of probabilistic data such as the number of EVs in each MG, the daily traveled distance by EVs, and the charging hour of each EV. The output of the GA is the power system control parameters, i.e. decision variables defined in Equation (12). SOC variables address the optimal charge/discharge of the battery and SEV shows the optimal starting time of EV charging.

The parameters of optimization are selected based on the knowledge of the authors and the GA toolbox of MATLAB is employed for this purpose. The selected GA parameters are presented in Table 1.

The following pseudo-code summarizes the procedure that is implemented on MMS optimal operation problem.

```
Loop (for all the scenarios)
        Collect Real Data for Loads
        Collect Real Data for PV Generation
        Generate Random Data for the Number of EVs in each MG
        Generate Random Data for Distance Moved by each EV
        Generate Random Data for the charging hour of each EV
        [SOC, SEV]← GA Optimization
        Save the Output of GA
```

The flowchart of the optimization process is shown in Figure 2.

4. SIMULATION RESULTS AND DISCUSSION

4. 1. System Description and Assumptions

As shown in Figure 3 a system consisting of three microgrids connected to the main grid is considered in this paper. These microgrids are geographically close to each other. The distance between the two parts of MMS is shown by L in Figure 3.

It is assumed that the scheduling horizon is one day (24 hours). The time step is 1 hour, and it is assumed that the load and distributed generation are constant in each step.

It is assumed that the EVs travel some round trips and the charging place of the EV is on its parking. It is assumed that the SOC values of the batteries are equal in the first and last hour of the scheduling horizon. It is assumed that each microgrid has one ESS with a capacity of 600 kWh. The ESS can discharge to 15% of its capacity. The charging/discharging rate of the ESS is considered 100 kWh/hour. ESS usage has two main reasons:

1. The maximum generation of PV and maximum consumptions of loads are different and also take place in

TABLE 1. GA parameters

Parameter	Value
Number of Iteration	200
Population Type	Double vector
Selection Operators	Stochastic uniform
Mutation Operators	Gaussian
Percent of Mutation	20%

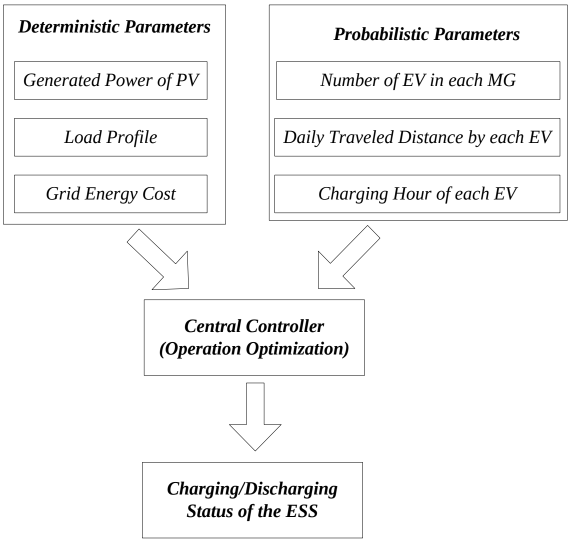


Figure 2. Optimization process flowchart

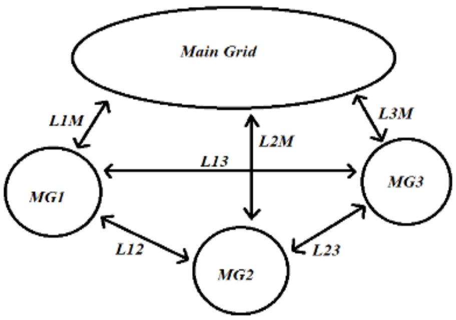


Figure 3. A 3-MG system

different time intervals. The ESS can solve this problem by charging in maximum PV generation time and discharge in the load maximum consumption times.

2. The main grid electricity price varies during the day. The ESS can be charged during low-price hours and can be discharged during peak-price hours.

The ESS has two operating modes with different objectives, i.e. energy management in the connected mode and frequency/voltage control in the islanded mode of the microgrid. In this paper, it is assumed that the energy management function of the ESS is investigated. It is considered that there are some photovoltaic, EVs and loads in each microgrid, which are presented in Table 2.

Load profiles of each MG are presented in Figure 4, which are based on real data which are collected from some residential loads in the city of Tehran, Iran. It is assumed that there are 4, 3, and 5 load centers in microgrids 1, 2, and 3, respectively. So, there are 4, 3, and 5 load profile curves in each part of Figure 4.

The PVs generation profiles are presented in Figure 5. These values are collected based on real data for the city of Tehran, Iran.

TABLE 2. Number of PV Sources and Loads in Each Microgrid

Microgrid	Installed PV capacity (kW)	Maximum Loads (kWh)
1	1400	1705
2	540	577
3	900	915

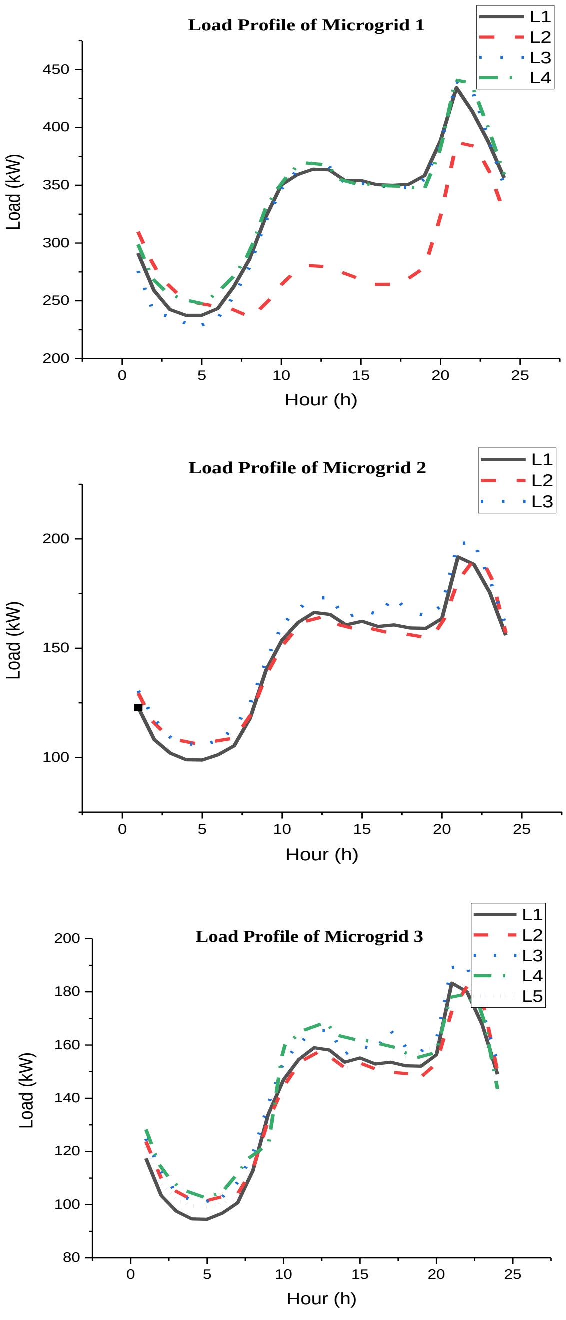


Figure 4. Load Profile of Microgrid

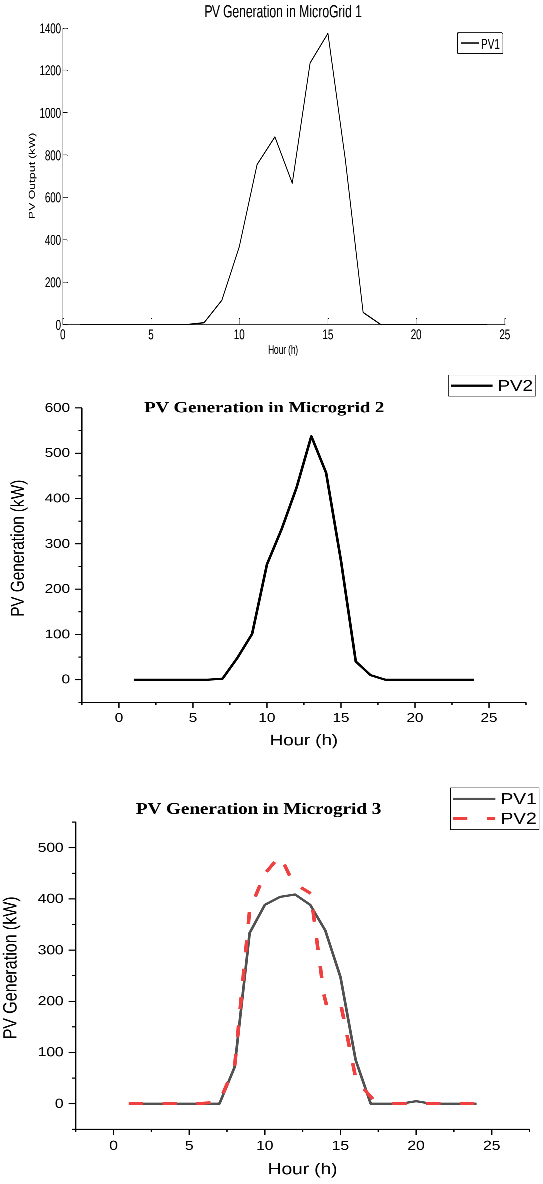


Figure 5. PV power generation

The average yearly interruption time is considered equal for all customers of each microgrid. The average yearly interruption time of microgrids 1, 2 and 3 are considered 21, 12 and 9 hours per year, respectively. These values are collected from literature [36].

4. 2. Monte Carlo Simulation The uncertain parameters are modeled with MCS in this paper with 1000 scenarios generated based on the random generation of uncertain parameters based on which the optimization problems are solved. The uncertain parameters are as follows:

- The number of EVs in each microgrid. The PDF of this parameter is considered a normal distribution function, with a mean value of 50 and a standard deviation of 20 [37].
- The distance travelled by EVs. The uniform distribution is considered for this parameter. It is considered that the EVs can travel between the microgrids and between their microgrid and the main grid. It is considered that travel is a round trip. The number of EVs which travel from each microgrid through other microgrids and the main grid is generated based on the random number generation process.
- Time of charging commencement of EVs. The uniform distribution addresses the statistical distribution of this parameter.

The dispersion of the number of EVs in each MG is shown in Figure 6.

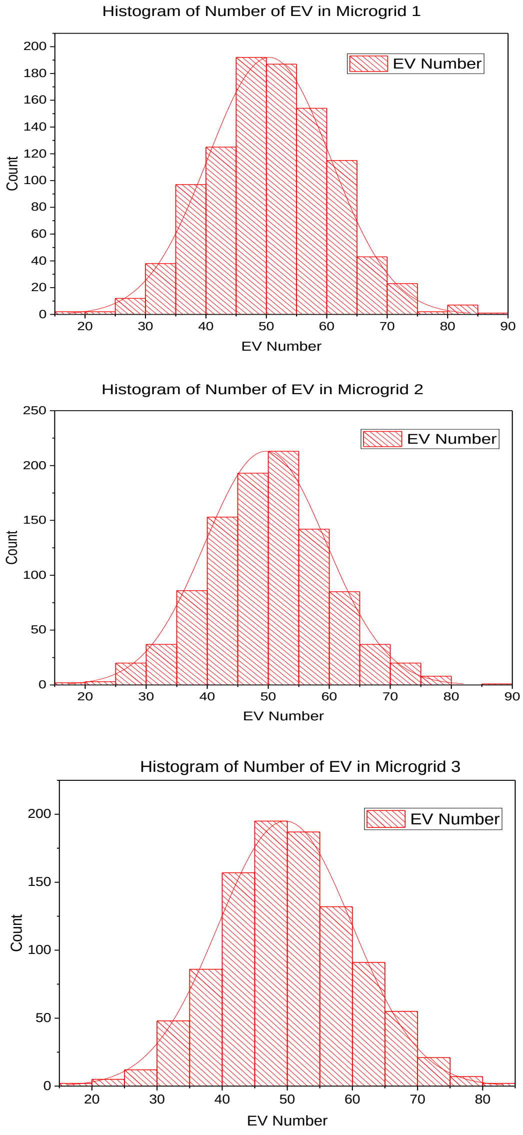


Figure 6. Dispersion of EV Numbers

The charging of EV batteries is reduced based on the traveled distance. The energy reduction of EV batteries based on each trip between the grids is presented in Table 3.

The weighting coefficients of OFs for microgrids are different. So, the cost of energy, emission, and reliability of microgrids are different. The selected weighting coefficients are presented in Table 4.

The results in this section are categorized into two parts: scenario 1, which shows the base case results of MMS, and scenario 2, which discusses the optimal results.

4. 3. Scenario 1: Base Case

For a better presentation of the effect of the quality of the optimization problem, at first, the MG is planned without the optimization problem. The results of this case are presented in Table 5.

TABLE 3. Energy Deployment of EV Batteries for Travel among MGs (kWh)

Grid	Microgrid1	Microgrid2	Microgrid3	Main Grid
MG 1	-	5	4	3
MG 2	5	-	7	5
MG 3	4	7	-	7
Main Grid	3	5	7	-

TABLE 4. Weighting Coefficients for different OF in MGs

MG number	OF1: Cost of Energy	OF2: Emission Cost	OF3: EENS
MG 1	0.7	0.2	0.1
MG 2	0.1	0.7	0.2
MG 3	0.2	0.1	0.7

TABLE 5. Average and Standard Deviation of Energy Cost for Each MG Before Optimization

Objective Function	MG number	Average ($)	Standard Deviation ($)	Max ($)	Min ($)
Cost of Energy	MG 1	6538.1	627.29	8927.3	4653.3
	MG 2	4609.4	699.94	6877.8	2710.3
	MG 3	5458.1	767.31	7760.9	3472.7
Emission	MG 1	524.23	11.62	558.78	493.55
	MG 2	350.39	8.40	375.33	331.50
	MG 3	467	9.09	487.94	440.90
EENS	MG 1	2425.2	13.6	2461.4	2392.6
	MG 2	1106.3	10.3	1134.8	1074.6
	MG 3	60.4	2.1	66.8	54.1

4. 4. Scenario 2: Optimal Operation of MMS

The weighting coefficients of OFs for microgrids are different. So, the cost of energy, emission and reliability of microgrids are different. The higher coefficient for the cost of energy causes a lower cost and a higher coefficient for the EENS leads to a higher cost of energy and hence higher reliability. The summary of the result is presented in Table 6 and the comparison is presented in Table 7.

As shown in Table 7, the optimization process reduces the cost of energy by 2.4%, 3.4%, and 2.8% in microgrids 1, 2, and 3, respectively. This reduction in cost causes a $474 daily reduction and a $173010 annual saving in energy cost.

TABLE 6. Average and Standard Deviation of Energy Cost for Each MG

OF	MG number	Average ($)	Standard Deviation ($)	Max ($)	Min ($)
Cost of Energy	MG 1	6380.8	627.30	8777.5	4496.3
	MG 2	4451.3	700.59	6719.9	2553
	MG 3	5300.2	767.76	7603.2	3308.9
Emission	MG 1	522.7	10.2	556.7	499.4
	MG 2	330.3	8.9	352.6	308.1
	MG 3	451.7	9.03	478	431.3
EENS	MG 1	2411.2	14.1	2445.2	2376.1
	MG 2	1084.3	8.5	1108.1	1060.2
	MG 3	56.2	2.3	61.8	51.1

TABLE 7. Comparison between scenarios 1 (Base Case) and 2 (Optimal Operation)

OF	MG Number	Average (optimized)	Average (non-optimized)	Percent of Improvement
Cost of Energy	MG 1	6380.8	6538.1	2.4%
	MG 2	4451.3	4609.4	3.4%
	MG 3	5300.2	5458.1	2.8%
Emission	MG 1	522.7	524.23	0.29%
	MG 2	330.3	350.39	5.73%
	MG 3	451.7	467	3.28%
EENS	MG 1	2411.2	2425.2	0.58%
	MG 2	1084.3	1106.3	1.99%
	MG 3	56.2	60.4	6.95%

In other words, the optimization process reduces the emission by 0.29%, 5.73%, and 3.28% in microgrids 1, 2 and 3, respectively and the reduction in EENS is 0.58%, 1.99%, and 6.95% in microgrids 1, 2 and 3, respectively.

As shown in Table 7, the effect of weighting coefficients, listed in Table 4, on the percentage of reduction of each part of OF is interesting. As an example, the reductions in the cost of energy, emission and EENS for the MG1 after applying the optimization process are 2.4%, 0.29%, and 0.58%, respectively. The interesting point is the relation between the reduction percent and the weighting coefficients of these three parts, which are 0.7, 0.2, and 0.1, respectively. As it is shown in Table 7, the higher value of the coefficient resulted in more reduction in OF optimization. This procedure is repeated for both microgrids 2 and 3.

To show the effects of weighting factors on the results, the weighting factors of MG 1 are changed and the microgrid OFs are obtained. Three scenarios are analyzed as follows:

1. The base case, in which 0.7, 0.2, and 0.1 are weighting factors for objective functions of 1, 2, and 3, respectively.
2. Scenario 1, in which 0.1, 0.7, and 0.2 are weighting factors for objective functions of 1, 2, and 3, respectively.
3. Scenario 2, in which 0.2, 0.1, and 0.7 are weighting factors for objective functions of 1, 2, and 3, respectively.

The results listed in Table 8 show the weighting factors' effects on the objective functions. The first column shows the number of scenarios and the second one shows the objective functions considered under that scenario. Being listed in the third column, the weighting factors are shown and the fourth and fifth columns show the average OF value considering all the random numbers generated by MCS, after and before the optimization process, respectively. The average value is selected as a descriptive index to show the performance of the optimization process. The last column shows the improvement percentage in the objective function value due to the optimization process.

As an example, the convergence process of one of the 1000 scenarios is shown in Figure 7.

Figures 8 and 9 depict the effect of changing the weighting factors on the value of the OFs., in which the changes of OF 1 and 2 versus the weighting factors variations are drawn. It is assumed that the weighting factor of OF 1 is fixed to 1.0 and that of the other OFs is changed from 0.1 to 4.0. The changes of OF 1 and 2 are shown in Figures 8 and 9, respectively. Each curve in these figures shows a fixed amount of objective function.

TABLE 8. Average and Standard Deviation of Energy Cost for MG1

Scenario	OF Number	W	Average (optimized)	Average (non-optimized)	Percent of Improvement
Base Case	OF1	0.7	6380.8	6538.1	2.40%
	OF2	0.2	522.7	524.23	0.29%
	OF3	0.1	2411.2	2425.2	0.58%
Scenario 1	OF1	0.1	6501.5	6538.1	0.56%
	OF2	0.7	516.52	524.23	1.47%
	OF3	0.2	2399.8	2425.2	1.05%
Scenario 2	OF1	0.2	6421.5	6538.1	1.78%
	OF2	0.1	523.41	524.23	0.16%
	OF3	0.7	2374.9	2425.2	2.07%

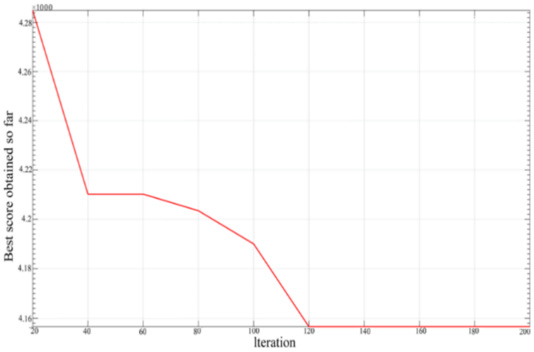


Figure 7. The convergence process of optimization algorithm

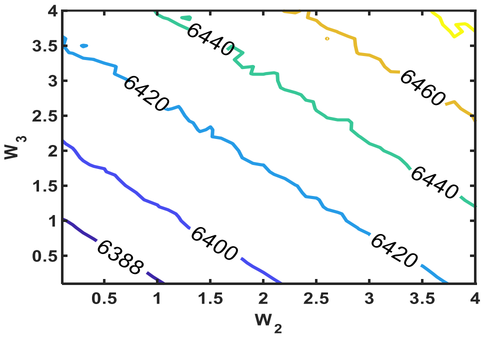


Figure 8. Changes of OF1 based on changing weighting factors

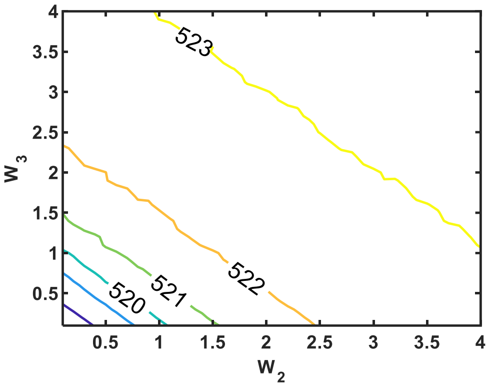


Figure 9. Changes of OF2 based on changing weighting factors

5. CONCLUSION

In this paper, an MMS, including three MGs with different operation objectives, was modeled and its operation was investigated. Three OFs considered for the microgrids are the cost of energy, greenhouse gas emission, and expected not-supplied energy. The number of EVs in each microgrid was considered by appropriately-shaped normal PDF and uniform density functions were adopted to consider the EVs traveled distance and their charging time. MCS was used to generate scenarios of uncertain parameters and the effect of uncertainty of EV numbers on the energy cost, EENS and gas emission cost was discussed. It was shown that the objective functions were decreased according to their weights, set by the MG operator. The optimal operation of MMS was also determined by adopting GA to the multi-objective MMS operation problem. Comparing two cases, i.e. base case and optimal operation showed that the optimization process led to a decrease in the cost of energy in MMS, enhancing the reliability index of the MG and greenhouse gas emission reduction.

6. REFERENCES

1. Gholami, M., Sanjari, M. and Gharehpetian, G., "Pmu-based voltage stability assessment in microgrids by anns considering single contingencies", ***International Review of Electrical Engineering-Iree***, Vol. 7, No. 6, (2012), 6317-6323.
2. Eini, M.K., Moghaddam, M.M., Tavakoli, A. and Alizadeh, B., "Stability analysis of ac/dc microgrids in island mode", ***International Journal of Engineering, Transactions A: Basics***, Vol. 34, No. 7, (2021), 1750. doi: 10.5829/ije.2021.34.07a.20.
3. Lasseter, R., Akhil, A., Marnay, C., Stephens, J., Dagle, J., Guttromsom, R., Meliopoulous, A.S., Yinger, R. and Eto, J., *Integration of distributed energy resources. The certs microgrid concept*. 2002, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).
4. Beer, S., Gómez, T., Dallinger, D., Momber, I., Marnay, C., Stadler, M. and Lai, J., "An economic analysis of used electric vehicle batteries integrated into commercial building microgrids", ***IEEE Transactions on Smart Grid***, Vol. 3, No. 1, (2012), 517-525. doi: 10.1109/TSG.2011.2163091.
5. Luthander, R., Shepero, M., Munkhammar, J. and Widén, J., "Photovoltaics and opportunistic electric vehicle charging in the power system–a case study on a swedish distribution grid", ***IET Renewable Power Generation***, Vol. 13, No. 5, (2019), 710-716. doi: 10.1049/iet-rpg.2018.5082.
6. Figueiredo, R.E., Monteiro, V., Ferreira, J.C., Afonso, J.L. and Afonso, J.A., "Smart home power management system for electric vehicle battery charger and electrical appliance control", ***International Transactions on Electrical Energy Systems***, Vol. 31, No. 4, (2021), e12812. doi: 10.1002/2050-7038.12812.
7. Madzharov, D., Delarue, E. and D'haeseleer, W., "Integrating electric vehicles as flexible load in unit commitment modeling", ***Energy***, Vol. 65, (2014), 285-294. doi: 10.1016/j.energy.2013.12.009.
8. Nikmehr, N. and Najafi-Ravadanegh, S., "Optimal operation of distributed generations in micro-grids under uncertainties in load and renewable power generation using heuristic algorithm", ***IET Renewable Power Generation***, Vol. 9, No. 8, (2015), 982-990. doi: 10.1049/iet-rpg.2014.0357.
9. Marzband, M., Alavi, H., Ghazimirsaeid, S.S., Uppal, H. and Fernando, T., "Optimal energy management system based on stochastic approach for a home microgrid with integrated responsive load demand and energy storage", ***Sustainable Cities and Society***, Vol. 28, (2017), 256-264. doi: 10.1016/j.scs.2016.09.017.
10. Wouters, C., Fraga, E.S. and James, A.M., "An energy integrated, multi-microgrid, milp (mixed-integer linear programming) approach for residential distributed energy system planning–a south australian case-study", ***Energy***, Vol. 85, (2015), 30-44. doi: 10.1016/j.energy.2015.03.051.
11. Dimeas, A.L. and Hatziargyriou, N.D., "Operation of a multiagent system for microgrid control", ***IEEE Transactions on Power Systems***, Vol. 20, No. 3, (2005), 1447-1455. doi: 10.1109/TPWRS.2005.852060.
12. Li, P., Xu, D., Zhou, Z., Lee, W.-J. and Zhao, B., "Stochastic optimal operation of microgrid based on chaotic binary particle swarm optimization", ***IEEE Transactions on Smart Grid***, Vol. 7, No. 1, (2015), 66-73. doi: 10.1109/TSG.2015.2431072.
13. Arefifar, S.A., Ordonez, M. and Mohamed, Y.A.-R.I., "Energy management in multi-microgrid systems—development and assessment", ***IEEE Transactions on Power Systems***, Vol. 32, No. 2, (2016), 910-922. doi: 10.1109/TPWRS.2016.2568858.
14. Kargarian, A. and Rahmani, M., "Multi-microgrid energy systems operation incorporating distribution-interline power flow controller", ***Electric Power Systems Research***, Vol. 129, (2015), 208-216. doi: 10.1016/j.epsr.2015.08.015.
15. Jha, S., Hussain, I., Singh, B. and Mishra, S., "Optimal operation of pv-dg-battery based microgrid with power quality conditioner", ***IET Renewable Power Generation***, Vol. 13, No. 3, (2019), 418-426. doi: 10.1049/iet-rpg.2018.5648.
16. Aghdam, F.H., Salehi, J. and Ghaemi, S., "Contingency based energy management of multi-microgrid based distribution network", ***Sustainable Cities and Society***, Vol. 41, (2018), 265-274. doi: 10.1016/j.scs.2018.05.019.
17. Parag, Y. and Ainspan, M., "Sustainable microgrids: Economic, environmental and social costs and benefits of microgrid deployment", ***Energy for Sustainable Development***, Vol. 52, (2019), 72-81. doi: 10.1016/j.esd.2019.07.003.
18. Niknam, T., Azizipanah-Abarghooee, R. and Narimani, M.R., "An efficient scenario-based stochastic programming framework for multi-objective optimal micro-grid operation", ***Applied Energy***, Vol. 99, (2012), 455-470. doi: 10.1016/j.apenergy.2012.04.017.

19. Khodaei, A., Bahramirad, S. and Shahidehpour, M., "Microgrid planning under uncertainty", ***IEEE Transactions on Power Systems***, Vol. 30, No. 5, (2014), 2417-2425. doi: 10.1109/TPWRS.2014.2361094.

20. Xiang, Y., Liu, J. and Liu, Y., "Robust energy management of microgrid with uncertain renewable generation and load", ***IEEE Transactions on Smart Grid***, Vol. 7, No. 2, (2015), 1034-1043. doi: 10.1109/TSG.2014.2385801.

21. Bahramirad, S., Reder, W. and Khodaei, A., "Reliability-constrained optimal sizing of energy storage system in a microgrid", ***IEEE Transactions on Smart Grid***, Vol. 3, No. 4, (2012), 2056-2062. doi: 10.1109/TSG.2012.2217991.

22. Sankarkumar, R.S. and Natarajan, R., "Energy management techniques and topologies suitable for hybrid energy storage system powered electric vehicles: An overview", ***International Transactions on Electrical Energy Systems***, Vol. 31, No. 4, (2021), e12819. doi: 10.1002/2050-7038.12819.

23. Bahmani-Firouzi, B. and Azizipanah-Abarghooee, R., "Optimal sizing of battery energy storage for micro-grid operation management using a new improved bat algorithm", ***International Journal of Electrical Power & Energy Systems***, Vol. 56, No., (2014), 42-54. doi: 10.1016/j.ijepes.2013.10.019.

24. Branco, H., Castro, R. and Lopes, A.S., "Battery energy storage systems as a way to integrate renewable energy in small isolated power systems", ***Energy for Sustainable Development***, Vol. 43, (2018), 90-99. doi: 10.1016/j.esd.2018.01.003.

25. Chen, S.X., Gooi, H.B. and Wang, M., "Sizing of energy storage for microgrids", ***IEEE Transactions on Smart Grid***, Vol. 3, No. 1, (2011), 142-151. doi: 10.1109/TSG.2011.2160745.

26. Al-Ghussain, L., Samu, R., Taylan, O. and Fahrioglu, M., "Sizing renewable energy systems with energy storage systems in microgrids for maximum cost-efficient utilization of renewable energy resources", ***Sustainable Cities and Society***, Vol. 55, (2020), 102059.

27. Xiao, J., Wang, P. and Setyawan, L., "Hierarchical control of hybrid energy storage system in dc microgrids", ***IEEE Transactions on Industrial Electronics***, Vol. 62, No. 8, (2015), 4915-4924. doi: 10.1109/TIE.2015.2400419.

28. Xu, Y., Zhang, W., Hug, G., Kar, S. and Li, Z., "Cooperative control of distributed energy storage systems in a microgrid", ***IEEE Transactions on Smart Grid***, Vol. 6, No. 1, (2014), 238-248. doi: 10.1109/TSG.2014.2354033.

29. Jing, W., Hung Lai, C., Wong, S.H.W. and Wong, M.L.D., "Battery-supercapacitor hybrid energy storage system in standalone dc microgrids: Areview", ***IET Renewable Power Generation***, Vol. 11, No. 4, (2017), 461-469. doi: 10.1049/iet-rpg.2016.0500.

30. Ahmadigorji, M. and Mehrasa, M., "A robust renewable energy source-oriented strategy for smart charging of plug-in electric vehicles considering diverse uncertainty resources", ***International Journal of Engineering, Transactions A: Basics***, Vol. 36, No. 4, (2023), 709-719. doi: 10.5829/ije.2023.36.04a.10.

31. Saber, A.Y. and Venayagamoorthy, G.K., "Intelligent unit commitment with vehicle-to-grid—a cost-emission optimization", ***Journal of Power Sources***, Vol. 195, No. 3, (2010), 898-911. doi: 10.1016/j.jpowsour.2009.08.035.

32. Zhang, N., Hu, Z., Han, X., Zhang, J. and Zhou, Y., "A fuzzy chance-constrained program for unit commitment problem considering demand response, electric vehicle and wind power", ***International Journal of Electrical Power & Energy Systems***, Vol. 65, No., (2015), 201-209. doi: 10.1016/j.ijepes.2014.10.005.

33. Lin, X., Sun, J., Ai, S., Xiong, X., Wan, Y. and Yang, D., "Distribution network planning integrating charging stations of electric vehicle with V2G", ***International Journal of Electrical Power & Energy Systems***, Vol. 63, (2014), 507-512. doi: 10.1016/j.ijepes.2014.06.043.

34. Rana, R., Singh, M. and Mishra, S., "Design of modified droop controller for frequency support in microgrid using fleet of electric vehicles", ***IEEE Transactions on Power Systems***, Vol. 32, No. 5, (2017), 3627-3636. doi: 10.1109/TPWRS.2017.2651906.

35. Derakhshandeh, S.Y., Masoum, A.S., Deilami, S., Masoum, M.A. and Golshan, M.H., "Coordination of generation scheduling with pevs charging in industrial microgrids", ***IEEE Transactions on Power Systems***, Vol. 28, No. 3, (2013), 3451-3461. doi: 10.1109/TPWRS.2013.2257184.

36. Berndt, E.R. and Wood, D.O., "Technology, prices, and the derived demand for energy", ***The review of Economics and Statistics***, (1975), 259-268. doi: 10.2307/1923910.

37. Khodayar, M.E., Wu, L. and Shahidehpour, M., "Hourly coordination of electric vehicle operation and volatile wind power generation in scuc", ***IEEE Transactions on Smart Grid***, Vol. 3, No. 3, (2012), 1271-1279. doi: 10.1109/TSG.2012.2186642.

Persian Abstract

چکیده

ادغام وسایل نقلیه الکتریکی در سیستم های قدرت به دلیل تأثیرات آنها بر چندین جنبه از عملکرد سیستم قدرت، مانند مدیریت ازدحام، کیفیت توان، تنظیم ولتاژ و تغییر زمان پیک، یک نگرانی برای اپراتورهای سیستم توزیع بوده است. در این مقاله پارامترهای عدم قطعیت مانند زمان شارژ، مسافت طی شده و محل اتصال EV ها در نظر گرفته شده و اثرات آنها بر عملکرد روزانه بهینه ریزشبکه ها مورد بحث قرار گرفته است. یک سیستم قدرت، شامل MGهای شبه مستقل کنترل شده از نظر جغرافیایی مجاور، که هر کدام تابع هدف عملیاتی متفاوتی دارند، در این مقاله مدل‌سازی شده‌اند. مجموعه ای از OF های اجتماعی-اقتصادی یعنی حداقل توان خرید از شبکه اصلی، حداکثر استفاده از توان سبز و حداقل انرژی مورد انتظار تامین نشده برای هر MG در نظر گرفته می شود که در فرآیند بهینه سازی با وزن های مختلف بر اساس خط مشی MG ظاهر می شود. اثر ادغام EV در سیستم چند ریزشبکه نیز در این مقاله بررسی شده و اثربخشی عملکرد سیاست‌های مدیریت عملیات مختلف در برابر یکپارچه‌سازی EV مورد بحث قرار گرفته است.

International Journal of Engineering

Journal Homepage: www.ije.ir

Low Embodied Carbon and Energy Materials in Building Systems: A Case Study of Reinforcing Clay Houses in Desert Regions

R. Taherkhani*[a], M. Alviri[b], P. Panahi[c], N. Hashempour[d]

[a] *Department of Civil Engineering, Imam Khomeini International University, Qazvin, Iran*
[b] *Department of Civil Engineering, Alborz University, Qazvin, Iran*
[c] *Department of Civil Engineering, Faculty of Engineering and Technology, Iran University of Science and Technology, Tehran, Iran*
[d] *Department of Civil Engineering, University of British Columbia (UBC), Vancouver, Canada*

PAPER INFO

Paper history:
Received 24 February 2023
Received in revised form 04 May 2023
Accepted 09 May 2023

Keywords:
Embodied Energy
Embodied Carbon
Sustainable Building
Construction Material
Life Cycle Assessment
Historical Textures

ABSTRACT

Over 40% of the world's energy consumption occurs in the construction sector. However, some countries do not address environmental criteria as design requirements in their construction codes. Accordingly, this research aims to provide a solution that reduces embodied energy and carbon while preserving historical and traditional textures of Iran. The comparison of embodied carbon and energy between new concrete and traditional buildings was performed by calculating the amount of construction materials. By examining both types of buildings, the reduction of embodied carbon and energy in a combined building system was evaluated. In the following, using SWOT analysis, the strategies of this combination were investigated. Clay building has less embodied energy and carbon than concrete one despite containing more mass of materials. According to SWOT analysis, the strategy of integrating clay and concrete systems is presented. The proposed system in compare to the concrete structure resulted in around 40% and 35% reduction in embodied carbon and energy, respectively. Extending this strategy throughout the country saves 13 million tons of embodied carbon and 130 million GJ of embodied energy. Finding a solution based on sustainability considerations to preserve historical texture is one of the basic concerns of countries where these textures form a part of their identity. The presented combined system, while paying attention to sustainable building and urban development, is a desirable solution to reduce buildings' embodied carbon and energy.

***doi**: 10.5829/ije.2023.36.08b.02*

NOMENCLATURE

E_e	Total embodied energy	$c_{p,j}$	Amount of emitted energy intensity for building material j
E_m	Embodied energy of manufacturing construction materials	$e_{p,j}$	Amount of emitted carbon for building material j
E_t	Energy consumption of transporting building materials	$c_{i,j}$	Amount of carbon emitted by producing the building materials j
E_p	Embodied energy related to building productions	$c_{t,l}$	Amount of carbon emission
C_e	Total embodied carbon	$e_{t,l}$	Amount of energy use
C_m	Embodied carbon related to the manufacturing of materials	d_l	Distance of transportation
C_t	Amount of carbon emission of transporting the materials	k	Number of building materials and elements
C_p	Value of embodied carbon emanated from different processes	n	Number of countries which material or element j is imported
$e_{i,j}$	Energy required for manufacturing the materials j in country i	**Greek Symbols**	
$q_{i,j}$	Amount of building materials j imported from the country i	μ_j	Replacement factor for building elements j
$Q_{p,j}$	Amount of building material j	λ_j	Factor for waste materials j

1. INTRODUCTION

Sustainable development is an internationally well-known philosophy defined in "Our Common Future" report published by the World Commission on Environment and Development (WCED), as "the development that meets the needs of the present generation without compromising the ability of future

*Corresponding Author Email: taherkhani@eng.ikiu.ac.ir (R. Taherkhani)

Please cite this article as: R. Taherkhani, M. Alviri, P. Panahi, N. Hashempour, Low Embodied Carbon and Energy Materials in Building Systems: A Case Study of Reinforcing Clay Houses in Desert Regions, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1409-1428

generations to meet their own needs" [1]. To achieve sustainable development, economic, social, environmental, and technological aspects are major factors that should be considered [2]. The significance of sustainable development planning causes it to be taken into account in different industries and various aspects of human life. The construction industry as one of the most efficient industries is not an exception to this consideration and sustainability is the major issue in which the construction industry is engaged with it [3, 4]. The relationship between sustainable development and the construction industry is undoubtedly evident [5] and the discourse of construction practitioners and decision-makers worldwide has begun to appreciate and acknowledge the advantages of sustainable building [6]. Sustainable construction was proposed to make the construction processes, activities, and practices more economically, socially, and environmentally responsive [7]. Research shows that among the sustainability dimensions, the main focus is on environmental sustainability [8]. In this regard, the green area is important to balance the ecosystem [9]. To achieve green areas, considerations related to buildings should be taken into account. Buildings, as the main components of cities, have a special and effective role in the emergence of environmental problems [10]. About 40 to 50 percent of the emission of greenhouse gases emanates from the construction industry [11]. In addition, the construction sector is responsible for one third of the global carbon dioxide (CO_2) emissions [12]. A significant amount of natural resources has been consumed by the building industries which is responsible for a noticeable amount of energy usage such that the building operations alone account for 30 to 40 percent of total energy use, globally [11, 12]. The environmental impacts of buildings need mitigation and adaptation strategies [13] and energy efficiency in the construction field needs to be seriously pursued, using approaches ranging from increasing research and development investment to maintaining appliance standards [14].

In order to satisfy sustainability objectives [15], cope with climate change, and reduce greenhouse gas emissions of buildings, the optimization of building design and operation is needed [16]. On the other hand, historical and traditional textures indicate the background and identity of some countries like Iran and are among the factors affecting foreign tourists. While many of the existing buildings in the historical textures are deteriorating and destroyed. Therefore, there is a serious need for solutions to preserve historical textures and maintain their attractiveness for tourists while respecting the technical aspects of construction. Considering that very limited research supports the provision of appropriate solutions to preserve the historical textures while highlighting sustainability considerations in Iran, it is felt necessary to address this issue.

In this regard, the current research, while emphasizing sustainable building and reviewing previous studies on embodied energy and carbon, will examine the life cycle energy assessment of buildings and the waste coefficient and lifespan of materials in the historical and traditional contexts of Iran. Finally, the importance of evaluating embodied carbon and energy in Iran's construction industry will be considered. Then, embodied carbon and energy evaluation has been carried out for both traditional clay houses and conventional concrete buildings. The necessity of modifying the construction codes with the aim of sustainable development in the country was investigated. By comparative studies, the obtained results and measuring the embodied carbon and energy increase after changing the constructional system and materials from traditional to conventional. According to SWOT analysis and the proposed strategies to preserve historical texture, to demonstrate the effect of substituting materials on the amount of energy and embodied carbon, the energy and carbon were analyzed considering a combined building with concrete structure, frame, and joist system, the adobe walls, and the traditional flooring. The presented combined building system, while paying attention to sustainable building with reduced embodied carbon and energy, preserves the structure of historical textures and its attractiveness for tourists.

1. 1. Embodied Energy

In the building sector, the energy and the greenhouse emissions embodied in the building materials are becoming dramatically important [17]. Hence, in the past few years, embodied energy (EE) has become a prominent research field. Due to the growing awareness that the energy initially consumed to produce goods and services might prevail in determining the whole amount of life-cycle energy [18]. Accordingly, several studies have focused on reducing greenhouse gas emissions and energy use during a building's life. It is crucial to assess energy requirements to come up with efficient energy-saving solutions [19]. Buildings are reckoned as major consumers of energy. There are various types of energy used during the life cycle of a building including embodied energy, maintenance, operational energies, disposal, and demolition energies. Embodied and operational energies account for a major portion in this regard. Embodied energy (EE) shows the whole energy consumed for the construction of a building, i.e., a sum of embodied energies of building materials, energies related to the transportation of materials, and building construction energy. Embodied energy contributes 10–20% of the lifecycle energy consumption of conventional buildings [20]. The embodied energy of building materials has a noticeable portion of the whole embodied energy in buildings. Hence, it is vital to choose suitable construction materials considering their embodied energy to diminish embodied energy of buildings. High embodied energy in buildings is expected by using energy-intensive construction materials such as steel, brick, glass, and cement [21].

Energy consumption in buildings includes two main components: operational and embodied energy. Operational energy is the energy needed for running a building by different operating processes such as cooling, heating, and lighting, whereas embodied energy of a building indicates the energy used by all of the processes associated with the production of the building from the mining and processing of natural resources to manufacturing, transportation, and product delivery. Embodied energy has been defined by different researchers, who have given various nuances to the concept, yet a general consensus exists emphasizing that embodied energy in building materials has increased its importance in a building's life cycle in comparison with operating energy thanks to the better energy performance of the buildings [22]. The consumption of embodied energy is a physical process highly related to material inventory flows which have been determined at the design or pre-construction stage. Dixit et al. [23] presented The first approach to standardization of the embodied energy of various materials. Determining embodied energy is a time-consuming and complex task. Moreover, there is no generally accepted method to accurately and consistently determine the embodied energy; thus, wide variations in measurement figures are inevitable [22]. Several research studies were conducted in the embodied energy field are presented in Table 1. As the table above presents, research on embodied energy has been conducted since 30 years ago in several countries, but Iran is not among them. According to these research projects, a wide range of embodied energy values (25.2 to 27208 MJ/m^2) has been mentioned, which relates to (1) discrepancies in the type of structural systems and materials, (2) differences in the amount of embodied energy of units of different materials in different countries and cities, (3) energy efficiency during operation, (4) building usage type and (5) mass construction. In general, green and wooden buildings possess the lowest embodied energy, while buildings with high energy efficiency during the operation period (low energy, very low energy, net-zero energy consumption, and passive house) have the highest embodied energy values. Some studies have

TABLE 1. Previous studies on embodied energy

Year	Author	Ref.	Building Type	Embodied Energy MJ/m^2	Life Span
1994	Buchanan and Honey	[24]	Conventional Building	5530	-
1995	Debnath et al.	[60]	Conventional Concrete Building	5000	-
1995	Suzuki et al.	[61]	Different houses in Japan	10400,2700	-
1997	Adalberth	[54]	Residential single-unit precast buildings	3014, 3487	50
2002	Thormark	[62]	All three projects with Low energy consumption	7033, 4388, 4079	50
2004	Mithraratne and Vale	[63]	Standard lighting (low energy consumption), Standard concrete (low energy consumption), High insulation (low energy consumption)	4424, 4709, 5088	100
2006	Casals	[64]	Average (Conventional), High Energy efficiency (low energy consumption)	3679.2, 14191.2	30
2007	Nässén, J et al.	[65]	Villa vision, Multi-unit buildings	6200, 5800	-
2008	Huberman, N. and D. Pearlmutter	[66]	Student Dormitory Complex	3280, 4910	-
2009	Utama and Gheewala	[67]	High Rise, Residential Buildings	1666.8, 1470.8	40
2010	Blengini and Carlo	[68]	Standard house (low-energy family house)	7560, 10990	70
2010	Gustavsson and Joelsson	[25]	Low-energy residential buildings	3504	50
2010	Ramesh et al.	[69]	Office and residential buildings	25.2, 385.2	50
2010	Gustavsson and Joelsson	[70]	Eight-story wood-framed apartment	3510	50
2011	Leckner and Zmeureanu	[71]	Conventional, Net Zero Energy House without the solar systems, Net Zero Energy House (NZEH) with solar combisystem	4820.4, 6020.4, 8936.4, 8780.4	40
2012	Dahlstrøm et al.	[72]	Passive house	7516.5, 7590, 7914.5, 7718	50
2013	Paulsen and Sposto	[73]	Social houses (low energy consumption)	7200	50

2013	Paleari et al.	[74]	Zero Energy Residential Buildings	16728	100
2013	Berggren et al.	[75]	Net Zero Energy Buildings	6912, 10584, 8208, 7344, 7344, 9504	60
2014	Stephan and Stephan	[76]	Low-rise residential buildings	27208	50
2017	Dissanayake et al.	[77]	House with recycled expanded polystyrene (EPS) based foam concrete wall panels	3460	-
2018	Vitale et al.	[78]	Residential Prefab LSF, Residential Traditional concrete	9900, 8500	50
2019	Praseeda et al.	[21]	Rural dwellings	2340-2800	50
2019	Tavares et al.	[79]	Prefabricated house with: Steel material, Concrete material, Timber, LSF	5624, 2151, 2335, 2619	100
2019	Thanu et al.	[80]	Conventional residential building	4060	-

recommended the use of wood and soil to diminish embodied energy [24]. For example, in 2010, Gustavsson and Joelsson [25] investigated an 8-story wooden building with a lifespan of 50 years and obtained 3500 MJ per square meter of embodied energy. Some research projects demonstrated the necessity of using local building materials to decrease embodied energy [26]. The traditional buildings in the desert regions of Iran are also made of indigenous materials such as clay and soil. These buildings have also roofs made of wood, and in this respect, they can be classified as low-carbon buildings. But there is a lack of studies in Iran measuring the embodied energy of the aforementioned buildings to compare the amount of energy.

1. 2. Embodied Carbon On a large scale, buildings account for 67% of embodied carbon emissions [27]. The emission of greenhouse gases like carbon dioxide causing climate change plays the most important role in sustainable development. The CO_2 emissions related to energy consumption have risen by 66% to reach a historic high of 33.1 Gt in 2018 compared to 1990 level [28]. Carbon emissions are usually denoted as CO_2 (i.e. CO_2 equivalent), which is a measurement unit according to the relative impact of a given gas on global warming (the so-called global warming potential). For instance, the 100-year global warming potential (GWP) of methane is equal to 25, which means that the effect of 1 kg of methane gas on climate change is equal to the influence of 25 kg of carbon dioxide on that. In other words, 1 kg of methane gas would count as 25 kg of CO_2 equivalent. Table 2 displays typical sources and GWP of various greenhouse gases over 100 years. Through a survey on a building with a lifespan of 40 years in 2009, Shukla et al. [29] concluded that using materials with low embodied energy rather than high embodied energy reduces carbon dioxide emissions to approximately 101 tons per year. In 2010 Ortiz-Rodríguez et al. [30] conducted a simultaneous study in Colombia and Spain on buildings with a lifespan of 50 years, showing that the energy and carbon of the construction period and the operational carbon for the building located in Colombia were, 4940 MJ per square meter, 238 kg carbon equivalent per square meter, and 599 kg per square meter, respectively. For the building located in Spain, these values are 4180 MJ per square meter, 192 kg of carbon equivalent per square meter and 2250 kg of carbon equivalent per square meter [30]. An investigation carried out in Portugal for buildings with a lifespan of 50 years in 2011-2012 revealed that greenhouse gas emissions are 13 kilograms of carbon equivalent per square meter per year [31].

In 2015, Atmaca and Atmaca [32] investigated the carbon content and embodied energy of the construction, operation, and demolition period of two buildings located in the urban and rural areas with a lifespan of 50 years. They obtained the percentage of operational energy as 73% and 76%, construction energy as 24% and 27%, and operational carbon as 59% and 74% [32]. As shown in Table 3, there are some research projects have been conducted in the field of embodied carbon. According to the above table, in a 2007 study of semi-detached houses in Scotland, Asif estimated carbon emissions of 618 kilograms per square meter. The results indicate that 99% of carbon emission is related to mortar and concrete [33]. In 2011, Monahan and Powell [34] investigated a building in which wood was the predominant material, but the most embodied carbon amounts were related to concrete, which indicates the significance of choosing low-carbon materials. In 2016, Luo et al. [35] revealed that as the building height increase, the amount of CO_2 emissions per unit area augments significantly. Also, the amount of CO_2 emissions per unit area of super-high-rise buildings is 1.5 times that of multi-story buildings; the CO_2 emissions in the field of Civil Engineering are responsible for 75% of the total construction materialization stage; and the carbon emissions of steel, concrete, mortar and wall materials account for 80% of the Civil Engineering sector [35]. Therefore, the strategy of preserving the historical texture discussed in this survey, in which buildings have a maximum of two stories, can be effective in reducing embodied carbon. A study by Gan et al. [36] in 2017 demonstrates that 10 to 20% of the reduction of embodied carbon can be fulfilled using cement substitutes. It is also shown that if recycled materials are employed, transportation will account for 20% of embodied carbon [36]. Therefore, the strategy of substituting concrete with low-carbon indigenous materials, used in this research, can be effective in reducing embodied carbon. Using indigenous materials also reduces the embodied carbon of transportation due to distance reduction. Teng's research results revealed that a reduction of wall thickness can diminish embodied carbon (with a 1.9% reduction potential) [37]. This strategy has been employed in the current study to reduce embodied carbon.

1. 3. Life Cycle Energy Assessment of Buildings

The interest in Life Cycle Assessment (LCA) has increased dramatically since the 1990s, especially with the advent of scientific publications. LCA is a tool to evaluate the environmental impacts and resources applied during the life cycle of a building, i.e., from the acquisition of raw materials, through the production and use phases, to waste management. The methodological development in LCA has been strong, and it is widely employed in practice. LCA is an exhaustive evaluation considering all attributes or aspects of the natural environment, human health, and resources. The LCA methodological development has been strong over the past decades [38]. Although the focus of LCA can be on social and economic effects, the environmental impacts have been the main focus of LCA. Engineers and designers designing and developing technical systems and products need to be able to study and size up life cycle assessment data about the alternatives they are considering, and the environmental sustainability

TABLE 2. Typical sources and GWP of various greenhouse gases [43]

Greenhouse gas	Chemical formula	GWP	Typical sources
Carbon dioxide	CO_2	1	Energy combustion, biochemical reactions
Nitrous oxide	N_2O	298	Fertilizers, car emissions, manufacturing
Methane	CH_4	25	Decomposition
Perfluorocarbon	PFC	7,390 - 12,200	Aluminum smelting
Hydrofluorocarbon	HFC	124 - 14,800	Refrigerants, industrial gases
Sulfur hexafluoride	SF_6	22,800	Switch gears, substations

TABLE 3. Previous studies on Embodied carbon

Year	Author	Ref.	Building Type	Embodied Carbon kg/m^2	Life Span
2007	Asif et al.	[33]	Semi-detached house	618	-
2008	Hacker et al.	[81]	Semi-detached house	332.70	100
2009	Blengini	[82]	Residential building	8 kgCO2E/m2 year	40
2010	Ortiz et al.	[30]	Dwelling in Colombia Dwelling in Spain	238 192	50
2011	Monahan et al.	[34]	Semi-detached house	405	-
2012	Monteiro	[83]	Single-family house in Portugal	13 kgCO2E/m2 year	50
2013	ChaoMao et al.	[84]	Semi-prefabricated construction conventional construction	336 368	-
2014	Lamnatou et al.	[85]	Building-integrated solar thermal collector	160 kg CO2.eq/m2	30
2015	Galua et al.	[86]	Green building	21000	20
2016	Luo et al.	[35]	78 office buildings	326.75	50
2017	Gan et al.	[36]	High-rise buildings	459 kg CO2-e/m2	30
2018	Kumanayake et al.	[87]	Office building	629.6 kg	_
2019	Teng et al.	[37]	Prefabricated high-rise public residential buildings	561	50
2020	Kayaçetin et al.	[27]	Residential house	409.2 kgCO2-eq/m2	50

specialists among them are also required to carry out the LCA studies [39]. When implementing LCA, the design/development phase is usually excluded, since it is often assumed not to contribute markedly. However, it should be considered that all decisions made in the phase of development/design can greatly affect the environmental impacts in the other life cycle stages. The design of a product can highly predetermine its behavior in the next phases. As a result, this paper focuses on the design stage.

When implementing sustainable development in the building sector, the focus needs to be on the long perspective entailing the significance of considering the whole life cycle of a building [8]. LCA is a strong tool to assess potential environmental influence from the extraction of materials and production, through construction and use or service phase to the waste treatment and end-of-life of the product [40]. Furthermore, LCA is one of the best tools to size up environmental impacts through all phases of the building according to a conclusion drawn and reported by the European Commission [41]. Some of the merits of using LCA assisting in terms of sustainability in the building sector are economic, social, and environmental. Environmental merits are followed by making a comparison between alternative products and providing information about environmental effects helping stakeholders to make informed decisions [42]. Consequently, buildings need to be assessed considering their whole life cycle, which entails both production and end-of-life stages and is not merely based on the energy demand throughout the use stage [17]. Most of the existing literature focused on the analysis of embodied energy of main construction materials such as steel, cement, and glass as the sources of embodied energy, and ignored other equipment inputs and materials [12].

In this regard, life cycle assessment concerning energy and carbon dioxide emission is divided into several categories as follows [43]:

1) Cradle-to-gate carbon emissions: Carbon emissions between the confines of the 'cradle' (earth) up to the factory gate of the final processing operation. This consists of mining, raw materials extraction, processing, and manufacturing.
2) Cradle-to-site carbon emissions: Sum of cradle-to-gate emissions and delivery to the installation and construction site.
3) Cradle-to-end of construction: Sum of cradle-to-site and assemblies on-site and construction.
4) Cradle-to-grave carbon emissions: Sum of cradle-to-end of construction and maintenance, renovation, demolition, disposals, and waste treatment.
5) Cradle-to-cradle: Cradle-to-grave emissions plus, converting the components into new components at the end of their life with an equal or lower quality.

Embodied carbon and energy are the emitted carbon and consumed energy measured through one of the above categories. Embodied carbon is usually presented in kilograms of CO_2 per kilogram of material or product, and embodied energy is expressed in megajoule energy per kilogram of material or product. The whole life cycle assessment provides important information, but there are lots of factors introducing more complexity to LCA in the building industries [44]. For example, the expected lifetime of buildings is usually more than 50 years which is a long lifetime, therefore, accurate prediction of all lifetime behavior of the project from cradle-to-grave is very difficult [45, 46]. There has been much research conducted on the life cycle energy assessment (LCEA) of buildings. Some important ones are presented in Table 4. The life cycle energy assessment is an exhaustive task, and cannot be fulfilled without calculating the embodied energy. Hence, to complete the life cycle analysis, there have been several studies calculating the amount of embodied energy around the world, as shown in Table 1, but Iran is not among them. Therefore, conducting such studies in Iran is of special necessity.

1. 4. Waste Coefficient and Lifespan of Materials

The construction industry produces nearly 35% of waste in landfills across the globe [47]. One of the most voluminous and heaviest waste streams produced in the European Union (EU) is construction and demolition waste (C&DW). It accounts for nearly a third of the waste produced which is more than 850 million tons [48]. In the UK, 44% of waste in 2013, was due to the construction sectsor [49]. Also, in 2014, the amount of C&DW generated by the UK was equal to 58 million tons [50]. The rate of C&DW generation (kg per capita per day) in Iran is six times more than that of the USA [51]. While the average of C&DW generated in the United States is 0.77 kg per capita per day, that average is equal to 4.64 kg per capita per day in Iran, according to reported data by Tehran Municipality Waste Management [52].

The definition of waste is important since the classification of substances as waste is the basis for the formulation of waste management policy and the application of regulatory controls to protect the environment as well as human health [53]. According to the EU Waste Framework Directive (European Community 1991), waste is defined as any substance or object that the holder discards or intends to discard or has to discard. The materials are considered waste under one of the following circumstances [53]:

- If the objects or substances have been discarded.
- If the objects or substances cannot be utilized anymore for their original design objective or elsewhere with the same design objectives.
- If the objects or substances are produced more than required.

TABLE 4. Previous studies on Life Cycle Energy Assessment (LCEA)

Year	Author	Ref.	Building Type	50 years Life Cycle Energy (GJ/m^2)	Energy Contribution
1997	Adalberth	[54]	Residential single-unit precast buildings	27.4, 31.7	Embodied energy: 11-12%, Renovation energy: 4-5%, Operational energy: 84%, Destruction energy: 0.3-0.5%
2002	Thormark	[62]	20 apartments	15.24	Embodied energy: 46%
2004	Mithraratne and Vale	[63]	Three residential concrete buildings with high insulation	17.02, 16.24, 11.83 (for 100 years)	Operational energy: in order 74%, 71%, 57%
2007	Citherlet and Defaux	[88]	Three variants of a family house	10-29	_
2007	Sartori and Hestnes	[89]	Conventional and low-energy buildings		Embodied energy: (Conventional) 2-38% (Low-energy) 9-46 %
2008	Utama and Gheewala	[90]	Houses made of clay bricks and concrete blocks	12.56, 13.24	Operational energy: in order 6.7 %, 6.2%
2009	Utama and Gheewala	[67]	Residential high-rise buildings with a double and single wall system	3.33, 5.41	Operational energy: in order 28%, 16%
2010	Ramesh et al.	[69]	Office and residential buildings	118.8-1404 kWh/m^2	Embodied energy: 7-107 kWh/m^2 Operational energy: 0-330 kWh/m^2 (about 80 to 90 %)
2010	Gustavsson and Joelsson	[70]	Eight-story wood-framed apartment building	1800-3672 kWh/m^2	Embodied energy: 45-60 %
2017	Ma et al.	[12]	Office building	345 *kWh/m2/year*	Embodied energy: 20 % Operational energy: 73 %
2019	Praseeda et al.	[21]	Rural dwellings	0.77-4.05	Embodied energy: 69 %, Operational energy: 0-2 %
2019	Petrovic et al.	[91]	Wooden single-family house	30.16 (for 100 years)	Operational energy: 64 %
2019	Hernandez et al.	[92]	Residential block	3.85	-
2019	Tettey et al.	[93]	Multi-story residential building with different materials	4060-11700 kWh/m^2 (for 80 years)	-

- If some of the materials and equipment remain and cannot be returned to the seller or sold to another person.
- If the materials or equipment cannot be operated after construction and installation.
- If the substances are discarded owing to rework, demolition during construction, low quality of the final product, modification of work, work changes, executive orders of principals, regulations, time delays, planning problems, budgeting and financing problems, productivity, and the quality of human resource and other such things.

The more waste a building has, the more amount of embodied carbon and energy it has. In the current study, the waste coefficient of the most widely used materials in Iran is obtained from questionnaires and interviews with professionals and is shown in Table 5. The lifespan of different materials is presented in Table 6 [54, 55].

TABLE 5. Waste coefficient of materials in Iran

Material	Waste coefficient	Material	Waste coefficient
Concrete	0.063	Polystyrene	0.0407
Steel	0.0645	Mosaic	0.0593
Cement- Slurry	0.1005	Stone	0.0959
Brick	0.0896	Cupper	0.0709
Coating	0.119	Mortar	0.1023
Ceramic tiles	0.0775	Aggregate	0.0468
Aluminum	0.0208	Glass	0.0329
Paint	0.0521	Bitumen	0.0903
Plastic	0.0078	Asphalt	0.0806

1. 5. The Importance of Investigation on Embodied Carbon and Energy in Iran's Construction Industry

There are some research projects conducted on embodied carbon and energy evaluation per unit of various materials. But for reasons such as different energy efficiency, export, import,

TABLE 6. The lifespan of materials

Material	Life Span	Material	Life Span
Cement	50	Iron	50
Concrete	50	Aluminum	30
Concrete – Cement replacement with fly ash (0-30)%	50	Bronze	30
Concrete – Cement replacement with furnace slag (0-30)%	50	Mosaic	40
Floor carpet - nylon	50	Ceramic	14
Vinyl flooring	50	Brick	50
Sealants and adhesives	50	Lead	50
Plastic - UPVC - Window	30	Copper	50
Aggregate	50	Brass	30
Sand	50	Wood	30
Soil	50	Linoleum	50
Clay	50	Isolation	50
Lime	50	Rubber	40
Asphalt	50	Coating	50
Bitumen	50	Glass	30
Facade Stone	50	Paint	10
Steel	50		

industrial and environmental conditions; these values vary for different countries and even different parts of a country. Few studies in Iran have investigated embodied carbon and energy while life cycle assessment accomplishment is impossible without considering this issue. Construction codes in Iran merely consider operational energy standards and embodied carbon and energy have not been regarded yet. Statistics show an increase in carbon emissions from 2003 to 2014 in Iran which will continue if not controlled [56]. Countries with high CO_2 emissions aim to reduce emissions by at least 25% until 2030; unfortunately, Iran is not among them [57]. Consumption of energy and waste of energy in Iran is higher than the world's average, and the contribution to air pollution is higher than expected as well. Based on statistics, China, the USA, India, Russia, Japan, Germany, South Korea, Iran, Saudi Arabia, and Indonesia are the top ten CO_2 emitters among all countries in the world according to their emission trends throughout the 1991–2015 period [57].

Therefore, Iran is among the top ten countries in the world with high CO_2 emissions. Using an annual increase of 5% as an assumption, the total amount of CO_2 emissions in Iran is predicted, by Mousavi et al. [56] to reach 985 million tons in 2025. Concerning the percentage change in CO_2 emissions, India, Indonesia, and Saudi Arabia without doubt are at the top of the increase in carbon emissions list (the percentage growth of their CO_2 emissions is either greater than or equal to 100%), followed by China, Iran, and South Korea. Although Saudi Arabia and Iran have not been committed to any Intended Nationally Determined Contribution (INDC) goals that would bring about international attention and discussion in the future, the situation of carbon reduction is grim in these countries. An adjustment in the structure of energy consumption is an urgent and inevitable requirement to reach the win-win combination of economic growth and CO_2 reduction, especially for countries such as Saudi Arabia and Iran which are petroleum-rich countries [57].

Although the building industry accounts for a considerable amount of carbon emission and energy consumption and the life cycle assessment in terms of energy and carbon is not fulfilled without considering embodied energy and carbon, the studies on embodied energy and carbon are very limited in Iran, and researchers merely focus on the operational energy standards, and embodied energy is not considered in energy codes. Therefore, in this paper, embodied carbon and energy evaluation has been carried out for both traditional clay houses and conventional concrete buildings, and the necessity of modifying the construction codes with the aim of sustainable development in the country was investigated by comparing the obtained results and measuring the embodied carbon and energy increase after changing the constructional system and materials from traditional to conventional. Figure 1 indicates a flow chart in which the research process of this article is illustrated.

2. METHODOLOGY

This research project aims to study the observance of environmental issues in Iran's desert regions by comparing the amount of energy consumption and carbon emission of concrete and traditional buildings and providing energy and carbon reduction strategies. In this survey, the positive effects of using optimal structural systems and materials concerning reducing carbon emission and energy consumption, and the amount of this reduction will be dealt with. The research flowsheet is shown in Figure 1. To achieve the above-mentioned objective, first, a traditional house was selected as a case study, and then a concrete building with a plan similar to that of a traditional building was designed using ETABS and SAFE software keeping the spaces as it was.

The case study is a clay house with a lime concrete foundation located in Yazd city, with a coordinate of 31°54'09"N, 54°22'02"E, and an altitude of 1212 meters above sea level. This house with an area of 383 square meters relates to 200 years ago in the Qajar period and is registered in the name of "Ehramianpour House" in the

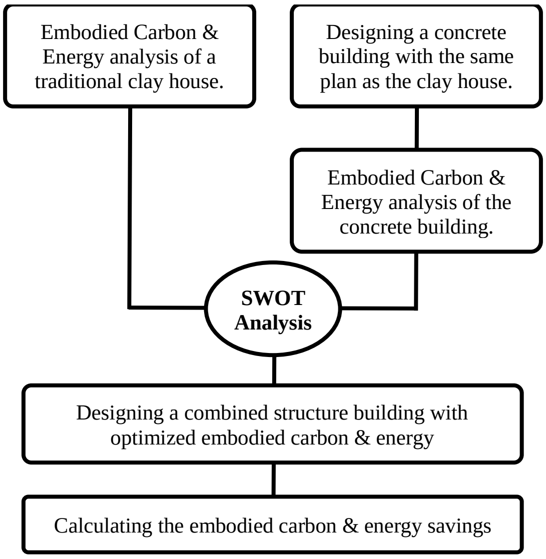

Embodied Carbon & Energy analysis of a traditional clay house.
Designing a concrete building with the same plan as the clay house.
Embodied Carbon & Energy analysis of the concrete building.
SWOT Analysis
Designing a combined structure building with optimized embodied carbon & energy
Calculating the embodied carbon & energy savings

Figure 1. Research process flow chart

Cultural Heritage Organization of Iran. The pictures, plans, and sections of this building are illustrated in Figures 2 to 4. A view of designing the concrete building with ETABS software is indicated in Figure 5. After choosing the case study and designing the concrete building, according to executive details, the types and weights of each material used were obtained. According to the weight amounts obtained, energy and embodied carbon analyses were performed. Analyzing embodied energy and carbon in this paper is based on a model presented by Chen in 2001 [55]. Embodied energy and carbon per mass unit of each material are also taken from a database (inventory of carbon & energy (ICE) Version 2.0) prepared by the University of Bath UK [58] presented in a supplementary file. To calculate the amount of energy and embodied carbon, a program was created using Excel software based on the aforementioned model and database. The concrete building uses materials with a large amount of energy and embodied carbon per mass unit and in the traditional building, due to the high thickness of the clay walls which usually reach 50cm, much more materials have been utilized. Therefore, drawing a comparison of energy and embodied carbon between these two buildings is a challenging task. Thanks to the existing limitations, including the University of Bath database version 2, which published life cycle information based on the cradle-to-gate stage, life cycle analysis in this paper is bound to that stage.

We expanded Chen's model to embodied carbon in which total embodied energy and carbon can be obtained using the following equations:

$$E_e = E_m + E_t + E_p \tag{1}$$

$$C_e = C_m + C_t + C_p \tag{2}$$

where E_e, E_m, E_t, and E_p stands for the total embodied energy, the embodied energy of manufacturing construction materials, the energy consumption of

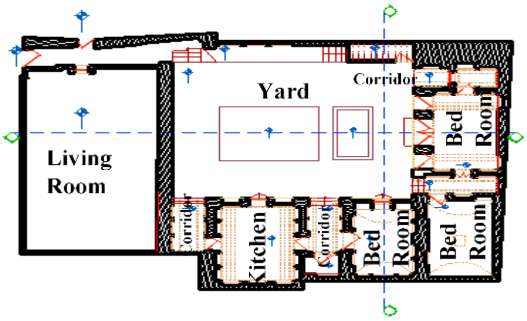


Figure 2. Clay building plan map

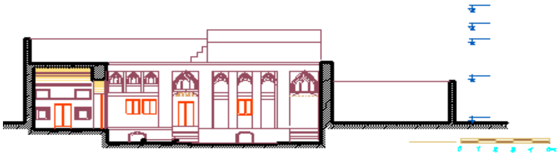


Figure 3. Clay building section

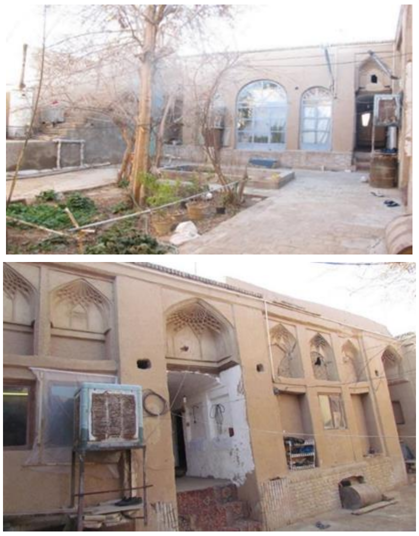


Figure 4. Clay building picture

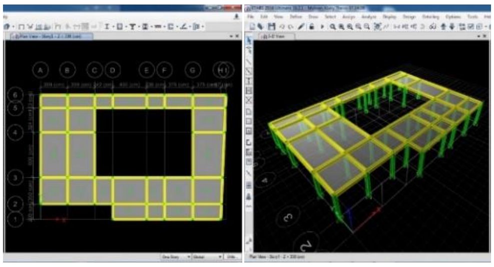


Figure 5. Designing the concrete building using ETABS

transporting building materials from and to the construction site, and embodied energy related to various processes throughout building productions, respectively. Moreover, C_e, C_m, C_t, and C_p represent the total embodied carbon, the value of embodied carbon related to the manufacturing of building materials, the amount of carbon emission of transporting the construction materials and building components, and the value of embodied carbon emanated from different processes, i.e. smoothing of soil and crane lifting, during building productions, respectively. E_m and C_m can be calculated using the following equations:

$$E_m = \sum_{j=1}^{k}(1+\lambda_j)\,\mu_j\left[\sum_{i=1}^{n} q_{i,j}e_{i,j}\right] \tag{3}$$

$$C_m = \sum_{j=1}^{k}(1+\lambda_j)\,\mu_j\left[\sum_{i=1}^{n} q_{i,j}c_{i,j}\right] \tag{4}$$

where k, $e_{i,j}$, $q_{i,j}$, and $c_{i,j}$ represent the number of building materials and elements, the energy required for manufacturing the building materials j in country i in MJ/kg, the amount of building materials j imported from the country i in kg, and the amount of carbon emitted by producing the building materials j in country i in MJ/kg, respectively. Also, n, μ_j, and λ_j denote the number of countries from which building material or element j is imported, the replacement factor for building elements j throughout the whole lifespan of a structure, and the factor for waste materials j produced during the implementation of the structure, respectively. It should be stated that μ_j must be higher than or equal 1, and (μ_j-1) stands for the factor for the recurring embodied energy of building material j. Some building components such as damaged doors and windows might be partially supplanted throughout the buildings' lifespan, while others, such as ceilings, walls, and floor finishes, might be required to be completely replaced every time. The replacement factor can be determined as follows:

$$\mu_j = L_b/l_j \tag{5}$$

The difference between Equations (5) and (6) yields the maintenance factor.

$$\mu_j = \lceil L_b/l_j \rceil \tag{6}$$

where L_b, l_j, are the buildings' lifespan, the mean value of lifespan of building components or materials j, and the mathematical operator that gives the least integer which is equal to or higher than a real number within.

E_t and C_t can be calculated using the following equations:

$$E_t = \sum_{j=1}^{k}(1+\lambda_j)\mu_j Q_j(\bar{e}_{t,j}+e_d) \tag{7}$$

$$C_t = \sum_{j=1}^{k}(1+\lambda_j)\mu_j Q_j(\bar{c}_{t,j}+c_d) \tag{8}$$

where E_t and Q_j are the amounts of energy needed for transportation of the building elements and materials in MJ/(kg. km) and the amount of building material j in kg, respectively. In addition, e_d and c_d indicate the amount of energy consumed and the amount of carbon emitted through demolishing the buildings and transporting the components of demolished buildings from the building site to the landfill, respectively. Subscripts t, $\bar{e}$, and $\bar{c}$ refer to transportation, the mean energy use and carbon emission for transportation of material to the building site in MJ/kg, which might be calculated by:

$$\bar{e}_{t,j} = \sum_{i=1}^{n}\frac{q_{i,j}}{Q_j}\left[\sum_l e_{t,l}d_l\right] \tag{9}$$

$$\bar{c}_{t,j} = \sum_{i=1}^{n}\frac{q_{i,j}}{Q_j}\left[\sum_l c_{t,l}d_l\right] \tag{10}$$

where $e_{t,l}$ and $c_{t,l}$ represent the amount of energy use and carbon emission related to the transportation of building materials by means of conveyance l in MJ/(kg·km). Also, d_l denotes the distance of transportation by the conveyance l in km. The required energy and carbon emitted for different processes throughout demolishing and producing the buildings can be calculated by:

$$E_p = \sum_{i=1}^{k} Q_{p,j}\, e_{p.j} \tag{11}$$

$$C_p = \sum_{i=1}^{k} Q_{p,j}\, c_{p.j} \tag{12}$$

where $Q_{p,j}$ indicates the amount of building material j dealt with in a process throughout demolishing and producing the building in kg, m^3, or MJ/m^2. $c_{p,j,}$ and $e_{p,j}$ stands for the amount of emitted carbon and required energy intensity for this process and building material j in MJ/kg, MJ/m^3, or MJ/m^2 usable floor area. In the next step, the traditional and concrete structural systems were analyzed using the SWOT analysis method, and a solution for optimizing embodied carbon and energy was provided considering the preservation of historical texture. Developing ideas exploring emerging opportunities, and guarding against threats while keeping the weaknesses and strengths of the organization in mind is the goal of this analysis [59]. Finally, by surveying the statistics of clay houses in the country, the effects of implementing the strategy of combining traditional and modern building systems in saving embodied energy and carbon were expressed.

3. RESULTS AND DISCUSSION

To obtain the amount of materials utilized in the building understudy, quantity surveying and estimating of this building was performed and the obtained results are presented in Table 7.

Next, the weights, as well as weight percentages of the materials, are presented from the highest to the lowest volume of materials used in the traditional building in Table 8 and Figure 6. Subsequently, ten materials with the most energy and embodied carbon can be observed in Table 9. The structure of this building has been made of adobe and mud mortar and its lining is made of cob. The soil has been utilized to prepare all of them. Expectedly,

TABLE 7. The outcomes obtained from quantity surveying and estimating of the traditional building

Material	Unit	Quantity
Mud-lime mortar	m^3	3.59
Wood	m^3	3.91
Glass	m^2	9.17
PVC water pipe	m	71
PVC Sewage Pipe	m	40
Power Cable	m	1000
Valves(brass)	kg	10
Mosaic	m^3	126.29
Plaster	m^3	12.28
Sun-dried brick	m^3	592.17
Mud mortar	m^3	252.37
Soil	m^3	145.82
Plaster of clay and straw	m^3	88.76
Lime mortar	m^3	58.22
Brick	n	59361
Soil plaster	m^3	29.1
Cement sand mortar	m^3	19.14

TABLE 8. The mass and percentage of materials used in traditional buildings

Material	Density kg/m^3	Weight kg (max to min)	Weight percentage
Sun-dried brick	1920	1136966.4	47.34890893
Mud mortar	2000	504740	21.01987208
Soil	2000	291640	12.14533323
Plaster of clay and straw	1600	142016	5.914249225
Lime mortar	1850	107707	4.485452634
Brick	1700	100913.7	4.202545995
Soil plaster	1600	46560	1.938988874
Cement sand mortar	2100	40194	1.673877122
Plaster	1300	15964	0.664819982
Mosaic	55 kg/m^2	6945.95	0.289263741
Mud-lime mortar	1300	4667	0.194356982
Wood	650	2541.5	0.105840641
Glass	25 kg/m^2	229.25	0.009547105
PVC plastic	-	134.23	0.005590002
Copper	0.0225 kg/m	22.5	0.000937011
Brass	-	10	0.00041645
Sum	-	2401251.53	100

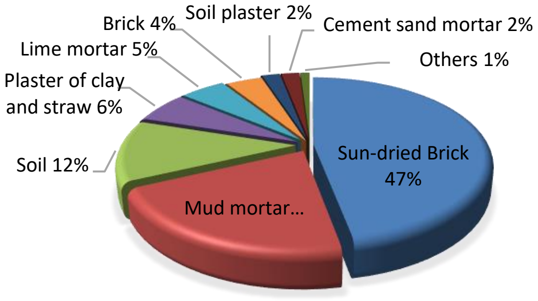


Figure 6. Pie chart of the weight percentage of materials used in traditional building

the soil has the most energy and embodied Carbon in the ranking table. For the accurate comparison between the two types of modern and traditional buildings; the structure and foundation of a one-story concrete house were designed and modeled exactly according to the plan of the traditional house keeping the existing spaces by Etabs and Safe Software with ACI 318-14 regulations.

TABLE 9. The rankings of the materials in terms of energy embodied carbon and equivalent carbon for the traditional building

Rank	Traditional building					
	Material	EE - MJ	Material	EC - kgCO2	Material	EC - kgCO2e
1	Soil	839852.6976	Soil	42925.80454	Soil	44792.14387
2	Brick	329863.3848	Brick	25289.52617	Brick	26389.07079
3	Plaster	90771.2423	Lime	11955.7044	Lime	12270.3282
4	Lime	83375.307	Cement	6468.82236	Cement	6911.892385
5	Wood	74909.86533	Plaster	5790.903386	Plaster	6334.264981
6	Cement	42977.79239	Wood	4915.959913	Wood	5009.597244
7	Sand	21610.29458	Sand	1280.610049	Sand	1360.648177
8	Mosaic	15611.18926	Mosaic	1199.238813	Mosaic	1271.193142
9	Plastic	10784.10075	Plastic	364.5920073	Plastic	433.0403152
10	Glass	5920.067209	Glass	339.4171866	Glass	359.150744

Then, it was analyzed concerning the amount and type of consumed materials. The dimensions of the cross-section of the columns and beams in the design were 30*30 cm and 35*30 cm, respectively. The roof type was selected as a joist system. The foundation of this building was designed as a strip footing with a width of 1 meter, and to control the punching shear, the depth of the foundation was designed to be 90 cm. Concerning the details obtained, the quantity surveying and estimating of the concrete building was carried out and the results are summarized in Table 10. Next, the weights and weight percentages of the materials are presented from the highest to the lowest volume of materials used in the concrete building in Table 11 and Figure 7.

TABLE 10. Results obtained from the concrete building quantity surveying and estimating

Material	Unit	Quantity
Concrete C25	m^3	206.04
Rock	m^3	148.35
Concrete C20	m^3	85.44
Cement Sand Mortar	m^3	61.84
Brick	m^3	76.30
Soil	m^3	42.39
Cement Block	n	2167.31
Soil Plaster	m^3	15.39
Deformed Bar	kg	21289.56
Granite	m^2	620.14
Clinker	m^3	28.02
Plaster	m^3	7.71
Bituminous Felt	m^2	377.47
Mosaic	m^2	88.48
Ceramic	m^2	230
Paint	m^3	0.0385
Tile	m^2	110.30
Wood	m^3	3.05
Glass	m^2	9.17
PVC Plastic	kg	134.23
Power Cable	m	1000
Valves (Brass)	kg	10
Nylon	m^2	83.15

Ten materials with the highest amounts of energy and embodied carbon in the concrete building are observed in Table 12.

Since the skeleton and the foundation of the building are made of reinforced concrete and its walls are made of pressed bricks, steel, concrete, and brick are at the top of the ranking table with the most energy and embodied carbon. Based on the data collected, the total equivalent of carbon and embodied energy, as well as the weight of the materials used, including the wastes, were compared in both traditional and concrete buildings in Table 13. As shown in Table 13, the weight of a concrete building is 1480 tons and the weight of its adobe counterpart is 2488 tons, which is about 1.7 times heavier. It is due to the high thickness of the adobe building walls, which sometimes reach 50 cm, and the use of materials with more mass in the adobe building as well. However, the results of carbon and embodied energy analysis show an increase of 1.73 times in the embodied energy, 2.28 times in the embodied carbon, and 2.33 times in the equivalent embodied carbon, with the change of system structure from traditional to concrete.

In other words, despite the lower mass of materials used in concrete buildings, the amount of carbon and embodied energy is markedly more compared to adobe buildings. It is because of using materials with energy, embodied carbon, and more units of mass.

As a result, establishing conventional concrete buildings instead of adobe buildings leads to increased carbon and embodied energy and a loss of historical context. It is also known that traditional houses such as ordinary rural buildings, need seismic retrofitting to become more resistant to earthquakes. This seismic retrofitting needs to be done in the context of sustainable development so as not to increase carbon emissions and energy consumption indiscriminately. Therefore, to choose the optimal method of strengthening these historical monuments, the right decision needs to be made, and in this regard, the SWOT technique will be employed. SWOT analysis is a systematic analysis seeking to provide a list of capabilities, weaknesses, opportunities, and threats, so the organizations can use these findings to find a strategy that fits their situation. From this model's viewpoint, a proper strategy maximizes the strengths and opportunities and minimizes weaknesses and threats.

TABLE 11. The mass and percentage of materials utilized in the concrete building

Material	Density (kg/m^3)	kg (max to min)	Weight percent	Material	Density (kg/m^3)	kg (max to min)	Weight percent
Concrete C25	2400	494496	35.5553378	Plaster	1300	10023	0.720675497
Rock	1400	207690	14.93336267	Bituminous Felt	15 kg/m^2	5662.05	0.407113709

Concrete C20	2390	204201.6	14.68253913	Mosaic	55 kg/m^2	4866.4	0.349904743
Cement Sand Mortar	2100	129864	9.337504021	Ceramic	21 kg/m^2	4830	0.347287504
Brick	1700	129710	9.326431086	Tile	20 kg/m^2	2206	0.158616197
Soil	2000	84780	6.095866375	Wood	650	1982.5	0.142546061
Cement Block	13.25 kg/n	28716.8575	2.064804506	Glass	25 kg/m^2	229.25	0.016483574
Soil Plaster	1600	24624	1.770519151	Plastic	-	134.23	0.009651429
Deformed Bar	7850	21289.56	1.530765663	Paint	1310	50.43	0.003632833
Granite	2800	17363.92	1.24850361	Copper	0.0225 kg/m	22.5	0.001617799
Clinker	550	15411	1.108084415	Brass	-	10	0.000719022
				Nylon	0.11 kg/m^2	9.15	0.000657905
				Sum	-	1388172.44	100

TABLE 12. The ranking of materials in terms of energy and embodied carbon and the equivalent carbon for concrete building

Rank	Concrete building					
	Material	**EE - MJ**	**Material**	**EC - kgCO2**	**Material**	**EC - kgCO2e**
1	Steel	661722.1039	Concrete	64892.07616	Concrete	69523.6485
2	Concrete	489895.8448	Steel	58693.84415	Steel	62772.95301
3	Brick	423991.7836	Brick	32506.03674	Brick	33919.34268
4	Ceramic	324901.8109	Cement Mortar	20900.31216	Cement Mortar	22331.84039
5	Stone	209318.4877	Ceramic	20035.61167	Ceramic	21118.61771
6	Bituminous Felt	180884.9125	Stone	12178.53019	Stone	13320.2674
7	Cement Mortar	138858.2383	Bituminous Felt	9415.145287	Bituminous Felt	9980.054004
8	Plaster	51188.74447	Plaster	3274.806625	Plaster	3579.854989
9	Wood	48075.94458	Wood	2529.644845	Wood	3182.71135
10	Soil	39295.53	Soil	2008.4382	Soil	2095.7616

TABLE 13. The total equivalent of embodied energy and carbon as well as the weight of materials used, including the construction waste

Comparison Criteria	Traditional Building		Concrete Building	
	Total	**per sqm**	**Total**	**per sqm**
EE - MJ	1521527.328	3972.656	2634913.490	6879.670
EC - kgCO$_2$	100888.823	263.417	229776.491	599.939
EC - kgCO$_2$e	105514.025	275.494	245546.029	641.112
Material weight including waste (kg)	2488035.475	6496.18	1480248.721	3864.88

Table 14 shows the SWOT analysis for concrete buildings with ordinary materials. Table 15 shows the SWOT analysis for traditional buildings with adobe materials. Therefore, according to the SWOT table and the proposed strategies to preserve historical texture and tourists attraction, and to show the effect of substituting materials on the amount of energy and embodied carbon, the energy and carbon were analyzed considering a combined building with concrete structure, frame and the joist system, the adobe walls, and the traditional flooring. Based on the analyzed information, the amount of embodied energy and carbon equivalent to total as well as the weight of materials used, including construction wastes in this combined building, and its difference from the conventional concrete building are presented in Table 16. The case-by-case comparison of saved weights, energy, and embodied carbon was performed and the percentage of savings for each material is separately shown in Table 17.

TABLE 14. SWOT analysis for concrete structures with the common materials

	Strengths (S)	**Weaknesses (W)**
SWOT analysis for concrete structures with the common materials	• High seismic retrofitting • High safety and security • High durability • Low maintenance costs	• High embodied energy (1.73 times higher than that of a traditional building according to the analysis) • A great amount of embodied carbon (2.28 times greater than that of a traditional building according to the analysis) • A great amount of construction waste • Lack of originality and non-observance of the tradition of Islamic Iranian architecture in such buildings
Opportunities (O)	**Strategies (SO)**	**Strategies (WO)**
• Familiarizing the executives and engineers with such materials, and how to implement them • The number of experts familiar with this structural system • Providing equipment for the implementation of such structures • Possessing a bylaw to match design issues	• Using concrete structures in dilapidated and historical textures to diminish financial as well as human losses • The combined use of concrete structures so as to maintain and renovate historical structures for greater durability	• Combining the technology of concrete frame construction with traditional facade rather than stone, traditional plan, and materials by the native architecture of each area • Creating new job opportunities by investing in line with the development of regulations, embodied energy, and carbon standards and monitoring their compliance • Informing engineers and project managers about the issue of embodied energy and compliance with standards
Threats (T)	**Strategies (ST)**	**Strategies (WT)**
• Not using indigenous materials • Non-compliance with environmental issues as well as the sustainable development model • Lack of attraction for tourists • The difficulty of construction waste recycling these materials	• Raising the awareness of the community, bringing the culture and a sense of trust in meeting energy and embodied carbon standards, and preserving the environment • Providing government funding, attracting private investment, and allocating funds to implement energy and carbon regulations • Development and the attraction of tourists by combining modern and traditional structures with Iranian architecture and preserving the historical texture in line with sustainable urban development	• Labeling all building materials in terms of energy and embodied carbon in factories for designer use • Replacing the common materials with the indigenous ones with less energy and carbon per unit of mass and easier recycling capability

TABLE 15. SWOT table for traditional structures with adobe materials

	Strengths (S)	**Weaknesses (W)**
SWOT analysis for traditional structures with adobe materials	• Low embodied energy (according to the analysis performed) • Low embodied carbon (according to the analysis performed) • Low construction waste and adaptation to climate • Possessing the originality and observance of the tradition of Iranian Islamic architecture in such buildings	• Low seismic retrofitting, high casualties during the natural catastrophe, and low safety Foundation • heavy roof • Lack of dry walls, lack of integrity, long and uncontrolled lengths, and long walls
Opportunities (O)	**Strategies (SO)**	**Strategies (WO)**
• Inexpensive and available indigenous materials Attract tourists	• Preserving historical and traditional textures and restoring them to preserve the originality of Iranian architecture and attract tourists	• Fixing major weaknesses in the structural system by combining concrete frame

- Materials are easily recycled - The necessity of considering environmental issues, following the model of sustainable development, and high executive potential	- Using indigenous materials due to the ease of access and coordination with the rural economy	construction technology with traditional adobe and finishing - Reduced construction waste in the construction sector using traditional architecture and materials in line with sustainable development
Threats (T)	**Strategies (ST)**	**Strategies (WT)**
- Lack of skilled experts familiar with this structural system - Modernism and forgetting the originality of architecture - No regulations to match the design issues of these structures - Ignorance from officials and the media regarding the culture of sustainable energy development and environmental issues	- Culturalization of preserving the originality of Iranian architecture and creating a sense of trust in this style of architecture by using modern technologies and standards - Training experts to use materials compatible with the climate of each region to create thermal and cooling insulation to save energy	- Encouraging engineers and allocating funds for research on seismic retrofitting of adobe buildings and improving the quality of rural life - Making regulations and implementing a plan to improve traditional buildings and reduce casualties due to natural catastrophes due to the impossibility of removing this system in rural areas that are far from facilities

TABLE 16. Energy and embodied carbon equivalent to total and the weight of materials used in the combined building compared to the conventional concrete building

Comparison Criteria	Combined Building		Combined building savings compare to concrete building		
	Total	per sqm	Total	per sqm	percentage
EE - MJ	1705856.092	4453.932	929057.398	2425.7	35.26%
EC - kgCO2	139289.164	363.679	90487.327	236.3	39.38%
Material weight including waste (kg)	2133260.538	5569.87	-653011.817	-1705.0	-44.12%

TABLE 17. Saved weight, energy, and embodied carbon in materials

Material	The amount of savings			The percentage of savings		
	Weight (kg)	Embodied Energy (MJ)	Embodied Carbon ($kgCO_2$)	Weight	Embodied Energy	Embodied Carbon
Concrete	63083.2	39742.4	5299.0	8.16%	8.11%	8.17%
Steel	1055.1	285767.3	30389.2	4.66%	43.19%	51.78%
Brick	133439.79	400319.37	30691.15	94.42%	94.42%	94.42%
Coating	13731.8	14456.4	887.8	35.42%	28.24%	27.11%
Ceramics and Tiles	6367.82	272906.44	16829.23	84.00%	84.00%	84.00%
Paint	53.06	18570.50	642.009	100.00%	100.00%	100.00%
Mosaic	-8262.49	-14459.4	-1218.7	0.00%	0.00%	0.00%
Stone	19028.95	209318.49	12178.53	100.00%	100.00%	100.00%
Mortar	112642.57	109263.29	16445.82	78.69%	78.69%	78.69%

Correspondingly, if the project is divided into three sections: skeleton frame, framework, and finishing, the skeleton frame includes concrete, steel, polystyrene, and aggregate; the framework includes pressed and clay brick, lining, plastics other than polystyrene, mortar except for slurry and bitumen, and the finishing includes cement-slurry, ceramic and tile, aluminum, paint, mosaic, stone, glass, and asphalt. Energy percentage and the embodied carbon and the weight of the materials in each of the work sections in the concrete building and the amount of savings in each section are presented in Table 18 and the diagrams in Figures 8 to 10.

TABLE 18. Saved weight, energy, and embodied carbon for each work section separately

Work sections	Concrete Building			The amount of savings			Percentage of weight saved
	Weight 1000ton	Embodied Energy 1000GJ	Embodied Carbon 1000 ton CO2	Weight 1000 ton	Embodied Energy 1000 GJ	Embodied Carbon 1000ton CO2	
Skeleton frame	0.96	1.18	0.125	0.06	0.33	0.04	6.71%
Framework	0.29	0.62	0.057	0.26	0.52	0.05	88.30%
Finishing	0.03	0.56	0.033	0.02	0.49	0.03	69.64%

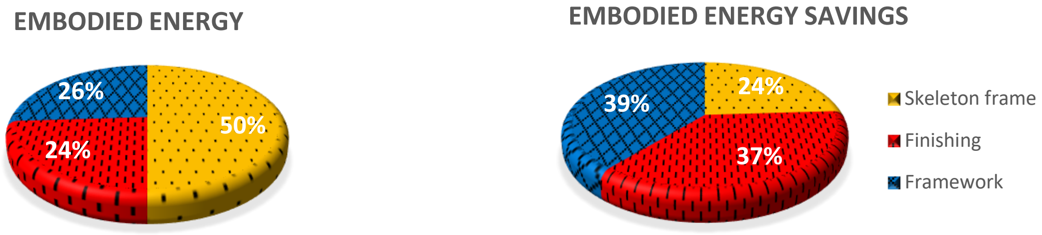


Figure 8. The embodied energy and its saving amount for each work section separately

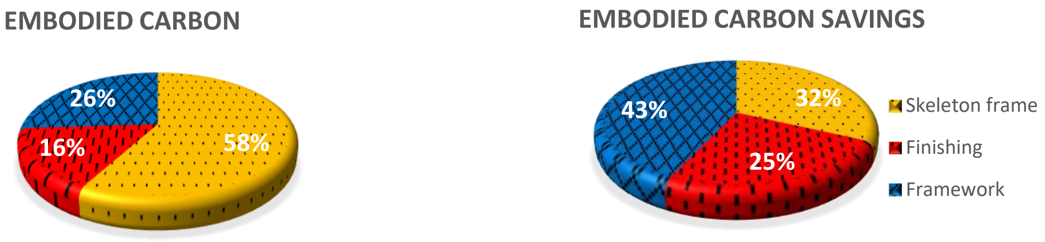


Figure 9. The embodied carbon and its saving amount for each work section separately

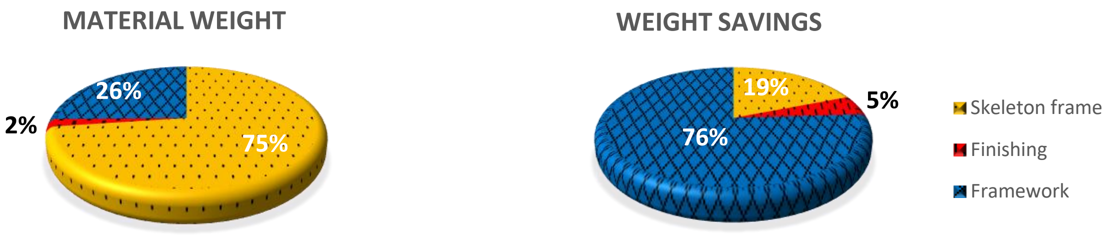


Figure 10. Material weight and its saved mount for each work section separately

4. CONCLUSION

Over time, the structural system of buildings has changed. This change caused an increase in energy and carbon dioxide. However, research in Iran has focused less on the necessity and importance of reforming this process. In this respect, a Microsoft Excel program was provided to determine embodied energy and carbon for all types of buildings, the results of which were validated in the selected case sample by manual calculations.

According to the strategy explained for historical and derelict buildings in Yazd and after SWOT analysis, the skeleton frame and roof of the concrete structure were combined with a framework and finishing to achieve the goals of reducing embodied energy and carbon (by removing and replacing materials such as bricks). Preserving the texture of the area is in line with sustainable urban development and maintaining the attractiveness of this texture for tourists. Because the frames of this combined structure are made of concrete, there is no need to implement thick load-bearing walls, moreover, the wall thickness is reduced to a minimum of 25 cm (a row of adobe considering the thickness of finishing with plaster of clay and straw).

According to Table 16, the weight of materials employed in the combined building has finally decreased by 14% compared to the traditional building. Observing the results of embodied energy and carbon analysis for the modified building indicates 39.38% savings in carbon and 35.26% in embodied energy. Consistent with the latest census of the Statistics Center of Iran in 2016, the number of residential units in which adobe is used is 10.54% of the total rural houses. It reveals that about 53.73 million square meters go to adobe houses. Considering the amount of energy and carbon saved in the combined plan, it is possible to reduce 13.66 million tons of carbon equivalent to 1.96 million tons of energy and store 130.34 million gigajoules of energy by developing this plan for the adobe texture in the country.

To show the effect of implementing the results of the current research project, it can be pointed out that the amount of energy saved in the proposed strategy is equivalent to the production of electric energy from Iran's largest power plant ‘Damavand Power Plant’ (Pakdasht Martyrs) with a capacity of 2868 MW in 2 years and 4 months. To develop the present study, the following suggestions are presented to researchers who intend to conduct additional research in this field:

- Investigating the effect of increasing the lifespan of building and durability of materials to reduce the replacement coefficient to save embodied energy and carbon.
- Calculate the amount of embodied carbon and energy of wooden houses and examine the possibility of replacing this structural system with current systems according to the climate.
- Investigating the role of advanced technologies in embodied energy and carbon optimization.

Comparison of steel and concrete structures concerning embodied energy and carbon.

5. REFERENCES

1. WCED, S. W. S. “World commission on environment and development.” ***Our Common Future***, Vol. 17, (1987), 1-91.
2. Mohamad, M. I., Nekooie, M. A., Ismail, Z. B., and Taherkhani, R. Amphibious urbanization as a sustainable flood mitigation strategy in south-east Asia. ***Advanced Materials Research,*** Vol. 622-623, (2013). https://doi.org/10.4028/www.scientific.net/AMR.622-623.1696
3. Taherkhani, R. *Development of a Social Sustainability Model in Industrial Building System*. Doctoral dissertation, Universiti Teknologi Malaysia, Johor Bahru, Malaysia 2013.
4. Taherkhani, R., Saleh, A. L., Mansur, S. A., Nekooie, M. A., Noushiravan, M., and Hamdani, M. “A Systematic Research Gap Finding Framework: Case Study of Construction Management.” ***Journal of Basic and Applied Scientific Research***, Vol. 2, No. 5, (2012), 5129-5136.
5. Taherkhani, R. “A Strategy towards Sustainable Industrial Building Systems (IBS): The Case of Malaysia.” ***Journal of Multidisciplinary Engineering Science and Technology***, Vol. 1, No. 4, (2014), 86-90. Retrieved from http://www.jmest.org/wp-content/uploads/JMESTN42350083.pdf
6. Taherkhani, R. “An integrated social sustainability assessment framework: the case of construction industry.” ***Open House International***, (2022). https://doi.org/10.1108/OHI-04-2022-0098
7. Oke, A., Aghimien, D., Aigbavboa, C., and Musenga, C. “Drivers of Sustainable Construction Practices in the Zambian Construction Industry.” ***Energy Procedia***, Vol. 158, (2019), 3246-3252. https://doi.org/10.1016/j.egypro.2019.01.995
8. Zimmermann, R. K., Skjelmose, O., Jensen, K. G., Jensen, K. K., and Birgisdottir, H. “Categorizing Building certification systems according to the definition of sustainable building.” ***IOP Conference Series: Materials Science and Engineering***, Vol. 471, No. 9, (2019), 92060. https://doi.org/10.1088/1757-899x/471/9/092060.
9. Rozana, Z., Khalid Ahmed, M., Zin, R. M., Zolfagharian, S., Nourbakhsh, M., Nekooie, M. A., and Taherkhani, R. “Sustainable Development Factors for Land Development in Universiti Teknologi Malaysia’s Campus.” ***OIDA International Journal of Sustainable Development***, Vol. 3, No. 9, (2012), 105-110.
10. Taherkhani, R., Hashempour, N., Shaahnazari, S., and Taherkhani, F. “Sustainable cities through the right selection of vegetation types for green roofs.” ***International Journal of Sustainable Building Technology and Urban Development***, Vol. 13, No. 3, (2022), 365-388. https://doi.org/10.22712/susb.20220027
11. Tabatabaee, S., and Weil, B. S. “Definition and Frameworks on a Life-Cycle Negative Growth Rate for Energy and Carbon in an Academic Campus.” In ***Handbook of Theory and Practice of Sustainable Development in Higher Education,*** 325-339, Springer. https://doi.org/10.1007/978-3-319-47877-7_22
12. Ma, J.-J., Du, G., Zhang, Z.-K., Wang, P.-X., and Xie, B.-C. “Life cycle analysis of energy consumption and CO2 emissions from a typical large office building in Tianjin, China.” ***Building and Environment***, Vol. 117, (2017), 36-48. https://doi.org/10.1016/j.buildenv.2017.03.005
13. Roohollah Taherkhani, Najme Hashempour, and Mitra Lot. “Sustainable-resilient urban revitalization framework: Residential buildings renovation in a historic district.” ***Journal of Cleaner Production***, Vol. 286, (2021), 124952. https://doi.org/DOI: 10.1016/j.jclepro.2020.124952
14. Farese, P. “How to build a low-energy future.” ***Nature***, Vol. 488, No. 7411, (2012), 275-277. https://doi.org/10.1038/488275a
15. Hashempour, N., Taherkhani, R., and Mahdikhani, M. “Energy performance optimization of existing buildings: A literature review.” ***Sustainable Cities and Society***, Vol. 54, (2020), 101967. https://doi.org/10.1016/j.scs.2019.101967

16. Aram, K., Taherkhani, R., and Šimelytė, A. "Multistage Optimization toward a Nearly Net Zero Energy Building Due to Climate Change." ***Energies***, Vol. 15, No. 3, (2022), 983. https://doi.org/10.3390/EN15030983

17. Weiler, V., Harter, H., and Eicker, U. "Life cycle assessment of buildings and city quarters comparing demolition and reconstruction with refurbishment." ***Energy and Buildings***, Vol. 134, (2017), 319-328. https://doi.org/10.1016/j.enbuild.2016.11.004

18. Copiello, S. "Economic implications of the energy issue: Evidence for a positive non-linear relation between embodied energy and construction cost." ***Energy and Buildings***, Vol. 123, , (2016), 59-70. https://doi.org/10.1016/j.enbuild.2016.04.054

19. Stephan, A., and Stephan, L. "Life cycle energy and cost analysis of embodied, operational and user-transport energy reduction measures for residential buildings." ***Applied Energy***, Vol. 161, , (2016), 445-464. https://doi.org/10.1016/j.apenergy.2015.10.023

20. Crishna, N., Banfill, P. F. G., and Goodsir, S. "Embodied energy and CO2 in UK dimension stone." ***Resources, Conservation and Recycling***, Vol. 55, No. 12, (2011), 1265-1273. https://doi.org/10.1016/j.resconrec.2011.06.014

21. Praseeda, K. I., Reddy, B. V. V., and Mani, M. "Embodied and operational energy of urban residential buildings in India." ***Energy and Buildings***, Vol. 110, (2016), 211-219. https://doi.org/10.1016/j.enbuild.2015.09.072

22. Cabeza, L. F., Barreneche, C., Miró, L., Morera, J. M., Bartolí, E., and Fernández, A. I. "Low carbon and low embodied energy materials in buildings: A review." ***Renewable and Sustainable Energy Reviews***, Vol. 1, No. 23, (2013), 536-542. https://doi.org/10.1016/j.rser.2013.03.017

23. Dixit, M. K., Fernández-Solís, J. L., Lavy, S., and Culp, C. H. "Need for an embodied energy measurement protocol for buildings: A review paper." ***Renewable and Sustainable Energy Reviews***, Vol. 16, No. 6, (2012), 3730-3743. https://doi.org/10.1016/j.rser.2012.03.021

24. Buchanan, A. H., and Honey, B. G. "Energy and carbon dioxide implications of building construction." ***Energy and Buildings***, Vol. 20, No. 3, (1994), 205-217. https://doi.org/10.1016/0378-7788(94)90024-8

25. Gustavsson, L., and Joelsson, A. "Life cycle primary energy analysis of residential buildings." ***Energy and Buildings***, Vol. 42, No. 2, (2010), 210-220. https://doi.org/10.1016/j.enbuild.2009.08.017

26. Kofoworola, O. F., and Gheewala, S. H. "Life cycle energy assessment of a typical office building in Thailand." ***Energy and Buildings***, Vol. 41, No. 10, (2009), 1076-1083. https://doi.org/10.1016/j.enbuild.2009.06.002

27. Kayaçetin, N. C., and Tanyer, A. M. "Embodied carbon assessment of residential housing at urban scale." ***Renewable and Sustainable Energy Reviews***, Vol. 117, (2020), 109470. https://doi.org/10.1016/j.rser.2019.109470

28. Sun, C., Chen, L., and Xu, Y. "Industrial linkage of embodied CO2 emissions: Evidence based on an absolute weighted measurement method." ***Resources, Conservation and Recycling***, Vol. 160, (2020), 104892. https://doi.org/10.1016/j.resconrec.2020.104892

29. Shukla, A., Tiwari, G. N., and Sodha, M. S. "Embodied energy analysis of adobe house." ***Renewable Energy***, Vol. 34, No. 3, (2009), 755-761. https://doi.org/10.1016/j.renene.2008.04.002

30. Ortiz-Rodríguez, O., Castells, F., and Sonnemann, G. "Life cycle assessment of two dwellings: One in Spain, a developed country, and one in Colombia, a country under development." ***Science of the Total Environment***, Vol. 408, No. 12, (2010), 2435-2443. https://doi.org/10.1016/j.scitotenv.2010.02.021

31. Monteiro, H., and Freire, F. "Environmental life-cycle impacts of a single-family house in Portugal: assessing alternative exterior walls with two methods." ***Gazi University Journal of Science***, Vol. 24, No. 3, (2011), 527-534.

32. Atmaca, A., and Atmaca, N. "Life cycle energy (LCEA) and carbon dioxide emissions (LCCO2A) assessment of two residential buildings in Gaziantep, Turkey." ***Energy and Buildings***, Vol. 102, (2015), 417-431. https://doi.org/10.1016/j.enbuild.2015.06.008

33. Asif, M., Muneer, T., and Kelley, R. "Life cycle assessment: A case study of a dwelling home in Scotland." ***Building and Environment***, Vol. 42, (2007), 1391-1394. https://doi.org/10.1016/j.buildenv.2005.11.023

34. Monahan, J., and Powell, J. C. "An embodied carbon and energy analysis of modern methods of construction in housing: A case study using a lifecycle assessment framework." ***Energy and Buildings***, Vol. 43, No. 1, (2011), 179-188. https://doi.org/10.1016/j.enbuild.2010.09.005

35. Luo, Z., Yang, L., and Liu, J. "Embodied carbon emissions of office building: a case study of China's 78 office buildings." ***Building and Environment***, Vol. 95, (2016), 365-371. https://doi.org/10.1016/j.buildenv.2015.09.018

36. Gan, V. J. L., Cheng, J. C. P., Lo, I. M. C., and Chan, C. M. "Developing a CO2-e accounting method for quantification and analysis of embodied carbon in high-rise buildings." ***Journal of Cleaner Production***, Vol. 141, (2017), 825-836. https://doi.org/10.1016/j.jclepro.2016.09.126

37. Teng, Y., and Pan, W. "Systematic embodied carbon assessment and reduction of prefabricated high-rise public residential buildings in Hong Kong." ***Journal of Cleaner Production***, Vol. 238, (2019), 117791. https://doi.org/10.1016/j.jclepro.2019.117791

38. Finnveden, G., Hauschild, M. Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., and Suh, S. "Recent developments in Life Cycle Assessment." ***Journal of Environmental Management***, Vol. 91, No. 1, (2009), 1-21. https://doi.org/10.1016/j.jenvman.2009.06.018

39. Hauschild, M. Z., Rosenbaum, R. K., and Olsen, S. I. Life Cycle Assessment. New York City, USA: Springer Cham. https://doi.org/10.1007/978-3-319-56475-3

40. Curran, M. A. "Life cycle assessment: a review of the methodology and its application to sustainability." ***Current Opinion in Chemical Engineering***, Vol. 2, No. 3, (2013), 273-277. https://doi.org/10.1016/j.coche.2013.02.002

41. Schlanbusch, R. D., Fufa, S. M., Häkkinen, T., Vares, S., Birgisdottir, H., and Ylmén, P. "Experiences with LCA in the Nordic building industry–challenges, needs and solutions." ***Energy Procedia***, Vol. 96, (2016), 82-93. https://doi.org/10.1016/j.egypro.2016.09.106

42. Petrovic, B., Myhren, J. A., Zhang, X., Wallhagen, M., and Eriksson, O. "Life Cycle Assessment of Building Materials for a Single-family House in Sweden." ***Energy Procedia***, Vol. 158, (2019), 3547-3552. https://doi.org/10.1016/j.egypro.2019.01.913

43. RICS. Methodology to calculate embodied carbon of materials. Coventry, UK: Royal Institution of Chartered Surveyors (RICS) *Information paper*, Vol. 32, (2012), 2012.

44. Robati, M., Daly, D., and Kokogiannakis, G. "A method of uncertainty analysis for whole-life embodied carbon emissions (CO_2-e) of building materials of a net-zero energy building in Australia." ***Journal of Cleaner Production***, Vol. 225, (2019), 541-553. https://doi.org/10.1016/j.jclepro.2019.03.339

45. Cabeza, L. F., Rincón, L., Vilariño, V., Pérez, G., and Castell, A. "Life cycle assessment (LCA) and life cycle energy analysis (LCEA) of buildings and the building sector: A review." ***Renewable and Sustainable Energy Reviews***, Vol. 29, (2014), 394-416. https://doi.org/10.1016/j.rser.2013.08.037

46. Tecchio, P., Gregory, J., Olivetti, E., Ghattas, R., and Kirchain, R. "Streamlining the Life Cycle Assessment of Buildings by

Structured Under-Specification and Probabilistic Triage." ***Journal of Industrial Ecology***, Vol. 23, No. 1, (2019), 268-279. https://doi.org/10.1111/jiec.12731

47. Solís-Guzmán, J., Marrero, M., Montes-Delgado, M. V., and Ramírez-de-Arellano, A. "A Spanish model for quantification and management of construction waste." ***Waste Management (New York, N.Y.)***, Vol. 29, No. 9, (2009), 2542-2548. https://doi.org/10.1016/j.wasman.2009.05.009

48. Barbudo, A., Ayuso, J., Lozano, A., Cabrera, M., and López-Uceda, A. "Recommendations for the management of construction and demolition waste in treatment plants." ***Environmental Science and Pollution Research***, Vol. 27, No. 1, (2020), 125-132. https://doi.org/10.1007/s11356-019-05578-0

49. Ajayi, S. O., and Oyedele, L. O. "Policy imperatives for diverting construction waste from landfill: Experts' recommendations for UK policy expansion." ***Journal of Cleaner Production***, Vol. 147, (2017), 57-65. https://doi.org/10.1016/j.jclepro.2017.01.075

50. Menegaki, M., and Damigos, D. "A review on current situation and challenges of construction and demolition waste management." ***Current Opinion in Green and Sustainable Chemistry***, Vol. 13, (2018), 8-15. https://doi.org/10.1016/j.cogsc.2018.02.010

51. Nikmehr, B., Hosseini, M. R., Rameezdeen, R., Chileshe, N., Ghoddousi, P., and Arashpour, M. "An integrated model for factors affecting construction and demolition waste management in Iran." ***Engineering, Construction and Architectural Management***, Vol. 24, No. 6, (2017), 1246-1268. https://doi.org/10.1108/ECAM-01-2016-0015

52. Saghafi, M. D., and Teshnizi, Z. A. H. "Building Deconstruction and Material Recovery in Iran: An Analysis of Major Determinants." ***Procedia Engineering***, Vol. 21, (2011), 853-863. https://doi.org/10.1016/j.proeng.2011.11.2087

53. *Guidance on the legal definition of waste and its application*. London, UK.

54. Adalberth, K. "Energy use during the life cycle of single-unit dwellings: examples." ***Building and Environment***, Vol. 32, No. 4, (1997), 321-329. https://doi.org/10.1016/S0360-1323(96)00069-8

55. Chen, T. Y., Burnett, J., and Chau, C. K. "Analysis of embodied energy use in the residential building of Hong Kong." ***Energy***, Vol. 26, No. 4, (2001), 323-340. https://doi.org/10.1016/S0360-5442(01)00006-8

56. Mousavi, B., Lopez, N. S. A., Biona, J. B. M., Chiu, A. S. F., and Blesl, M. "Driving forces of Iran's CO_2 emissions from energy consumption: An LMDI decomposition approach." ***Applied Energy***, Vol. 206, (2017), 804-814. https://doi.org/10.1016/j.apenergy.2017.08.199

57. Dong, C., Dong, X., Jiang, Q., Dong, K., and Liu, G. "What is the probability of achieving the carbon dioxide emission targets of the Paris Agreement? Evidence from the top ten emitters." ***Science of the Total Environment***, Vol. 622, (2018), 1294-1303. https://doi.org/10.1016/j.scitotenv.2017.12.093

58. Hammond, G., and Jones, C. "Inventory of Carbon & Energy (ICE) Version 2.0, Sustainable Energy Research Team (SERT), Department of Mechanical Engineering, University of Bath UK."

59. Muralidharan, K. "Six Sigma Project Management." In ***Six Sigma for Organizational Excellence,*** 19-37, Springer.

60. Debnath, A., Singh, S. V, and Singh, Y. P. "Comparative assessment of energy requirements for different types of residential buildings in India." ***Energy and Buildings***, Vol. 23, No. 2, (1995), 141-146. https://doi.org/10.1016/0378-7788(95)00939-6

61. Suzuki, M., Oka, T., and Okada, K. "The estimation of energy consumption and CO_2 emission due to housing construction in Japan." ***Energy and Buildings***, Vol. 22, No. 2, (1995), 165-169. https://doi.org/10.1016/0378-7788(95)00914-J

62. Thormark, C. "A low energy building in a life cycle—its embodied energy, energy need for operation and recycling potential." ***Building and Environment***, Vol. 37, No. 4, (2002), 429-435. https://doi.org/10.1016/S0360-1323(01)00033-6

63. Mithraratne, N., and Vale, B. "Life cycle analysis model for New Zealand houses." ***Building and Environment***, Vol. 39, No. 4, (2004), 483-492. https://doi.org/10.1016/j.buildenv.2003.09.008

64. Casals, X. G. "Analysis of building energy regulation and certification in Europe: Their role, limitations and differences." ***Energy and Buildings***, Vol. 38, No. 5, (2006), 381-392. https://doi.org/10.1016/j.enbuild.2005.05.004

65. Nässén, J., Holmberg, J., Wadeskog, A., and Nyman, M. "Direct and indirect energy use and carbon emissions in the production phase of buildings: an input-output analysis." ***Energy***, Vol. 32, No. 9, (2007), 1593-1602. https://doi.org/10.1016/j.energy.2007.01.002

66. Huberman, N., and Pearlmutter, D. "A life-cycle energy analysis of building materials in the Negev desert." ***Energy and Buildings***, Vol. 40, No. 5, (2008), 837-848. https://doi.org/10.1016/j.enbuild.2007.06.002

67. Utama, A., and Gheewala, S. H. "Indonesian residential high rise buildings: A life cycle energy assessment." ***Energy and Buildings***, Vol. 41, No. 11, (2009), 1263-1268. https://doi.org/10.1016/j.enbuild.2009.07.025

68. Blengini, G. A., and Di Carlo, T. "The changing role of life cycle phases, subsystems and materials in the LCA of low energy buildings." ***Energy and Buildings***, Vol. 42, No. 6, (2010), 869-880. https://doi.org/10.1016/j.enbuild.2009.12.009

69. Ramesh, T., Prakash, R., and Shukla, K. K. "Life cycle energy analysis of buildings: An overview." ***Energy and Buildings***, Vol. 42, No. 10, (2010), 1592-1600. https://doi.org/10.1016/j.enbuild.2010.05.007

70. Gustavsson, L., Joelsson, A., and Sathre, R. "Life cycle primary energy use and carbon emission of an eight-storey wood-framed apartment building." ***Energy and Buildings***, Vol. 42, No. 2, (2010), 230-242. https://doi.org/10.1016/j.enbuild.2009.08.018

71. Leckner, M., and Zmeureanu, R. "Life cycle cost and energy analysis of a Net Zero Energy House with solar combisystem." ***Applied Energy***, Vol. 88, No. 1, (2011), 232-241. https://doi.org/10.1016/j.apenergy.2010.07.031

72. Dahlstrøm, O., Sørnes, K., Eriksen, S. T., and Hertwich, E. G. "Life cycle assessment of a single-family residence built to either conventional-or passive house standard." ***Energy and Buildings***, Vol. 54, (2012), 470-479. https://doi.org/10.1016/j.enbuild.2012.07.029

73. Paulsen, J. S., and Sposto, R. M. "A life cycle energy analysis of social housing in Brazil: Case study for the program 'MY HOUSE MY LIFE.'" ***Energy and Buildings***, Vol. 57, (2013), 95-102. https://doi.org/10.1016/j.enbuild.2012.11.014

74. Paleari, M., Lavagna, M., and Campioli, A. "Life cycle assessment and zero energy residential buildings." In ***PLEA 2013***.

75. Berggren, B., Hall, M., and Wall, M. "LCE analysis of buildings-Taking the step towards Net Zero Energy Buildings." ***Energy and Buildings***, Vol. 62, (2013), 381-391. https://doi.org/10.1016/j.enbuild.2013.02.063

76. Stephan, A., and Stephan, L. "Reducing the total life cycle energy demand of recent residential buildings in Lebanon." ***Energy***, Vol. 74, (2014), 618-637. https://doi.org/10.1016/j.energy.2014.07.028

77. Dissanayake, D., Jayasinghe, C., and Jayasinghe, M. T. R. "A comparative embodied energy analysis of a house with recycled expanded polystyrene (EPS) based foam concrete wall panels." ***Energy and Buildings***, Vol. 135, (2017), 85-94. https://doi.org/10.1016/j.enbuild.2016.11.044

78. Vitale, P., Spagnuolo, A., Lubritto, C., and Arena, U. "Environmental performances of residential buildings with a

structure in cold formed steel or reinforced concrete." ***Journal of Cleaner Production***, Vol. 189, (2018), 839-852. https://doi.org/10.1016/j.jclepro.2018.04.088

79. Tavares, V., Lacerda, N., and Freire, F. "Embodied energy and greenhouse gas emissions analysis of a prefabricated modular house: The 'Moby' case study." ***Journal of Cleaner Production***, Vol. 212, (2019), 1044-1053. https://doi.org/10.1016/j.jclepro.2018.12.028

80. Thanu, H. P., Kumari, H. G. K., and Rajasekaran, C. "Sustainable Building Management by Using Alternative Materials and Techniques." In ***Sustainable Construction and Building Materials*** (pp. 583-593). Springer.

81. Hacker, J. N., De Saulles, T. P., Minson, A. J., and Holmes, M. J. "Embodied and operational carbon dioxide emissions from housing: A case study on the effects of thermal mass and climate change." ***Energy and Buildings***, Vol. 40, No. 3, (2008), 375-384. https://doi.org/10.1016/j.enbuild.2007.03.005

82. Blengini, G. A. "Life cycle of buildings, demolition and recycling potential: A case study in Turin, Italy." ***Building and Environment***, Vol. 44, No. 2, (2009), 319-330. https://doi.org/10.1016/j.buildenv.2008.03.007

83. Monteiro, H., and Freire, F. "Life-cycle assessment of a house with alternative exterior walls: Comparison of three impact assessment methods." ***Energy and Buildings***, Vol. 47, (2012), 572-583. https://doi.org/10.1016/j.enbuild.2011.12.032

84. Mao, C., Shen, Q., Shen, L., and Tang, L. "Comparative study of greenhouse gas emissions between off-site prefabrication and conventional construction methods: Two case studies of residential projects." ***Energy and Buildings***, Vol. 66, (2013), 165-176. https://doi.org/10.1016/j.enbuild.2013.07.033

85. Lamnatou, C., Notton, G., Chemisana, D., and Cristofari, C. "Life cycle analysis of a building-integrated solar thermal collector, based on embodied energy and embodied carbon methodologies." ***Energy and Buildings***, Vol. 84, (2014), 378-387. https://doi.org/10.1016/j.enbuild.2014.08.011

86. Galua, R. D., and Tobias, E. G. "An assessment of sustainability of a green residential building in an urban setting: focus in Pueblo de Oro, Cagayan de Oro City." ***Advances in Environmental Sciences***, Vol. 7, No. 1, (2015), 60-69.

87. Kumanayake, R., Luo, H., and Paulusz, N. "Assessment of material related embodied carbon of an office building in Sri Lanka." ***Energy and Buildings***, Vol. 166, (2018), 250-257. https://doi.org/10.1016/j.enbuild.2018.01.065

88. Citherlet, S., and Defaux, T. "Energy and environmental comparison of three variants of a family house during its whole life span." ***Building and Environment***, Vol. 42, No. 2, (2007), 591-598. https://doi.org/10.1016/j.buildenv.2005.09.025

89. Sartori, I., and Hestnes, A. G. "Energy use in the life cycle of conventional and low-energy buildings: A review article." ***Energy and Buildings***, Vol. 39, No. 3, (2007), 249-257. https://doi.org/10.1016/j.enbuild.2006.07.001

90. Utama, A., and Gheewala, S. H. "Life cycle energy of single landed houses in Indonesia." ***Energy and Buildings***, Vol. 40, No. 10, (2008), 1911-1916. https://doi.org/10.1016/j.enbuild.2008.04.017

91. Petrovic, B., Myhren, J. A., Zhang, X., Wallhagen, M., and Eriksson, O. "Life cycle assessment of a wooden single-family house in Sweden." ***Applied Energy***, Vol. 251, (2019), 113253. https://doi.org/10.1016/j.apenergy.2019.05.056

92. Hernandez, P., Oregi, X., Longo, S., and Cellura, M. "Life-Cycle Assessment of Buildings." In ***Handbook of Energy Efficiency in Buildings***, 207-261, Elsevier.

93. Tettey, U. Y. A., Dodoo, A., and Gustavsson, L. "Effect of different frame materials on the primary energy use of a multi storey residential building in a life cycle perspective." ***Energy and Buildings***, Vol. 185, (2019), 259-271. https://doi.org/10.1016/j.enbuild.2018.12.017

Persian Abstract

چکیده

بیش از 40 درصد انرژی مصرفی جهان در بخش ساخت و ساز مصروف می گردد. با این حال، برخی از کشورها معیارهای زیست محیطی را به عنوان الزامات طراحی در کدهای ساخت و ساز خود در نظر نمی گیرند. بر این اساس، هدف این تحقیق ارائه راهکاری است که با حفظ بافت‌های تاریخی و سنتی ایران، انرژی و کربن نهفته را کاهش دهد. لذا مقایسه کربن و انرژی نهفته بین ساختمان‌های بتنی جدید و ساختمان های سنتی با محاسبه میزان مصالح ساختمانی انجام شد. ضمن بررسی هر دو نوع ساختمان، کاهش کربن و انرژی نهفته در یک سیستم ساختمان ترکیبی مورد ارزیابی قرار گرفت. در ادامه با استفاده از تحلیل SWOT راهبردهای این ترکیب بررسی شد. ساختمان سنتی گِلی علیرغم داشتن جرم بیشتر مواد، انرژی و کربن نهفته کمتری نسبت به ساختمان بتنی دارد. با توجه به تحلیل SWOT، استراتژی یکپارچه سازی سیستم های ساختمانی سنتی گِلی و جدید بتنی ارائه شد. سیستم پیشنهادی به ترتیب حدود 40٪ و 35٪ کاهش کربن و انرژی نهفته را در مقایسه با سازه بتنی ایجاد می‌کند. گسترش این استراتژی در سراسر کشور باعث صرفه جویی 13 میلیون تن کربن و 130 میلیون گیگاژول انرژی نهفته می شود. یافتن راه حلی مبتنی بر ملاحظات پایداری برای حفظ بافت تاریخی یکی از دغدغه های اساسی کشورهایی است که این بافت ها بخشی از هویت آنها را تشکیل می دهند. در همین راستا سیستم ترکیبی ارائه شده، ضمن توجه به ساختمان پایدار و توسعه شهری، راه حلی مطلوب برای کاهش کربن و انرژی نهفته ساختمان است.

International Journal of Engineering

Journal Homepage: www.ije.ir

A Robust Control Chart for Monitoring Autocorrelated Multiple Linear Profiles in Phase I

F. Sogandi*[a], A. Amiri[b]

[a] *Department of Mechanical and Industrial Engineering, University of Torbat Heydarieh, Razavi Khorasan Province, Iran*
[b] *Department of Industrial Engineering, Shahed University, Tehran, Iran*

PAPER INFO

Paper history:
Received 12 March 2023
Received in revised form 13 May 2023
Accepted 14 May 2023

Keywords:
Profile Monitoring
Auto Correlated Multiple Linear Profiles
Robust Estimator
Phase I

ABSTRACT

Many problems do not have one or more variables that determine quality characteristics. In these situations, as a solution method, a profile is descibed by linking independent variables to the response variable. One of the common assumptions in most monitoring schemes is the assumption of independent residuals. Contravention of this assumption can lead to misleading results of the control chart. On the other hand, when the data are contaminated, the classical methods of estimating the parameters do not perform well. Such situations require robust estimation methods. Hence, this paper proposes a robust method to estimate the process parameters for Phase I monitoring autocorrelated multiple linear profiles. The developed control chart is appraised in the absence and presence of contaminated data through comprehensive simulation studies. The results showed that the robust estimator decreases the impact of contaminated data on the performance of the proposed control chart for all outlier percentages and shift magnitudes. Generally, in all three scenarios, including outliers in the model parameters and error variance, the robust approach performs better than the comparative method.

***doi**: 10.5829/ije.2023.36.08b.03*

1. INTRODUCTION

Control charts are an essential tool for quality practitioners to improve industrial and service processes. For example, Sogandi and Vakilian [1] used control chart to estimate a step change in Gamma regression profiles. Sometimes, the quality characteristic of a product or process can be described by a relationship between response and predictor variable(s) typically known as a profile. Profiles can be categorized based on their functional forms into polynomial profiles, simple linear profiles, multiple linear profiles, generalized linear model profiles and so on. As the first review papers on profile monitoring, Woodall et al. [2] and Woodall [3] provided a comprehensive introduction and research gaps on profile monitoring. In this respect, Saghaei et al. [4] surveyed different types of profiles, and introduced the definition and applications of profile monitoring. In real applications, John and Vaibha [5] also demonstrated the application of the control chart for monitoring the quality characteristics exhibiting a nonlinear profile during time. Sogandi and Vakilian [1] and Khedmati et al. [6] surveyed AR(1) autocorrelated structure to estimate a change point in simple linear and polynomial profile, respectively. Niaki et al. [7] also provided a control chart based on the generalized linear test to monitor coefficients of the simple linear profiles. More recently, Abbasi et al. [8] presented a new monitoring scheme for non-parametric profiles using an adaptive Exponentially Weighted Moving Average (EWMA) control chart. This control chart, EWMA is developed under a type II censoring life test by mohammadipour et al. [9]. For the sake of brevity, other related research about profile monitoring is referred to Maleki et al. [10], in which an overview is performed on research published during the period 2008–2018.

In the aforementioned studies, the profile parameters are often estimated by methods, which perform appropriately without outliers. However, in many real cases, there may exist some contamination on the

*Corresponding Author Institutional Email: f.sogandi@torbath.ac.ir (F. Sogandi)

Please cite this article as: F. Sogandi, A. Amiri, A Robust Control Chart for Monitoring Autocorrelated Multiple Linear Profiles in Phase I, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1429-1439 samples due to many reasons, such as the worker's fault. Applying the classical methods of parameter estimation in the presence of outliers would lead to inaccurate estimations and as a result, the erroneous performance of the monitoring scheme. To deal with these challenges, robust estimations are better properties than classical estimations. As one of the pioneering robust works, Khoo [11] suggested two time-weighted robust monitoring schemes for the process variance that the interquartile sample range for the control limits. In the profile monitoring field, Xuemin et al. [12] suggested a robust distribution-free approach to monitor linear profiles using rank-based regression to monitor nonparametric profiles. For simple linear profiles, Ebadi and Shahriari [13] used two robust methods, including the M-estimator and Huber estimator, in Phase I data with contamination. Similarly, Shahriari et al. [14] applied two methods for robust estimation of complex profiles using a nonparametric method for Phase I monitoring. After that, Shahriari and Ahmadi [15] proposed a robust estimation of complicated profiles. Also, Hakimi et al. [16] employed robust approaches using the M-estimator and the redescending M-estimator for Phase I monitoring of the logistic regression profile to reduce the impact of contaminated data. Furthermore, Ahmadi et al. [17] proposed a robust wavelet-based profile monitoring in Phase II in a two-stage process. For a simple linear profile, Hassanvand et al. [18] used two robust M-estimators for the parameter estimation to eliminate the detrimental impact of outliers in Phase I monitoring. After that, Kordestani et al. [19] suggested a monitoring scheme for monitoring multivariate simple linear profiles based on a robust estimation method. Moheghi et al. [20] considered robust estimation to monitor model parameters in GLM-based profiles with contaminated data. In recent years, Khedmati and Niaki [21] considered simple linear profiles based on robust parameter estimation in multistage processes in Phase-I. They proposed two robust methods, namely the MM-estimator and Huber’s M-estimator with outliers in historical data. Despite of the many studies in profile monitoring, there are few works for robust profile monitoring with autocorrelation within profile data.

The critical assumption in many profile monitoring procedures is that the observations within or between profiles are independent. However, there are many cases in the real world where this assumption is violated. So far, some work has shown correlations within or between profiles. In Phase I monitoring, Jensen et al. [22] suggested a mixed model to describe the autocorrelation structure within each profile. Moreover, Jensen and Birch [23] used nonlinear hybrid models to extend a monitoring scheme for autocorrelated nonlinear profiles. Afterward, Soleimani et al. [24] suggested a transformation to remove the autocorrelation structure between observations within simple linear profiles. In a similar method, Soleimani and Noorossana [25] studied the impact of autocorrelation in Phase II monitoring in multivariate simple linear profiles. Another research in this scope is Narvand et al. [26] in which they extended a Phase II monitoring scheme for auto-correlated linear profiles. In this paper, they used Hotelling’s T^2, multivariate cumulative sum, and multivariate EWMA control charts to monitor the process. In Phase II monitoring, Soleimani and Noorossana [27] developed a control chart for the multivariate simple linear profiles considering autocorrelation between observations for each profile. In the same type of autocorrelation, Yang et al. [28] suggested two Shewhart multivariate control charts to monitor a linear profile as well. To eliminate the effects of autocorrelation, Soleimani et al. [29] proposed three methods based on time series models for monitoring multivariate simple linear profiles with autocorrelation between profiles. Also, they demonstrated that the presence of outliers has a deleterious effect on the control chart performance.

Among a few works concentrating on robust methods for profile monitoring, only Kamranrad and Amiri [30] developed a robust control chart for auto-correlated simple linear profiles. Ahmadi et al. [31] suggested a control chart for Phase II monitoring of multiple linear profiles in which two robust estimate methods, the M-estimator, and fast-τ-estimator, were used. They showed their robust control chart based on M-estimator performs better than the fast-τ-estimator under high contamination data. To the best of the authors' knowledge, there is no more research on robust estimation for autocorrelated profiles monitoring. Hence, in this research, we considereed the robust monitoring of autocorrelated multiple linear profiles in Phase I. On this subject, the robust estimation approach will be appraised using the control signal probabilities. Besides, we survey the benefits of using the proposed approach against the classical estimation method with and without outliers. The structure of this paper is as follows: The second section provides the statistical model and corresponding assumptions of the considered process. Then, the classical and robust estimators were reviewed for the model parameters of autocorrelated multiple linear profiles. Section 3 proposed robust control chart for monitoring autocorrelated multiple linear profiles. Section 4 related to the performance evaluation by some simulation results to validate the proposed robust control chart. Finally, our concluding remarks and future studies provided in section 5.

2. STATISTICAL MODEL AND ASSUMPTION

In this section, we model the problem and describe the corresponding assumptions. Let *m* samples of observations be available, and *n* fixed values of the predictor variable in each sample. We define the autocorrelated multiple linear profile model for the j^{th} sample profile in which (x_i, y_{ij}) is the observation vector $(j=1,2,...,m)$. Assume that the process is in a state of statistical control, the autocorrelation within the profile can be modeled using Equation (1):

$$y_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_{ij},$$
$$\varepsilon_{ij} = \varphi\varepsilon_{(i\text{-}1)j} + a_{ij}, \tag{1}$$

where y_{ij} is i^{th} observation in j^{th} sample profile $(i=1,2,...,n)$. Let x_{ip} the p^{th} value of the independent variable for i^{th} observation which is fixed from sample to sample. Also, $\beta_k(k=0,1,2,...,p)$ are the parameters of the regression model in the autocorrelated multiple linear profile. ε_{ij}'s are the autocorrelated error terms and a_{ij}'s are independent identically normal distributed with mean zero and variance σ^2. We assume that there is autocorrelation within a multiple linear profile and the autocorrelation structure is a first-order autoregressive (AR(1)) model. In the following subsection, we will show how to eliminate the autocorrelation structure between observations within multiple linear profile.

2. 1. Autocorrelation Elimination Method

The autocorrelation structure among error terms leads to autocorrelation between data in each profile. Hence, a transformation method should be used to remove the impact of autocorrelation. In this regard, each observation is transformed using Equation (2):

$$y'_{ij} = y_{ij} - \varphi y_{(i\text{-}1)j} \tag{2}$$

According to Equations (1) and (2) can be easily written for (i-1)th observation in the j^{th} profile.

$$y_{(i\text{-}1')j} = \beta_0 + \beta_1 x_{(i\text{-}1')1} + \beta_2 x_{(i\text{-}1')2} + ... + \beta_p x_{(i\text{-}1')p} + \varepsilon_{(i\text{-}1)j}, \tag{3}$$

By replacing Equations (1) and (3) into Equation (2), and simplification it, for each observation, we will obtain Equation (4):

$$y'_{ij} = \beta_0(1-\varphi) + \beta_1(x_{1i} - \varphi x_{1(i\text{-}1')}) + ...$$
$$+\beta_p(x_{pi} - \varphi x_{p(i\text{-}1')}) + (\varepsilon_{ij} - \varphi\varepsilon_{(i\text{-}1)j}), \tag{4}$$

leading to Equation (5):

$$y'_{ij} = \beta'_0 + \beta'_1 x'_{1i} + \beta'_1 x'_{2i} + ... + \beta'_p x'_{pi} + a_{ij}, \tag{5}$$

In which $\beta'_0 = \beta_0(1-\varphi)$, and a_{ij}'s are independent random variables with mean zero and variance σ^2. Moreover, $\beta'_k = \beta_k$, $x'_{ki} = x_{ki} - \varphi x_{k(i\text{-}1')}$, for each explanatory variable $(k=1,2,...,p)$. As it is clear Equation (5) is a multiple linear profile with independent error terms. In the next section, the proposed methods of parameter estimation are given.

2. 2. Robust Estimation of Model Parameters

Usually, for uncontaminated cases, the ordinary least-square estimation (LSE) method is utilized to estimate the model parameters. For each sample, the least-square estimator for $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ is achieved using minimizing the sum of squared errors $\left|(\mathbf{y}-\mathbf{x\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{x\beta})\right|$, and it is given by Equation (6).

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^{\mathbf{T}}\mathbf{x}\right)^{-1}\mathbf{x}^{\mathbf{T}}\mathbf{y}. \tag{6}$$

Equation (6) should be derived using all samples, even out-of-control profiles. This effect is known as the masking effect and leads to changing the value of statistics. The signicant outliers intensify this impact. To deal with this challenge, a robust estimation method should be used. If few contaminated data exist in a random sample, there are two methods to cope with this sample, including eliminating it and keeping it in which some information may be eliminated, or inaccurate estimates may be achieved. Hence, applying robust estimations is rational because they give unbiased estimations, under both contaminated data and outlier free. On this subject, researchers have applied robust regression with slighter sensitivity to outliers using appropriate weighting method. Many robust estimators have been suggested so far. Among these methods, M-estimator is the most introduced method introduced by Huber [32] because it has higher efficiency. On the other hand, Ahmadi et al. [32] showed the M-estimator is better than the fast-τ-estimator in high contamination for Phase II monitoring of multiple linear profiles. Hence, we estimate the parameters of autocorrelated multiple linear profile using the M-estimator, which are a generalization of maximum likelihood estimation.

Usually, s is the median absolute deviation, a robust estimator, defined as follows by Abu-Shawiesh [33] according to Equation [7]:

$$s = \frac{med\left|e_i \text{ - } med\left(e_i\right)\right|}{0.6745}. \tag{7}$$

Considering s as a robust scale estimate, M-estimator could be calculated using minimizing a function $\rho(.)$ of regression residuals according to Equation (8):

$$\min\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right), \tag{8}$$

in which ρ is a function of Huber or bisquare weight function. The bisquare function, as one of the main weighting functions is used here. The $\psi(x)$ is the derivative of $\rho(.)$ and the other functions, from a family of bisquare function given by Shahriari et al. [34]

according to Equations (9) and (10):

$$\rho(x)=\begin{cases} 1-\left(1-\left(\frac{x}{k}\right)^2\right)^3 & |x|\le k \\ 1 & |x|>k \end{cases}, \tag{9}$$

$$w(x)=\begin{cases} \left(1-\left(\frac{x}{k}\right)^2\right)^2 & |x|\le k \\ 0 & |x|>k \end{cases}, \tag{10}$$

where k value should be chosen so that the resultant estimate would have a suitable asymptotic σ^2. Shahriari et al. [34] proved that these functions apply well with k=4.68.

2. 3. Proposed Robust Monitoring Scheme for Autocorrelated Multiple Linear Profiles

We use T_I^2 which is based on intra-profile pooling and sample average in Phase I monitoring. Consider $\hat{\boldsymbol{\beta}}_j$ can be shown by $(\hat{\beta}_{0j},\hat{\beta}_{1j},\hat{\beta}_{2j},...,\hat{\beta}_{pj})$ vector for each profile. Note that estimation of $\text{var}(\hat{\boldsymbol{\beta}}_j)$ is equal to Equation (11). For more details see Yeh et al. [35].

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}_j)=\left(\mathbf{X}^{\mathbf{T}}\mathbf{W}_{\mathbf{j}}\mathbf{X}\right)^{-1}. \tag{11}$$

Hence, an estimate of variance-covariance matrix could be obtained by taking the average of values of $\hat{\text{var}}(\hat{\boldsymbol{\beta}}_j)$ according to the $\mathbf{S}_I=\frac{1}{m}\sum_{j=1}^{m}\hat{\text{var}}(\hat{\boldsymbol{\beta}}_j)$. In a similar way, the estimation of average parameters is equal to the $\bar{\hat{\boldsymbol{\beta}}}=\frac{\sum_{i=1}^{m}\hat{\boldsymbol{\beta}}_j}{m}$ across all m samples. Therefore, T_I^2 control chart is obtained by Equation (12) to monitor the regression model parameters in autocorrelated multiple linear profiles.

$$T_{I,j}^2=\left(\hat{\boldsymbol{\beta}}_j-\bar{\hat{\boldsymbol{\beta}}}\right)^{\mathbf{T}}\mathbf{S}_I^{-1}\left(\hat{\boldsymbol{\beta}}_j-\bar{\hat{\boldsymbol{\beta}}}\right). \tag{12}$$

The proposed T_I^2 control chart trigger a statistical alarm when $T_{I,j}^2>UCL$ in which Upper Control Limit (UCL) is obtained by $UCL=(p+1)F_{p+1,m(n-p-1),\alpha}$. In this regard, Figure 1 depicts a general graphical scheme about the robust control chart for Phase I monitoring of autocorrelated multiple linear profiles.

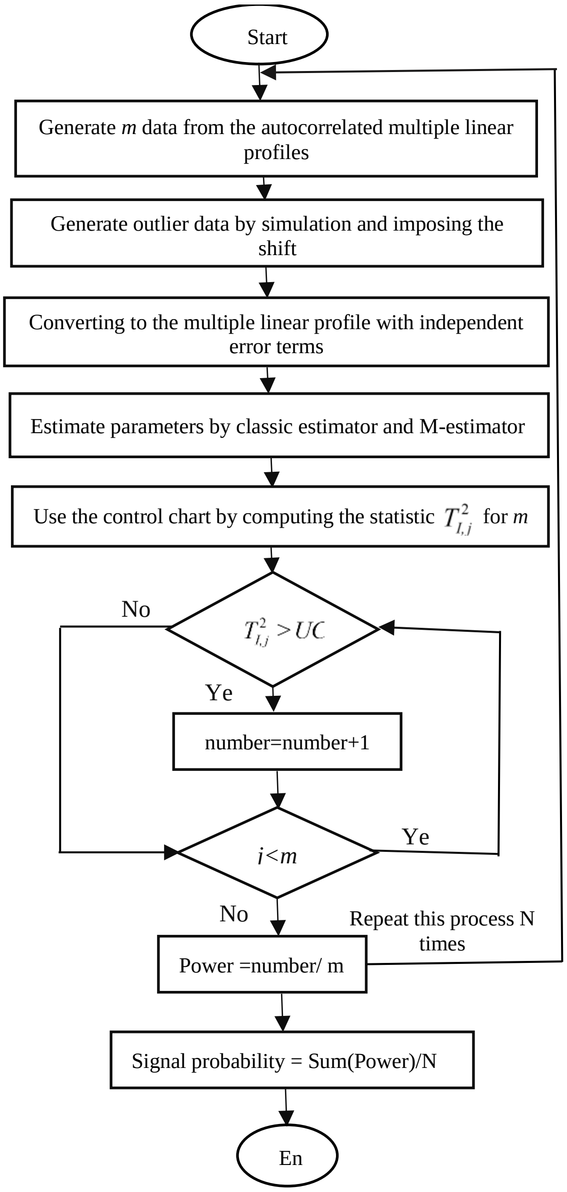

Start
Generate m data from the autocorrelated multiple linear profiles
Generate outlier data by simulation and imposing the shift
Converting to the multiple linear profile with independent error terms
Estimate parameters by classic estimator and M-estimator
Use the control chart by computing the statistic $T_{I,j}^2$ for m
No
$T_{I,j}^2>UC$
Ye
number=number+1
i<m
Ye
No
Repeat this process N times
Power =number/ m
Signal probability = Sum(Power)/N
En

Figure 1. Flowchart of proposed robust monitoring scheme

3. SIMULATION STUDY AND PERFORMANCE EVALUATION

In this section, taking into account contaminated data, some simulation studies are provided to evaluate the performance of the proposed monitoring scheme in Phase I. The number of runs in Monte-Carlo simulation is 10000 in R software. On this subject, to apply classical and robust estimators, a simulation example of an autocorrelated multiple linear regression model is utilized to generate the data by Equation (13):

$$\begin{aligned} y_{ij}&=3+1.2x_{1i}+1.3x_{2i}+\varepsilon_{ij},\\ \varepsilon_{ij}&=0.8\varepsilon_{(i-1)j}+a_{ij}, \end{aligned} \tag{13}$$

In which a_{ij} is the independent random variable, and follow a Normal distribution with mean 0 and $\sigma^2=1$. Let explanatory variables equal to $\mathbf{x}_0=(1,1,...,1)$, $\mathbf{x}_1=(0.2,0.4,0.6,...,4)$ and $\mathbf{x}_2=(0.1,0.3,0.5,...,2)$. Consider that 10 observations are generated for each level of explanatory variables. Hence, the total number of observations in each profile is 200. To appraise the estimation of the model parameters, 30 random samples are generated under different shifts and given contamination percentages. After that, a percentage of the simulated data is contaminated by shifting the model parameters of the autocorrelated multiple linear profile as $\beta_0+\lambda,\ \beta_1+\lambda, \beta_2+\lambda.$

In this regard, different contamination percentages are considered using global outliers to evaluate the robust and classical estimates. Then, the mean and standard deviation of estimates in autocorrelated multiple linear profile are calculated under global outlying conditions. In the global contamination, a given percent of observations in all profiles should be replaced with contaminated data. For this aim, (*c*) percent of the data of each profile include outliers, and (100-*c*) percent of them are simulated by the pre-specified autocorrelated multiple linear profile. In other words, 10 levels were randomly selected from all the profiles, and in this regard, even levels are considered. According to the conducted simulation study, Table 1 shows the accuracy and standard deviation of both estimators in the presence of outliers in which λ ($\lambda = 0.3, 0.6, 0.9, 1.2, 1.5$) is the shift magnitude in the intercept.

Based on the simulation results provided in Table 1, in the absence of contamination, robust and classical estimators are almost identical. Also, it can be inferred that the proposed robust estimation method outperforms the classical estimation method in the presence of contamination. That is, the robust estimator gives more accurate estimates of β_0 compared to the estimator obtained by the LSE method regardless of outlier percentages and shift magnitudes. The conventional criterion used in Phase I monitoring for performance comparison of control charts is the probability of signal. Hence, we calculate signal probability of T^2 control chart after estimation of the regression coefficients. When there are no outliers in the process, the upper control limit of the T^2 control chart is set equal to 10.83 considering α=0.005. In this regard, Table 2 gives simulation results for different shifts with contamination in the intercept parameter.

TABLE 1. Performance evaluation of classical and robust estimations under contamination in β_0

Method		**Classic**		**Robust**	
c	**Shift (λ)**	**Parameter estimation**	**Standard deviation**	**Parameter estimation**	**Standard deviation**
5	**0**	2.911	1.142	2.999	0.940
	0.3	3.104	1.147	3.007	0.941
	0.6	3.155	1.175	3.025	0.943
	0.9	3.221	1.189	3.031	0.944
	1.2	3.284	1.200	3.080	0.955
	1.5	3.419	1.208	3.137	0.966
10	**0**	3.006	1.164	2.974	0.939
	0.3	3.247	1.170	3.006	0.940
	0.6	3.303	1.178	3.038	0.945
	0.9	3.382	1.183	3.127	0.952
	1.2	3.558	1.193	3.206	0.972
	1.5	3.576	1.231	3.267	0.988
15	**0**	3.040	1.142	2.962	0.936
	0.3	3.255	1.156	3.054	0.946
	0.6	3.370	1.163	3.172	0.954
	0.9	3.414	1.181	3.231	0.982
	1.2	3.777	1.191	3.279	0.984
	1.5	3.945	1.284	3.428	1.002
20	**0**	3.054	1.181	2.976	0.949
	0.3	3.243	1.182	3.031	0.956
	0.6	3.472	1.191	3.193	0.960
	0.9	3.738	1.201	3.330	0.976
	1.2	3.834	1.206	3.578	0.981
	1.5	4.125	1.246	3.761	0.986
25	**0**	3.077	1.281	3.072	0.959
	0.3	3.343	1.290	3.111	0.966
	0.6	3.482	1.131	3.273	0.970
	0.9	3.838	1.322	3.360	0.986
	1.2	3.935	1.336	3.777	0.991
	1.5	4.204	1.346	3.851	0.992

TABLE 2. Performance of T^2 control chart for shifts of various magnitudes in the presence of contamination in β_0

Method		**Classic**	**Robust**
c	**Shift (λ)**	**Signal probability**	**Signal probability**
5	**0**	0.005	0.005
	0.3	0.328	0.616
	0.6	0.461	0.679
	0.9	0.563	0.741
	1.2	0.667	0.771
	1.5	0.754	0.809
10	**0**	0.005	0.005
	0.3	0.444	0.733
	0.6	0.643	0.834
	0.9	0.820	0.916
	1.2	0.889	0.961
	1.5	0.960	0.987

15	**0**	0.005	0.005
	0.3	0.566	0.787
	0.6	0.802	0.925
	0.9	0.931	0.981
	1.2	0.982	0.977
	1.5	0.992	0.998
20	**0**	0.005	0.005
	0.3	0.629	0.856
	0.6	0.871	0.957
	0.9	0.969	0.996
	1.2	0.985	0.999
	1.5	0.999	0.999
25	**0**	0.005	0.005
	0.3	0.658	0.876
	0.6	0.882	0.966
	0.9	0.973	0.998
	1.2	0.986	0.999
	1.5	0.999	0.999

According to Table 2, when there is contamination in the intercept parameter, the robust control chart will show considerably better performance than classical control chart. Moreover, it shows that the presence of outliers in the clean observations causes to increase the signal probabilities. Also, whatever the magnitude of shifts increases, the signal probability values will be larger in both estimators. Similar to the previous tables, when there is contamination in β_1, Tables 3 and 4 summarize the estimators and signal probability of T^2 control chart, respectively.

TABLE 3. Performance evaluation of classical and robust estimations under contamination in β_1

Method		**Classic**		**Robust**	
c	**Shift (λ)**	**Parameter estimation**	**Standard deviation**	**Parameter estimation**	**Standard deviation**
5	**0**	1.055	1.124	1.154	0.923
	0.3	1.244	1.165	1.186	0.941
	0.6	1.487	1.171	1.234	0.950
	0.9	1.625	1.185	1.493	0.957
	1.2	1.989	1.158	1.711	0.929
	1.5	2.226	1.164	2.024	0.969
10	**0**	1.045	1.170	1.182	0.964
	0.3	1.300	1.208	1.259	0.967
	0.6	1.519	1.167	1.381	0.951
	0.9	1.736	1.180	1.464	0.952
	1.2	2.044	1.185	1.988	0.973
	1.5	2.299	1.197	2.074	0.988
15	**0**	1.002	1.148	1.173	0.948
	0.3	1.359	1.152	1.229	0.944
	0.6	1.617	1.159	1.267	0.957
	0.9	1.805	1.167	1.427	0.974
	1.2	2.010	1.160	1.551	0.979
	1.5	2.358	1.186	2.035	0.982
20	**0**	1.013	1.168	1.208	0.933
	0.3	1.367	1.204	1.276	0.935
	0.6	1.606	1.152	1.426	0.949
	0.9	1.853	1.156	1.848	0.950
	1.2	2.002	1.166	2.015	0.974
	1.5	2.393	1.226	2.054	0.991

TABLE 4. Performance of T^2 control chart for shifts of various magnitudes in the presence of contamination in β_1

Method		**Classic**	**Robust**
c	**Shift (λ)**	**Signal probability**	**Signal probability**
5	**0**	0.005	0.005
	0.3	0.262	0.569
	0.6	0.274	0.591
	0.9	0.292	0.638
	1.2	0.314	0.655
	1.5	0.349	0.752
10	**0**	0.005	0.005
	0.3	0.285	0.605
	0.6	0.382	0.668
	0.9	0.427	0.724
	1.2	0.508	0.764
	1.5	0.566	0.815
15	**0**	0.005	0.005
	0.3	0.336	0.634
	0.6	0.466	0.761
	0.9	0.577	0.837
	1.2	0.675	0.895
	1.5	0.799	0.943
20	**0**	0.005	0.005
	0.3	0.405	0.742
	0.6	0.593	0.833
	0.9	0.749	0.940
	1.2	0.837	0.966
	1.5	0.922	0.994

The obtained results from simulation runs show both classical and robust estimation methods are almost similar under the clean data. However, robust estimator decreases the effect of outliers on the mean of estimated parameters with outliers. In other words, robust estimator values are closer to the in-control β_1 than the classical estimator under different shifts and outlier observations. Moreover, comparing the standard deviation of them demonstrates that the robust estimation method is better than the least-square estimation method in the presence of outliers. Note that the classical estimation method of standard deviation performs roughly better than the robust estimation method without contamination. Afterward, outliers are generated with shift in β_2 and the corresponding mean and standard deviation values are given in Table 5. Also, simulation results of signal probability of the proposed control chart are given in Table 6.

TABLE 5. Performance evaluation of classical and robust estimations under contamination in β_2

Method		Classic		Robust	
c	Shift (λ)	Parameter estimation	Standard deviation	Parameter estimation	Standard deviation
5	0	1.024	1.147	1.301	0.936
	0.3	1.382	1.167	1.342	0.941
	0.6	1.482	1.170	1.404	0.953
	0.9	1.715	1.181	1.615	0.961
	1.2	2.001	1.185	1.935	0.979
	1.5	2.206	1.198	2.055	0.999
10	0	1.052	1.129	1.328	0.914
	0.3	1.452	1.114	1.378	0.933
	0.6	1.771	1.114	1.562	0.934
	0.9	1.839	1.117	1.781	0.937
	1.2	2.116	1.149	2.027	0.935
	1.5	2.220	1.157	2.129	0.936
15	0	1.041	1.140	1.306	0.916
	0.3	1.440	1.145	1.310	0.925
	0.6	1.592	1.147	1.415	0.932
	0.9	1.877	1.149	1.762	0.961
	1.2	1.994	1.163	1.896	0.968
	1.5	2.307	1.185	2.220	0.980
20	0	1.051	1.177	1.321	0.901
	0.3	1.176	1.197	1.354	0.913
	0.6	1.527	1.148	1.452	0.947
	0.9	1.740	1.176	1.671	0.955
	1.2	1.992	1.180	1.987	0.942
	1.5	2.136	1.197	2.096	0.960

TABLE 6. Performance of T^2 control chart for shifts of various magnitudes in the presence of contamination in β_2

Estimation method		Classic	Robust
c	Shift (λ)	Signal probability	Signal probability
5	0	0.005	0.005
	0.3	0.246	0.575
	0.6	0.252	0.576
	0.9	0.257	0.579
	1.2	0.273	0.586
	1.5	0.372	0.599
10	0	0.005	0.005
	0.3	0.286	0.608
	0.6	0.299	0.644
	0.9	0.342	0.666
	1.2	0.407	0.677
	1.5	0.443	0.743
15	0	0.005	0.005
	0.3	0.325	0.654
	0.6	0.428	0.731
	0.9	0.510	0.788
	1.2	0.593	0.828
	1.5	0.671	0.887
20	0	0.005	0.005
	0.3	0.371	0.686
	0.6	0.525	0.824
	0.9	0.623	0.906
	1.2	0.734	0.955
	1.5	0.844	0.988

Similarly, Table 5 demonstrates satisfactory performance for robust estimator under global outliers, as the proposed robust estimator reduces their impact. The robust estimator with no outliers has 0.936 standard deviation, which is lower than the classical estimator (1.147). A close match between the robust estimator and the corresponding actual value is shown in Table 5. Also, Table 6 shows that the developed T^2 chart by a robust estimator is a more efficient scheme than the T^2 chart based on the classical estimator in Phase I monitoring.

To take account into the impact of contminations on the variance of error terms, let (1-c) percent of ε_{ij}'s independently follow a Normal distribution as with N(0, σ^2). Besides, let *c* percent of the residuals generate an another Normal distribution. In other words, a model (say uncontaminated case) in which all observation are from N(0,1). A model for symmetric variance disturbances in which each observation has (1-*c*)% probability of being drawn from N(0,1) distribution and a *c*% probability of being drawn from N(0,9).The Mean Squared Error (MSE) criterion is applied to appraise the capability of error terms variance estimators. A smaller MSE value indicates a more accurate estimation of the parameter. Figure 2 shows the MSE of σ^2 estimations if there is contamination in the variance of the error terms.

Figure 2 illustrates when shift magnitudes and outlier percentages increase, robust approach performs better than classical approach. Furthermore, low contamination in variance of the ε_{ij}'s does not significantly affect classical estimation of parameters. While, by increasing the contaminated error terms variance, the classical estimator of ε_{ij}'s variance becomes significantly different from the actual value. Despite the satisfactory performance of the classical estimator for some low values of σ and *c*, with moderate and large contamination rates, it has worse performance than the robust estimation method. In these simulation studies, the maximum estimates for variance of the ε_{ij}'s based on the classical estimator was 11.32. However, this value is 1.58 for the robust estimator. Moreover, as shown in Figure 3, the robust scheme increases the contamination percent. Besides, we taken into account other simulations with different σ and *c* values. For the brevity, these simulation studies, not given here, support the results shown in these figures.

4. A REAL CASE

To show the practicality and effectiveness of the proposed robust control chart, we present a real case derived from the automotive industry given by Amiri et al. [36]. Specifically, when evaluating an automobile engine's performance, a crucial quality characteristic is how torque production relates to engine speed stated in revolutions per minute. We have 26 engines available for the initial phase of analysis pertaining to the engine data. Within each engine, we establish a set of speed values, including 1500, 2000, 2500, 2660, 2800, 2940, 3500, 4000, 4500, 5000, 5225, 5500, 5775, and 6000 RPM, and collect corresponding torque measurements. Consequently, we obtain a profile of interest consisting of 14 data points per engine. They showed for this data set that a quadratic polynomial works well according to Equation (14). In is worth mentioning that polynomial profiles is a special case of multiple linear profiles.

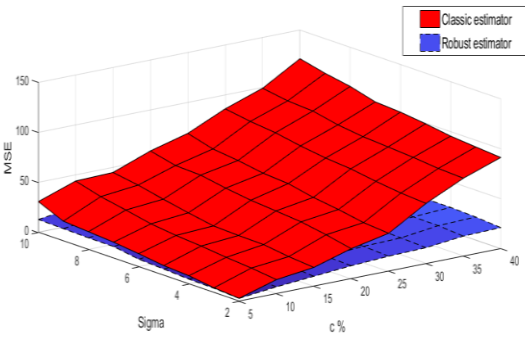


Figure 2. The comparison of MSE of variance estimations for *c*% contamination in error terms distribution and different shifts

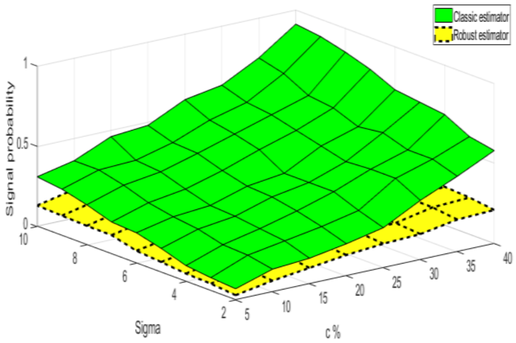


Figure 3. Performance of T^2 control chart for *c*% contamination in error terms distribution and different shifts

$$y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \varepsilon_j, \tag{14}$$

In which y_j denotes torque values, and x_j show RPM values in which $x_j = x$. The model parameters are estimated for each profile by high values of the adjusted coefficient of determination. Figure 4 depicts a scatterplot showcasing the data for one specific engine, identified as Engine number 1791. The figure serves as an example to demonstrate that the speed values have been adjusted for mean correction in order to mitigate the impact of multicollinearity. The variance inflation factors showed that there is no multi-collinearity between explanatory variables. They used a run chart to check independence of residuals over time assumption and showed that the clustering and trend hypothesis tests are significant and as a result it can be concluded that the residuals are correlated. In other words, the process of evaluating model adequacy revealed that we are dealing with a situation where there is a correlation between the residuals and therefore between the observations in each profile. Hence, the data set can be modeled by autocorrelated multiple linear profiles.

Amiri et al. [36] showed an AR(1) error structure using Draftman's display for the data set. Besides, the estimate of the mean vector and the covariance matrix under the classical and robust methods are reported in Table 7.

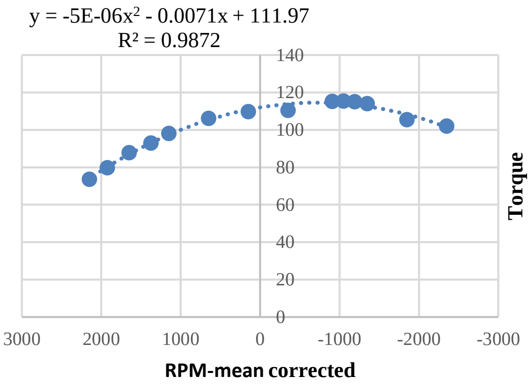


Figure 4. A second-order multiple linear profile fit for engine number 1791

TABLE 7. The parameter estimates of the real case obtained by the classical and the robust estimator

Estimates		Sample mean		Sample standard deviation	
Outlier percentages		Classic	Robust	Classic	Robust
0	β_0	111.2589	111.2586	1.5299	1.529876
	β_1	−0.005985	−0.005775	0.000496	0.000489
	β_2	−0.0000049	−0.0000051	0.0000031	0.00000309
20	β_0	114.577	111.4313	1.69988	1.52969
	β_1	−0.007798	−0.005784	0.005096	0.000477
	β_2	−0.0000041	−0.00000508	0.00004	0.0000032

The classical and robust estimates indicate that the process mean as well as the process standard deviation are influenced by outliers. To examine the stability of the process and identify any unusual profiles among the 26, we can establish UCL equal to 0.7089 at 95% confidence level for Phase II of the control chart. Then, we can employ $T^2_{I,j}$ based on classic and robust approach. Hence, we use the T^2 control chart to monitor the process mean. The control chart statistics are also determined for the sample data points using Equation (12) with the classic and robust estimators and plotted Figure 5. From Figure 5, it is clear that robust control charts gave a quick out of control signal by last two data points whereas classical estimator-based control chart did not signal any alarm.

5. CONCLUSION AND FURTHER STUDIES

An assumption commonly used in profile monitoring schemes is that residuals will be independent. The control chart can be misled if this assumption is violated. From another standpoint, when the data are contaminated, the classical estimation methods do not perform well. Hence,this paper proposed a robust approach for Phase I monitoring of autocorrelated multiple linear profiles.

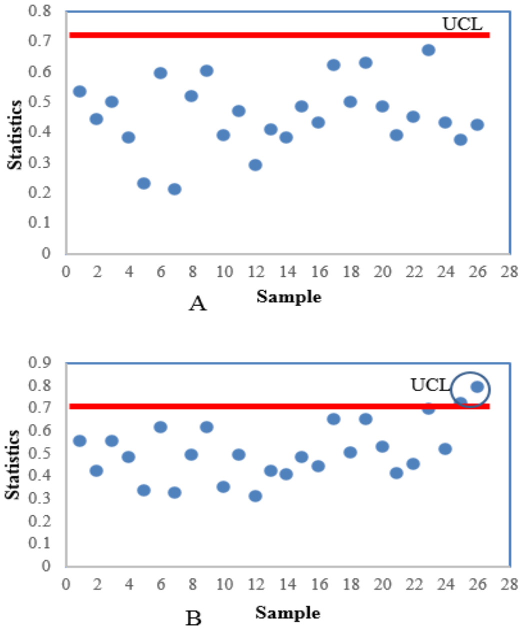


Figure 5. Plots of sample data for T^2 control chart with classical (A), and robust (B), estimators

The proposed control chart was appraised in the absence and presence of contaminated data through extensive simulation studies. The simulation results showed that without outliers, the classical and robust method performed partly the same in parameter estimation of the model. Considering these results, when the outlier magnitude increased, the estimations achieved by the classical estimator deviated considerably from their actual values. While, the estimates computed based on the robust estimator were close to the reference values. We showed the LSE method was affected by the outlier data, however, the M-estimator decreased their effect. Generally, in all scenarios including contaminations in the model coefficients and error variance, the robust approach performed better than the classical method. Besides, the performance of the T^2 control chart with the classical and the robust estimates was appraised by extensive comparison under different shift magnitudes with and without outliers. When the regression parameters were estimated using the robust method, the capability of the proposed T^2 control chart enhanced under different shifts in the parameters of regression model.

Considering the proposed robust approaches for multistage processes can be a fruitful subject for future research. Also, investigating the performance of the proposed robust estimator under autocorrelation between profiles and contamination data can be considered as a further research.

6. REFERENCES

1. Sogandi, F. and Vakilian, F., "Isotonic change point estimation in the ar (1) autocorrelated simple linear profiles", ***International Journal of Engineering, Transactions A: Basics***, Vol. 28, No. 7, (2015), 1059-1067. doi: 10.5829/idosi.ije.2015.28.07a.12.

2. Woodall, W.H., Spitzner, D.J., Montgomery, D.C. and Gupta, S., "Using control charts to monitor process and product quality profiles", ***Journal of Quality Technology***, Vol. 36, No. 3, (2004), 309-320. doi: 10.1080/00224065.2004.11980276.

3. Woodall, W.H., "Current research on profile monitoring", ***Production***, Vol. 17, (2007), 420-425. doi: 10.1080/07408170600998769.

4. Saghaei, A., Noorossana, R. and Amiri, A., "Statistical analysis of profile monitoring, Wiley Online Library, (2013).

5. John, B. and Agarwal, V., "A regression spline control chart for monitoring characteristics exhibiting nonlinear profile over time", ***The TQM Journal***, Vol. 31, No. 3, (2019), 507-522. doi: 10.1108/TQM-11-2018-0183.

6. Khedmati, M., Soleymanian, M.E., Keramatpour, M. and Niaki, S., "Monitoring and change point estimation of ar (1) autocorrelated polynomial profiles", ***International Journal of Engineering, Transactions A: Basics***, Vol. 26, No. 9, (2013), 933-942. doi: 10.5829/idosi.ije.2013.26.09a.11.

7. Niaki S.T.A., Abbasi, B. and Arkat, J., "A generalized linear statistical model approach to monitor profiles", ***International Journal of Engineering, Transactions A: Basics,***, Vol. 20, No. 3, (2007). doi: 10.5829/idosi.ije.2007.20.03.006.

8. Abbasi, S.A., Yeganeh, A. and Shongwe, S.C., "Monitoring non-parametric profiles using adaptive ewma control chart", ***Scientific Reports***, Vol. 12, No. 1, (2022), 14336. doi: 10.1038/s41598-022-39011-0.

9. Mohammadipour, P., Farughi, H., Rasay, H. and Arkat, J., "Designing exponentially weighted moving average control charts under failure censoring reliability tests", ***International Journal of Engineering,Transactions B: Applications***, Vol. 34, No. 11, (2021), 2398-2407. doi: 10.5829/ije.2021.34.11b.03.

10. Maleki, M.R., Amiri, A. and Castagliola, P., "An overview on recent profile monitoring papers (2008–2018) based on conceptual classification scheme", ***Computers & Industrial Engineering***, Vol. 126, (2018), 705-728. doi: 10.1016/j.cie.2018.08.030.

11. Khoo, M.B., "Robust time weighted control charts for the process variance", ***International Journal of Reliability, Quality and Safety Engineering***, Vol. 12, No. 05, (2005), 439-458. doi: 10.1142/S0218539305001977.

12. Zi, X., Zou, C. and Tsung, F., "A distribution-free robust method for monitoring linear profiles using rank-based regression", ***IIE Transactions***, Vol. 44, No. 11, (2012), 949-963. doi: 10.1080/0740817X.2011.605559.

13. Ebadi, M. and Shahriari, H., "Robust estimation of parameters in simple linear profiles using m-estimators", ***Communications in Statistics-Theory and Methods***, Vol. 43, No. 20, (2014), 4308-4323. https://doi.org/10.1080/03610926.2012.721914

14. Shahriari, H., Ahmadi, O. and Samimi, Y., "Estimation of complicated profiles in phase i, clustering and s-estimation approaches", ***Quality and Reliability Engineering International***, Vol. 32, No. 7, (2016), 2455-2469. doi: 10.1002/qre.2049.

15. Shahriari, H. and Ahmadi, O., "Robust estimation of complicated profiles using wavelets", ***Communications in Statistics-Theory and Methods***, Vol. 46, No. 4, (2017), 1573-1593. doi: 10.1080/03610926.2015.1114893.

16. Hakimi, A., Amiri, A. and Kamranrad, R., "Robust approaches for monitoring logistic regression profiles under outliers", ***International Journal of Quality & Reliability Management***, (2017). doi: 10.1108/IJQRM-02-2015-0024.

17. Ahmadi, O., Shahriari, H. and Samimi, Y., "A robust wavelet based profile monitoring and change point detection using s-estimator and clustering", ***Journal of Industrial and Systems Engineering***, Vol. 11, No. 3, (2018), 167-189. doi: 10.1080/24725854.2018.1565149.

18. Hassanvand, F., Samimi, Y. and Shahriari, H., "A robust control chart for simple linear profiles in two-stage processes", ***Quality and Reliability Engineering International***, Vol. 35, No. 8, (2019), 2749-2773. doi: 10.1002/qre.2543.

19. Kordestani, M., Hassanvand, F., Samimi, Y. and Shahriari, H., "Monitoring multivariate simple linear profiles using robust estimators", ***Communications in Statistics-Theory and Methods***, Vol. 49, No. 12, (2020), 2964-2989. doi: 10.1080/03610926.2018.1568453.

20. Moheghi, H.R., Noorossana, R. and Ahmadi, O., "Glm profile monitoring using robust estimators", ***Quality and Reliability Engineering International***, Vol. 37, No. 2, (2021), 664-680. doi: 10.1002/qre.2689.

21. Khedmati, M. and Niaki, S.T.A., "Phase-i robust parameter estimation of simple linear profiles in multistage processes", ***Communications in Statistics-Simulation and Computation***, Vol. 51, No. 2, (2022), 460-485. https://doi.org/10.1080/03610918.2019.1653916

22. Jensen, W.A., Birch, J.B. and Woodall, W.H., "Monitoring correlation within linear profiles using mixed models", ***Journal of Quality Technology***, Vol. 40, No. 2, (2008), 167-183. doi: 10.1080/00224065.2008.11918540.

23. Jensen, W.A. and Birch, J.B., "Profile monitoring via nonlinear mixed models", ***Journal of Quality Technology***, Vol. 41, No. 1, (2009), 18-34. doi: 10.1080/00224065.2009.11917898.

24. Soleimani, P., Noorossana, R. and Amiri, A., "Simple linear profiles monitoring in the presence of within profile autocorrelation", ***Computers & Industrial Engineering***, Vol. 57, No. 3, (2009), 1015-1021. doi: 10.1016/j.cie.2009.06.019.

25. Soleimani, P. and Noorossana, R., "Investigating effect of autocorrelation on monitoring multivariate linear profiles", ***International Journal of Industrial Engineering & Production Research,*** , Vol. 23, No. 3, (2012), 187-193. doi: 10.22059/ijiepr.2012.30681.

26. Narvand, A., Soleimani, P. and Raissi, S., "Phase ii monitoring of auto-correlated linear profiles using linear mixed model", ***Journal of Industrial Engineering International***, Vol. 9, (2013), 1-9. doi: 10.1186/2251-712X-9-22.

27. Soleimani, P. and Noorossana, R., "Monitoring multivariate simple linear profiles in the presence of between profile autocorrelation", ***Communications in Statistics-Theory and Methods***, Vol. 43, No. 3, (2014), 530-546. https://doi.org/10.1080/03610926.2012.665554

28. Zhang, Y., He, Z., Zhang, C. and Woodall, W.H., "Control charts for monitoring linear profiles with within-profile correlation using gaussian process models", ***Quality and Reliability Engineering International***, Vol. 30, No. 4, (2014), 487-501. doi: 10.1002/qre.1589.

29. Soleimani, P., Noorossana, R. and Niaki, S., "Monitoring autocorrelated multivariate simple linear profiles", ***The International Journal of Advanced Manufacturing Technology***, Vol. 67, No. 5-8, (2013), 1857-1865. doi: 10.1007/s00170-012-4648-4.

30. Kamranrad, R. and Amiri, A., "Robust holt-winter based control chart for monitoring autocorrelated simple linear profiles with contaminated data", ***Scientia Iranica***, Vol. 23, No. 3, (2016), 1345-1354. doi: 10.24200/sci.2016.2142.

31. Ahmadi, M.M., Shahriari, H. and Samimi, Y., "A novel robust control chart for monitoring multiple linear profiles in phase ii", ***Communications in Statistics-Simulation and Computation***, Vol. 51, No. 11, (2022), 6257-6268. https://doi.org/10.1080/03610918.2020.1799228

32. Huber, P.J., "Robust regression: Asymptotics, conjectures and monte carlo", ***The Annals of Statistics***, (1973), 799-821. doi: 10.1214/aos/1176342503.

33. Abu-Shawiesh, M.O., "A simple robust control chart based on mad", ***Journal of Mathematics and Statistics***, Vol. 4, No. 2, (2008), 102. doi: 10.3844/jmssp.2008.102.107.

34. Shahriari, H., Ahmadi, O. and Shokouhi, A.H., "A two-step robust estimation of the process mean using m-estimator", ***Journal of Applied Statistics***, Vol. 38, No. 6, (2011), 1289-1301. https://doi.org/10.1080/02664763.2010.498502

35. Yeh, A.B., Huwang, L. and Li, Y.-M., "Profile monitoring for a binary response", ***IIE Transactions***, Vol. 41, No. 11, (2009), 931-941. https://doi.org/10.1080/07408170902735400

36. Amiri, A., Jensen, W.A. and Kazemzadeh, R.B., "A case study on monitoring polynomial profiles in the automotive industry", ***Quality and Reliability Engineering International***, Vol. 26, No. 5, (2010), 509-520. doi: 10.1002/qre.1127.

Persian Abstract

چکیده

بسیاری از مسائل صرفاً دارای یک یا چند متغیر نیستند که بتوانند مشخصه های کیفیت را تعیین کنند. در این مواقع، به عنوان یک راه حل، یک پروفایل با پیوند دادن متغیرهای مستقل به متغیر پاسخ معرفی می‌شود. یکی از مفروضات رایج در اکثر رویه های کنترلی، فرض مستقل بودن باقیمانده‌ها است. نقض این فرض می تواند منجر به نتایج گمراه کننده در نمودار کنترل شود. از سوی دیگر، زمانی که داده ها آلوده هستند، روش های کلاسیک تخمین پارامترهای عملکرد خوبی ندارند. چنین شرایطی نیازمند روش های برآورد قوی است. از این رو، این مقاله یک روش قوی برای تخمین پارامترهای فرآیند برای نظارت بر نمپ روفایل‌های خطی چندگانه خود همبسته برای فاز I پیشنهاد می‌کند. نمودار کنترل توسعه‌یافته در غیاب و حضور داده‌های آلوده از طریق مطالعات شبیه‌سازی جامع ارزیابی می‌شود. نتایج شبیه سازی‌های گسترده، نشان داد که برآوردگر قوی تأثیر داده‌های آلوده را بر عملکرد نمودار کنترل پیشنهادی برای همه درصدهای دورافتاده و مقادیر مختلف تغییرات کاهش می‌دهد. به طور کلی، در هر سه سناریو، از جمله نقاط پرت در پارامترهای مدل و واریانس خطا، رویکرد قوی بهتر از روش کلاسیک عمل می کند.

International Journal of Engineering

Journal Homepage: www.ije.ir

Multimodal Spatiotemporal Feature Map for Dynamic Gesture Recognition from Real Time Video Sequences

S. Reddy P., C. Santhosh*

Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur AP, India

PAPER INFO

Paper history:
Received 21 March 2023
Received in revised form 15 May 2023
Accepted 16 May 2023

Keywords:
2D Video Data
3D Video Data
Human Action Recognition
Visual Saliency
Deep Learning

ABSTRACT

The utilization of artificial intelligence and computer vision has been extensively explored in the context of human activity and behavior recognition. Numerous researchers have investigated and suggested various techniques for human action recognition (HAR) to accurately identify actions from real-time videos. Among these techniques, convolutional neural networks (CNNs) have emerged as the most effective and widely used for activity recognition. This work primarily focuses on the significance of spatial information in activity/action classification. To identify human actions and behaviors from large video datasets, this paper proposes a two-stream spatial CNN approach. One stream, based on RGB data, is fed with the spatial information from unprocessed RGB frames. The second stream is powered by graph-based visual saliency maps generated by GBVS (Graph-Based Visual Saliency) method. The outputs of the two spatial streams were combined using sum, max, average, and product feature fusion techniques. The proposed method is evaluated on well-known benchmark human action datasets, such as KTH, UCF101, HMDB51, NTU RGB-D, and G3D, to assess its performance Promising recognition rates were observed on all datasets.

doi: *10.5829/ije.2023.36.08b.04*

1. INTRODUCTION

Human action recognition from 2D RGB videos has been an active area of research, and various solutions have been proposed, ranging from feature-based approaches to machine learning-based models [1]. However, most of these solutions have primarily focused on basic human movements like walking, jumping, and crawling, and the datasets used have been limited to molecular images, 2D videos (offline and/or online), and 3D depth videos. While the 3D depth and 3D mocap data are popular. They are not practical for real-time use, which raises concerns about their feasibility. Consequently, our research is focused on using 2D videos from lab-captured and online action datasets, and we are exploring the use of feature extraction from 2D videos for identifying human actions.

Classifiers such as Support Vector Machine (SVM), graph matching (GM) [2], adaptive graph matching (AGM) [3], adaptive kernel matching (AKM) [4], artificial neural networks (ANNs), Hidden Markov models (HMM), and Convolutional neural networks (CNN) have been employed for modelling spatial, temporal and concatenation of both in RGB videos. However, the impact of CNN's and their derivatives on action recognition models utilizing 2D video data has been significant [5]. Abaei Kashan et al. [6] investigated the strengths of two descriptors, namely local binary pattern (LBP) and Histogram of Oriented Gradient (HOG), to extract local characteristics. The considered descriptors are widely used in computer vision applications. Azimi et al. [7] proposd method for fully automated image segmentation, which involves layer segmentation and the use of a fully convolutional network (FCN) for task-specific segmentation. CNNs are not only capable of recognizing efficiently but also have ability of extracting a set of actions from an input video sequence when provided with the ideal filter length on every convolution layer [8]. Trained human beings are capable of performing highly complex action sequences, which poses a challenge for extracting a multitude of

*Corresponding Author Email: csanthosh@kluniversity.in
(C. Santhosh)

Please cite this article as: S. Reddy P., C. Santhosh, Multimodal Spatiotemporal Feature Map for Dynamic Gesture Recognition from Real Time Video Sequences, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1440-1448

variations in such videos captured during live performances. To accurately interpret complex poses, it has become essential to extract spatial and temporal data with precision. In this work the human action recognition from video sequences are identified by the spatial information.Conv-Nets are a powerful tool to recognize human actions from 2D video sequences [9]. Numerous neural networks have been proposed for this task, many of which are regarded as state-of-the-art in terms of performance [10]. However, some of these networks suffer from drawbacks such as overfitting, inadequate data representation, and suboptimal selection of network architecture [11]. The other challenges include the selection of video frames per class, which in this case, changes depending on the actor performance and the unconstraint nature of videos captured using different imaging sensors.

We propose a multi stream CNN architecture with two streams to address the above challenges effectively. The two streams extract spatial information for classification. The first spatial stream is fed with RGB video frames and the other with graph based visual saliency maps of the action sequences.

Organization of manuscript is as follows: section 1 introduces human recongntion models with the necessity for 3D and 2D video sequences and section 2 deals about the layers on CNN. Whereas the dual stream CNN architecture was illustrated in the section 3 which follows the network architecture and training. Section 5 deals with the results and discussion followed by conclusion and future scope section.

1. 1. Literature Survey

Convolution parameters optimization for CNNs, referred as CPOCNN which assigns adaptive upper-bounds of convolution parameters depends on data dimension in current layer and number of remained layers to reach the output layer is propsed by Chegeni et al. [12]. Several fusion methods have been proposed to combine the features extracted from the spatial and temporal streams. The most common methods are early fusion, late fusion, and hybrid fusion. Early fusion combines the features from the two streams at the input level. Late fusion, on the other hand, combines the scores obtained from the two streams at the output level [13]. Hybrid fusion combines both early and late fusion methods to combine the features from the two streams. Recently, several studies have proposed novel methods to improve the performance of the two-stream CNN approach [14]. Scherer et al. [13] proposed a Spatial Temporal Inception module that combines the spatial and temporal streams at different levels of the network. This method has shown to improve the performance of the two-stream CNN approach on the UCF101 [15] and HMDB51 [9] datasets. Liu et al. [16] proposed a Multi-Scale Temporal Attention module that learns the importance of different temporal scales of the video. This method has shown to improve the performance of the two-stream CNN approach on the UCF101, HMDB51, and Kinetics datasets.

Spatio-Temporal ConvNet is a deep learning architecture that processes both spatial and temporal information to classify the action. This architecture consists of 3D convolutional layers that capture the temporal dynamics of the video [17]. Two-Stream ConvNet is an architecture that processes spatial and temporal information separately. The spatial stream processes the raw frames of the video and extracts spatial features. The temporal stream processes the optical flow, which represents the motion information between consecutive frames, to extract temporal features [18]. The two streams are then fused to classify the action. Two-Stream ConvNet has shown to improve the performance of action recognition compared to using only one stream. Long-Term Temporal Convolutions is a technique that improves the performance of action recognition by capturing long-term temporal information. This technique consists of 1D convolutional layers that operate on the temporal dimension of the video [14].

Transfer Learning is a research problem in machine learning that retainsknowledge obtained from the solution of a problem (source domain) to be applied to different but relatively similar problems (target domain). TL has been the mostpopular approach in CNN models in recent years [19]. The 1D convolutional layers have a larger kernel size than the typical 3x3 kernel used in spatio-temporal convolutions, which allows them to capture longer temporal information [20]. Long-Term Temporal Convolutions have shown to improve the performance of action recognition on datasets that involve long-term action [21]. Temporal Segment Networks (TSN) that aggregates features from multiple segments of the video. TSN uses a multi-scale architecture that processes the video at different temporal scales to capture both short-term and long-term temporal information. TSN has shown to achieve state-of-the-art performance on the Kinetics dataset [22].

2. LAYERS OF CNN

The popularity of CNN has increased due to its ability to process large amounts of data. The convolutional layer is the most crucial component of CNN, where the input is convolved using convolution kernels. These kernels act as filters and are followed by a non-linear activation, as defined in Equation (1). This enables CNN to identify distinct representations in speech or image data:

$$act_{i,j} = f\left(\sum_{k=1}^{K} \sum_{l=1}^{L} w_{k,l} \cdot x_{i+k,j+l} + b\right) \quad (1)$$

where, act(i,j) is the respective activation, the weight matrix of the kernel is denoted with w (k,l) of size k×l.

A small bias value b is added and it passed through a non-linear activation f [23].

Rectified linear units (ReLUs) nonlinear function is shown in Equation (2) and is utilized in the convolutional layers to generate the feature maps.

$$\sigma(x) = \text{maximum } (0,x) \quad (2)$$

In general, more hidden features in the input samples can be extracted as the more convolution kernels are included. In contrast to the regular CNN, the model discussed in this paper substituted a global average pooling (GAP) layer for the fully-connected layer that was previously placed behind the convolutional layer. CNN typically ends with single or multiple fully-connected layers that may transform multi-dimensional feature maps into one-dimensional feature vectors. Since each node present in the fully connected layer is linked to a node in the top layer, the fully connected layer's weight parameters may take up the greatest space. The GAP layer implements a global averaging pooling operation on every feature map, in contrast to the fully-connected layer. The GAP layer has no parameters that can be optimized.

During the training process of a neural network, the weight parameters in the top layer continuously change, leading to a continuous shift in the input data distribution for each layer. This dynamic change in distribution poses a challenge for network training and can impede convergence. To address this issue, a batch normalization layer (BN) is introduced after the Global Average Pooling (GAP) layer. The purpose of the BN layer is to adaptively modify the weight parameters to accommodate the evolving data distribution, thereby aiding in the faster convergence of the model.

The BN layer normalizes and reconstructs the input data on each batch of training samples in order to ensure the stability of the output from the previous layer and to improve the speed and accuracy of training.

A Softmax layer as a classifier plus a fully connected layer make up a output layer of CNN [24]. The fully-connected layer should be added at the end of the model because it has a significant advantage. Each node of the fully connected layer is linked to the nodes of the top layer in order to integrate the features that were extracted from the upper layer. In this way, it compensates for the GAP layer's drawbacks.

The Softmax sits underneath the fully connected layer and it transforms the output of the top layer into a probability vector whose value indicates the max likelihood that the current sample belongs to each class. The output of Softmax is given Equation (3).

$$Soft_i = \frac{e^{y_i}}{\sum_{c=1}^{C} e^{y_c}} \quad (3)$$

where, y is the fully connected layer output, C is the number of classes considered.

3. DUAL – STREAM CNN ARCHITECTURE

An action video sequence comprises both spatial and temporal variations observed across a collection of video frames. The spatial component represents the discrete appearances of objects within the frames, while the temporal component captures the movements exhibited by these objects over time. In this work, the action recognition Convolutional Neural Network (CNN) is designed with two spatial streams, as depicted in Figure 1. These streams are constructed using Conv-Nets with SoftMax layers, and late fusion models are employed to calculate similarity scores. Four late fusion models, namely averaging, maximum, product, and sum, are considered in this study.

The architecture comprises of two streams, each containing 8 convolutional layers, two dense layers, and a SoftMax layer. In order to combine the outputs from the SoftMax layers of both streams, a score fusion model utilizing multiple fusion models has been proposed. The proposed architecture is faster and accurate in identifying complex human actions from videos. The model is investigated on multiple action datasets for checking its persistence to multiple types of input video sequences. Robustness of the proposed model is contemplated against various other CNN architectures to ascertain its usefulness in detecting complex actions.

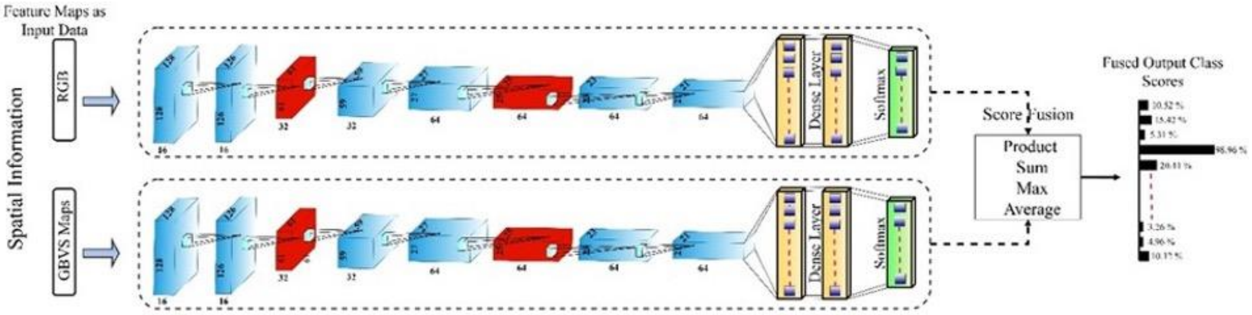


Figure 1. Proposed dual stream architecture

Previous CNN models use RGB sequence as a spatial stream. The main shortcoming is observed that the RGB data in the video sequence poorly represented due to low light indoor environment during a real action performance. To outplay the above shortcoming, we propose to add one more stream which is insensitive to color brightness components in a video sequence.

However, the spatial stream gets graph based visual saliency (GBVS) representing spatial distribution of the action. For the purpose of human action recognition, one of the stream is fed with the GBVS feature maps. The RGB spatial information is totally inconsistent in action sequences captured during a real time performance. The inconsistency appears due to poor lighting and indistinguishable background. To balance this effect, spatial saliency maps are added as an additional stream for the existing RGB stream [23].

To demonstrate the merits of the 2-stream type coding by training a CNN, later the trained CNN is tested to analyze the performance, view invariance and check robustness. We utilized five diverse publicly accessed datasets, KTH , UCF101 [15], HMDB51 [9], G3D [24] and NTU RGB-D [25] for examining the proposed model. The next section briefly enlightens the proposed work and its performance in recognizing human actions from video sequences by performing experiments as an individual streams and 2-streams.

3. 1. Spatial stream-1: RGB stream Conv-Net The proposed architecture incorporates two spatial streams. In the first spatial stream, the RGB frames of an action video sequence are directly utilized. These action sequences are sourced from online datasets, which encompass videos captured under diverse conditions, including complex and simple backgrounds, variations in lighting conditions, and instances of object occlusions.

Figure 2 shows a set of action video frames captured in the wild, which is part of online available datasets.

The frames show a multitude of variations in the action poses with uncontrollable effects from lighting, erratic movements, object occlusion and background variations. The color plays a vital component in identifying a dancer from the background. The first spatial stream is a set of frames in each action class to identify the static pose based on color, texture and shape of the action. Due to large distractions in the data, we propose to double the spatial information model by inducing a second spatial stream based on feature maps extracted from saliency maps.

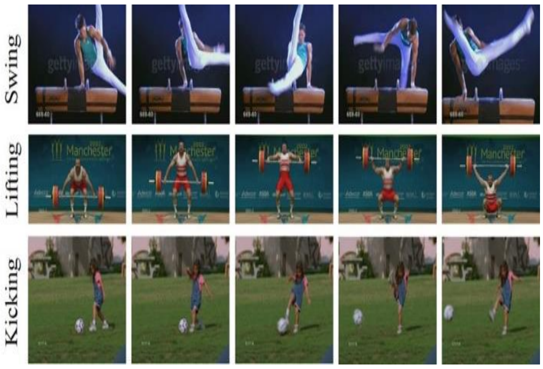


Figure 2. Real time action performances from online datasets

3. 2. Spatial stream-2 : GBVS stream Conv-Net Graph based visual saliency algorithm is applied to extract the saliency maps from the RGB video frames using graph cut based model. The algorithm was used by Kishore et al. [26] generated the visual saliency maps of video objects. Figure 3 shows the saliency maps created for the sample action video frames from the online datasets.

The saliency maps describe the spatial distribution of the action in the video frame. The second CNN stream takes the video frames shown in Figure 3. This stream of spatial contents is immune to variations such as lighting, color, and background as a reinforcement to the RGB spatial stream. These are a set of low-level features describing the action in an abstract manner.

Incorporating model based spatial representation using GBVS, greatly improves the end-to-end convolution based deep learning methods. When these streams are operated at a time, it also solves the problem of overfitting in the first stream with the weight vectors from the second stream.

4. CNN NETWORK ARCHITECTURE AND TRAINING

The proposed Convolution Neural Network is influenced from the VGG network developed by Simonyan and Zisserman [23] which is extremely deep CNN model that accomplished the state-of-the-art accuracy on Large Scale Visual Recognition Challenge, 2014 classification and localization tasks. VGG net is a densely layered CNN consisting of 16 to 19 weighted layers and a small window of 3×3 aligned along the entire convolution

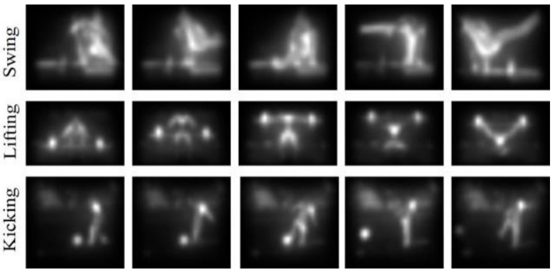


Figure 3. GBVS based saliency maps for action frames of Figure 2

layers. The proposed CNN model is more similar to the indigenous architectures by Ciresan et al. [27] and Dean et al. [28]. Although, the presented Conv-Net architecture is developed from the motivation of VGG net, the depth of weight layers is constrained to 8 and fully connected layers are bounded to 2. Python aided with Keras and TensorFlow libraries are utilized for the construction of this architecture.

The VGG is the origin of every individual stream in the Convnet architecture but limited to the 10 layers. After performing multiple tests using diverse network models such as VGG, Res-Net, Alex-Net and Inception it resulted into the 8 convolution layers preceded by 2 fully connected layers. These are built from the scratch and the overall development of model is done with Keras and TensorFlow in Python 3.6 platform. Additionally, for image classification tasks we trained few pre-trained networks.

4. 1. Dual-stream Conv-Net Training During experiments, the filter configuration of the designated layers varied linearly from 3×3 for initial two layers, 5×5 for the next two, 7×7 and 9×9 for the remaining 4 layers for all sets of training batches. The training algorithm is incepted from Dean et al. [28]. The multinomial logistic regression objective is optimized by training and with the aid of the mini-batch gradient descent when supplied with momentum of 0.9. To normalize the weight decay for 128 frame size at training period, the penalization multiplier is fixed to a range of 2 to 0.0002. Bicubic interpolation is utilized to resize all input frames and they been numbered according to the sequence in each class. For the initial 8 layers the drop out regularization is set to 0.5. When validation accuracy stabilized at a certain constant, the initial learning rate is positioned to 0.02 and declined by a factor of 10. The learning rate plummeted for three times as result it altered from 0.02 to 0.005 to 0.001 and the training has been paused after 10000 epochs i.e., 311.25k iterations and for every 1480 epochs i.e., 150k iterations the learning rate got plunging down. Whereas, when trained with alternate datasets consisting of 500 input size the training terminated at 4800 epochs i.e., 100.125k iterations and learning rate dropped down for every 1220 epochs (10.517k iterations).

The nets required less epochs due to medium scale frame size. In addition, the frame size was raised to the 224×224 in the anticipated labels with negligible or no advancement. In every layer the weights are assigned arbitrarily and the gaussian distribution function and variance are set to zero mean and 0.01 respectively for every layer. The filter outputs from 8 convolutional layers in both the streams are conceptualized in Figure 4.

4. 2. Training Batch Index In order to validate the proposed CNN architecture, we conducted experiments using several 2D action datasets, including KTH, UCF101, HMDB51, G3D, and NTU RGB-D. Each dataset consisted of 20 different classes with 50 diverse actions, providing a comprehensive evaluation of the developed model. For each dataset, we performed individual training to generate a specific model. Subsequently, the trained model was employed to test the actions within the respective dataset. In order to assess cross-data authentication, we also conducted training with one dataset and testing with another dataset. However, it was ensured that the actions being tested were common across all datasets. HMDB51 and NTU RGB-D datasets were particularly chosen due to their extensive collection of data in various views and diverse subjects. These datasets allowed us to produce comprehensive results and gain a clearer understanding of the proposed method. Throughout this study, the majority of the results were obtained from experiments conducted on these popular datasets, HMDB51 and NTU RGB-D, to provide a detailed analysis of the proposed approach.

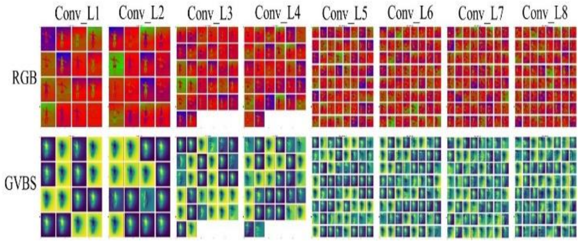


Figure 4. Feature maps visualization of various Convolutional layers

4. 3. Testing the Proposed Dual-stream Conv-Net Once the CNN model is trained, it can be used for testing by inputting a batch of 20 action videos. The testing process involves classifying the actions in the videos and assigning corresponding labels. In all the datasets, the 2D video samples have frame sizes of 128x128 pixels. The output of the SoftMax layer is a class score vector, where each element represents the likelihood of the input video belonging to a particular class. Since the proposed Conv-Net consists of two spatial streams, separate class scores are obtained for each stream. To combine the class scores from the two streams, a late fusion approach is applied. This fusion involves using four popular fusion models: maximum, average, sum, and product. Each fusion model generates a single score for each class based on the class scores from the two streams. The network is then tested using multiple experiments focused on human actions. Additionally, pre-trained networks can be utilized for testing by retraining them using available action videos from online datasets. This approach leverages the pre-trained weights and further fine-tunes the network on the specific action videos to improve performance and adapt to the task at hand.

5. EXPERIMENTATION RESULTS AND DISCUSSION

The execution is done with the aid of Keras and TensorFlow toolboxes which can be accessed in python 3.6 substantial adjustments during testing and training. The proposed method is tested on 20 classes of actions from KTH, UCF101, HMDB51, G3D and NTU RGB-D action datasets. Five training models had been produced for the five 5 datasets separately. Performance of each Conv-Net is for a particular dataset is validated with respect to percentage of identification on the complete testing dataset.

5. 1. Evaluation of the Proposed Conv-Net on 2D Action Datasets

In this section, the performance of the proposed Dual-stream architecture is evaluated using five publicly available 2D action datasets. Twenty classes with 50 action sequences from the KTH, UCF101, HMDB51, G3D, and NTU RGB-D datasets are selected and labeled appropriately. The CNN architecture remains consistent for each video in the datasets.Each video is segmented into 656 frames, with each frame having a size of 128x128 pixels. Cross-subject testing is conducted on the proposed CNN architecture using 20 test videos from each of the databases. The results of the testing are summarized in the form of a confusion matrix, which shows the classification accuracy and misclassifications for the five test classes. Figure 5 illustrates the confusion matrix, providing a visual representation of the performance of the proposed CNN architecture across the five datasets.

Table 1 gives recognition of the proposed CNN architectures as individual spatial stream and mixed dual-streams across view and subjects. The simulation shows that the proposed architecture utilizes the advantage of both spatial features.

The average fused scores in Table 1, point to the advantage of using multi stream networks compared to single stream. The dual-stream CNN model showed highest recognition rates compared to single stream models. This proves the universal fact in machine learning that wider heterogeneous trainings provide higher recognition.

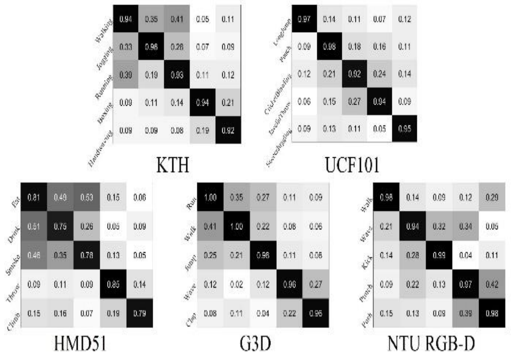


Figure 5. Confusion matrices obtained for few test samples from each dataset on the proposed 2 stream CNN

To visualize the effects of fusion on multi stream CNN architectures, four different score fusion models such as average, maximum, sum and product were tested. Table 2 summarizes the results of the experiment. Product fusion model has proved to produce good score fusions compared to sum, average and max score fusions.

5. 2. Evaluation of the Proposed Conv-Net against other Deep-Nets

The dataset is applied on various Conv-Nets such as VGG 16 [23], Alex-Net [29], ImageNet [30], RNN [31], LSTM [32]. At this stage, we study the influence of network model and pre-training on the performance of the task. For this, three cases are formulated: pretrained, pretrained + retrained and training from scratch. The results show training from scratch networks perform better compared to pretrained and pre + retrained. However, training from scratch is computationally intensive due to large data samples fed into the system.

The other model we tested is retraining the pretrained model with new data. Results are summarized in Table 3.

Recognition rates obtained from the state-of-the-art deep nets are compared against the proposed dual-stream architecture on action datasets. The results are averaged over the different testing instances including cross subject, cross view and cross datasets. The recognition rates are tabulated in Table 4.

TABLE 1. Comparison of recognition rates obtained in different individual streams and multi streams of proposed CNN

Dataset	Recognition Rates (%)		
	Spatial	Stream	Dual-stream
	RGB	GBVS	RGB + GBVS
KTH	85.37	83.47	92.99
UCF101	89.51	88.25	91.35
HMDB51	66.55	60.18	90.85
G3D	87.43	83.41	92.87
NTU RGB-D	90.29	85.51	94.81

TABLE 2. Testing the score fusion models in 2 Stream networks

Dataset	Score Fusion Method			
	Sum	Max	Average	Product
KTH	92.43	88.57	90.91	92.89
UCF101	95.02	90.12	92.46	91.35
HMDB51	75.63	68.25	70.59	90.45
G3D	89.26	87.58	88.92	92.87
NTU RGB-D	94.59	88.92	91.26	93.81

TABLE 3. Gives the recognition on human action datasets with various Deep Nets

Training Type	Architecture	Action Data				
		HMDB51	NTU RGB-D	KTH	UCF101	G3D
Pretrained	VGG 16	83.74	84.45	83.39	83.91	83.73
	AlexNet	77.91	81.75	81.05	79.73	81.08
	ImageNet	79.29	80.78	80.91	80.07	79.56
Pretrained + Retrained	VGG 16	87.13	88.01	87.17	86.93	86.92
	AlexNet	81.52	85.31	84.23	82.56	84.29
	ImageNet	82.85	84.34	82.91	83.49	81.75
Training from Scratch	RNN	85.25	86.29	86.04	85.72	85.82
	LSTM	85.81	87.21	86.29	85.97	86.01
	Ours	90.45	93.81	92.89	91.35	92.87

All architectures are realized with Keras and Tensorflow. The proposed CNN architecture gives highest recognition on both the action datasets, which is due to inclusion of multiple spatial streams for decision making. The above presented results show that the proposed dual-stream CNN gives consistent performance for each of the action with cross view and cross subject variations compared to single stream nets.

5. 3. Evaluation of the Proposed Conv-Net against other Multi Stream Deep-Nets

Finally, human action video data is inputted to various multi stream architectures and the average recognition rates were calculated with different late fusion rules.

Table 5 reports the results, showing the proposed multi stream architectures provide better recognition compared other deepNets in literature. The average recognition of our nets touched 94%.

TABLE 4. Comparison of recognition rates for action datasets with different deep learning architectures

Architecture	Action Datasets				
	HMDB51	NTU RGB-D	KTH	UCF101	G3D
VGG-VD16 [24]	84.73	89.72	87.46	85.36	87.16
AlexNet [27]	79.99	84.51	83.29	80.99	82.62
ImageNet [28]	80.52	82.96	81.17	81.05	80.94
CNN [9]	85.59	88.95	90.74	86.46	91.25
Proposed	**90.45**	**93.81**	**92.89**	**91.35**	**92.87**

TABLE 5. Performance of multi-stream nets on NTU RGB-D data

Technique	Feature sets	Recognition Rate (%)
Two-stream CNN (fusion-SVM) [9]	Opticalflow + RGB	88.30
Two-stream CNN (fusion-Averaging) [9]	Opticalflow + RGB	86.70
Spatio-temporal ConvNet (Slow Fusion) [17]	High resolution + Low resolution RGBs	89.60
Two-stream ConvNet + LSTM [18]	Opticalflow + Raw RGB frames	88.80
Long-term temporal convolutions (LTC) [21]	MPEG flow + RGB	91.60
DSSCA-SSLM [32]	Depth Maps + RGB	74.56
c-ConvNet [22]	Depth Maps + RGB	89.09
Proposed (Spatial Stream-1)	**RGB**	**85.06**
Proposed (Spatial Stream-2)	**GBVS**	**86.13**
Proposed (Spatial dual-Stream)	**RGB + GBVS**	**94.81**

This is due to the unique architecture having multiple streams in spatial features. Each stream identifies a set of features which are locally crafted by the uniqueness of that algorithm. RGB frame gives static distribution of color brightness in the spatial domain which is supplemented with object action distribution in space to enhance the effect of filters to identify a complex human action correctly.

6. CONCLUSIONS AND FUTURE SCOPE

This work proposed and presented a dual-stream architecture for recognizing complex human actions from 2D videos sequences. The proposed two-stream ConvNets separate spatial information in videos with two streams for each of the spatial data. The two spatial streams are RGB action frames and GVBS based spatial saliency maps. The outputs of 2 SoftMax layers are score fused to generate similarity score. The results showed that the proposed multi streams can recognize human actions accurately and are robust to changing backgrounds in unconstraint videos. Extensive testing proves that the proposed two-stream ConvNets can handle a variety of 2D video data with ease producing consistent outcomes. The average recognition rate on the entire datasets for the proposed two-stream ConvNet is around 94.81%.The dual-stream architecture for human activity recognition has the potential to be further developed and improved to recognize more complex activities. Incorporating additional modalities, attention mechanisms, multi-task learning, and transfer learning can all help to improve the performance of the model for recognizing complex human activities.

7. REFERENCES

1. Afsar, P., Cortez, P. and Santos, H., "Automatic visual detection of human behavior: A review from 2000 to 2014", ***Expert Systems with Applications***, Vol. 42, No. 20, (2015), 6935-6956. https://doi.org/10.1016/j.eswa.2015.05.023
2. Zhou, F. and De la Torre, F., "Factorized graph matching", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 38, No. 9, (2015), 1774-1789. doi: 10.1109/TPAMI.2015.2501802.
3. Yang, X. and Liu, Z.-Y., "Adaptive graph matching", ***IEEE Transactions on Cybernetics***, Vol. 48, No. 5, (2017), 1432-1445. doi: 10.1109/TPAMI.2015.2501802.
4. Popovici, V. and Thiran, J., "Adaptive kernel matching pursuit for pattern classification", in Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, Acta Press. (2004), 235-239.
5. Liu, H., Ju, Z., Ji, X., Chan, C.S., Khoury, M., Liu, H., Ju, Z., Ji, X., Chan, C.S. and Khoury, M., "A view-invariant action recognition based on multi-view space hidden markov models", ***Human Motion Sensing and Recognition: A Fuzzy Qualitative Approach***, (2017), 251-267. doi: 10.1109/TPAMI.2015.2501802.
6. Abaei Kashan, A., Maghsoudi, A., Shoeibi, N., Heidarzadeh, M. and Mirnia, K., "An automatic optic disk segmentation approach from retina of neonates via attention based deep network", ***International Journal of Engineering, Transactions A: Basics***, Vol. 35, No. 4, (2022), 715-724. doi: 10.5829/IJE.2022.35.04A.11.
7. Azimi, B., Rashno, A. and Fadaei, S., "Fully convolutional networks for fluid segmentation in retina images", in 2020 International Conference on Machine Vision and Image Processing (MVIP), IEEE. (2020), 1-7.
8. Srihari, D., Kishore, P., Kumar, E.K., Kumar, D.A., Kumar, M.T.K., Prasad, M. and Prasad, C.R., "A four-stream convnet based on spatial and depth flow for human action classification using rgb-d data", ***Multimedia Tools and Applications***, Vol. 79, (2020), 11723-11746. doi: 10.1109/TPAMI.2015.2501802.
9. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T., "Hmdb: A large video database for human motion recognition", in 2011 International conference on computer vision, IEEE. (2011), 2556-2563.
10. Längkvist, M., Karlsson, L. and Loutfi, A., "A review of unsupervised feature learning and deep learning for time-series modeling", ***Pattern Recognition Letters***, Vol. 42, No., (2014), 11-24. doi: 10.1109/TPAMI.2015.2501802.
11. Simonyan, K. and Zisserman, A., "Two-stream convolutional networks for action recognition in videos", ***Advances in Neural Information Processing Systems***, Vol. 27, (2014).
12. Chegeni, M.K., Rashno, A. and Fadaei, S., "Convolution-layer parameters optimization in convolutional neural networks", ***Knowledge-Based Systems***, Vol. 261, (2023), 110210. https://doi.org/10.1016/j.knosys.2022.110210
13. Scherer, M., Magno, M., Erb, J., Mayer, P., Eggimann, M. and Benini, L., "Tinyradarnn: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars", ***IEEE Internet of Things Journal***, Vol. 8, No. 13, (2021), 10336-10346. https://doi.org/10.1162/neco.1997.9.8.1735
14. Savadi Hosseini, M. and Ghaderi, F., "A hybrid deep learning architecture using 3d cnns and grus for human action recognition", ***International Journal of Engineering, Transactions B: Applications***, Vol. 33, No. 5, (2020), 959-965. doi: 10.5829/ije.2020.33.05b.29.
15. Soomro, K., Zamir, A.R. and Shah, M., "Ucf101: A dataset of 101 human actions classes from videos in the wild", arXiv preprint arXiv:1212.0402, (2012).
16. Liu, H., Zhou, A., Dong, Z., Sun, Y., Zhang, J., Liu, L., Ma, H., Liu, J. and Yang, N., "M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar", ***IEEE Internet of Things Journal***, Vol. 9, No. 5, (2021), 3397-3415. https://doi.org/10.1162/neco.1997.9.8.1735
17. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., "Large-scale video classification with convolutional neural networks", in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2014), 1725-1732.
18. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G., "Beyond short snippets: Deep networks for video classification", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2015), 4694-4702.
19. Zohrevand, A., Imani, Z. and Ezoji, M., "Deep convolutional neural network for finger-knuckle-print recognition", ***International Journal of Engineering, Transactions A: Basics***,

Vol. 34, No. 7, (2021), 1684-1693. doi: 10.5829/IJE.2021.34.07A.12

20. Parvez M, M., Shanmugam, J., Sangeetha, M. and Ghali, V., "Coded thermal wave imaging based defect detection in composites using neural networks", ***International Journal of Engineering, Transactions A: Basics***, Vol. 35, No. 1, (2022), 93-101. doi: 10.5829/ije.2022.35.01A.08.

21. Varol, G., Laptev, I. and Schmid, C., "Long-term temporal convolutions for action recognition", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 40, No. 6, (2017), 1510-1517. https://doi.org/10.1162/neco.1997.9.8.1735

22. Wang, P., Li, W., Wan, J., Ogunbona, P. and Liu, X., "Cooperative training of deep aggregation networks for rgb-d action recognition", in Proceedings of the AAAI conference on artificial intelligence. Vol. 32, (2018).

23. Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, (2014). doi: 10.1109/TPAMI.2015.2501802.

24. Auli, M., Galley, M., Quirk, C. and Zweig, G., "Joint language and translation modeling with recurrent neural networks", in Proc. of EMNLP. (2013).

25. Bloom, V., Makris, D. and Argyriou, V., "G3d: A gaming action dataset and real time action recognition evaluation framework", in 2012 IEEE Computer society conference on computer vision and pattern recognition workshops, IEEE. (2012), 7-12.

26. Kishore, P., Kumar, D.A., Sastry, A.C.S. and Kumar, E.K., "Motionlets matching with adaptive kernels for 3-d indian sign language recognition", ***IEEE Sensors Journal***, Vol. 18, No. 8, (2018), 3327-3337. doi: 10.5591/978-1-57735-516-8/IJCAI11-210.

27. Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M. and Schmidhuber, J., "Flexible, high performance convolutional neural networks for image classification", in Twenty-second international joint conference on artificial intelligence, Citeseer. (2011).

28. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M.a., Senior, A., Tucker, P. and Yang, K., "Large scale distributed deep networks", ***Advances in Neural Information Processing Systems***, Vol. 25, (2012).

29. Girshick, R., Donahue, J., Darrell, T. and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2014), 580-587.

30. Shahroudy, A., Liu, J., Ng, T.-T. and Wang, G., "Ntu rgb+ d: A large scale dataset for 3d human activity analysis", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2016), 1010-1019.

31. Hochreiter, S. and Schmidhuber, J., "Long short-term memory", ***Neural Computation***, Vol. 9, No. 8, (1997), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

32. Shahroudy, A., Ng, T.-T., Gong, Y. and Wang, G., "Deep multimodal feature analysis for action recognition in rgb+ d videos", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 40, No. 5, (2017), 1045-1058. doi: 10.1109/TPAMI.2017.2691321.

Persian Abstract

چکیده

استفاده از هوش مصنوعی و بینایی کامپیوتری به طور گسترده در زمینه فعالیت های انسانی و تشخیص رفتار مورد بررسی قرار گرفته است. بسیاری از محققان تکنیک های مختلفی را برای تشخیص اقدامات انسانی (HAR)برای شناسایی دقیق اقدامات از ویدیوهای بلادرنگ بررسی و پیشنهاد کرده اند. در میان این تکنیک‌ها، شبکه‌های عصبی کانولوشنال (CNN)به عنوان مؤثرترین و پرکاربردترین شبکه‌های عصبی برای تشخیص فعالیت ظاهر شده‌اند. این کار در درجه اول بر اهمیت اطلاعات مکانی در طبقه بندی فعالیت/عمل متمرکز است. برای شناسایی اعمال و رفتارهای انسانی از مجموعه داده‌های ویدیویی بزرگ، این مقاله یک رویکرد فضایی CNN دو جریانی را پیشنهاد می‌کند. یک جریان، بر اساس داده های RGB، با اطلاعات مکانی از فریم های RGB پردازش نشده تغذیه می شود. جریان دوم توسط نقشه های برجستگی بصری مبتنی بر نمودار ایجاد شده توسط روش (GBVSبرجستگی بصری مبتنی بر نمودار) طراحی شده است. خروجی‌های دو جریان فضایی با استفاده از تکنیک‌های مجموع، حداکثر، میانگین و ترکیب ویژگی محصول ترکیب شدند. روش پیشنهادی بر روی مجموعه داده‌های عملکرد انسانی معیار شناخته شده، مانند KTH، UCF101، HMDB51، NTU RGB-D، و G3D ارزیابی می‌شود تا عملکرد آن ارزیابی شود نرخ‌های تشخیص امیدوارکننده‌ای در همه مجموعه‌های داده مشاهده شد.

International Journal of Engineering

Journal Homepage: www.ije.ir

Development and Calibration of an Efficiency Factor Model for Recycled Aggregate Concrete Struts

A. D. Chaudhari*, S. Suryawanshi

Department of Civil Engineering, S. V. National Institute of Technology, Surat, Gujarat, India

PAPER INFO

Paper history:
Received 22 March 2023
Received in revised form 22 April 2023
Accepted 23 April 2023

Keywords:
Bottle-shaped Strut
Strut-and-tie Model
Deep Beam
Recycled Concrete Aggregate

ABSTRACT

In the strut-and-tie (STM) method of design, the internal mechanism of flow of forces is represented by hypothetical truss in which the behavior of the beam is controlled by the strut connecting load and support points. The strength of such strut is correlated to the shear capacity of the deep beam through a factor called the strut efficiency factor. Different efficiency factor models have been recommended by various internationally accepted codes. However, none of the codes takes into account the effect of recycled aggregates in concrete. Although some codes yield conservative results, these predictions are not sensitive enough to the recycled aggregate content. Therefore, an efficiency factor model sensitive to recycled aggregate concrete and easy to operate is much desired. In this work, published results of laboratory tests on deep beam specimens made of concrete consisting of recycled aggregates were considered for the analysis, employing a suitable strut-and-tie model. All these deep beams were originally designed by sectional or empirical method. Based on regression analysis of the outcomes of the STM analysis, an efficiency factor model has been proposed which takes into account the effect of recycled aggregates in concrete. Subsequently, scaled deep beam specimens containing recycled aggregate concrete were cast and tested in the laboratory in order to calibrate the proposed strut efficiency factor model. The yield of proposed efficiency factor model was compared with the predictions of the selected internationally accepted code provisions. It is found that the predictions of proposed efficiency factor model give consistent and comparable results.

***doi**: 10.5829/ije.2023.36.08b.05*

1. INTRODUCTION

A strut-and-tie model (STM) is a method used in structural engineering to analyze and design reinforced concrete structures, especially for structural members containing D-regions such as corbels, beam-column joints, deep beams, pile caps, etc. [1-4]. Theoretically, STM is a lower bound method in which the mechanism of load transfer is represented by a set of struts and ties attached with node under the condition of plane stress. The capacity of the elements, such as struts and ties, of STM is then calculated, taking into account equilibrium and constitutive relations. The strut connecting load point and support point, hereafter called the bottle-shaped strut, plays a key role in the failure mechanism of deep beams. While transferring the load, due to direct compression between load and support point indirect tension is generated which reduces the strength of such strut. To represent this reduction in strength, the coefficient, or factor, 'β_s' is applied to the strut. Various codes name this factor as follows: ACI 318-14 [5] defines it as strut coefficient, Eurocode 2 [6] describes it as strength reduction factor, JSCE [7] guidelines simply name it reduction factor, and AS-3600 [8] expresses it as strut efficiency factor. Although different codes assign different nomenclature to this factor, in this work it is termed the strut efficiency factor and used in the forthcoming description. In general, the crushing strength of a concrete strut is referred to as its effective strength and is given by the following formula (Equation (1)):

$$f_{cu} = \beta_s f_c' \qquad (1)$$

where, β_s is an efficiency factor having a value between 0 to 1, f_{cu} is the effective concrete compressive strength

*Corresponding Author Institutional Email: ac2565@gmail.com (A. D. Chaudhari)

Please cite this article as: A. D. Chaudhari, S. Suryawanshi, Development and Calibration of an Efficiency Factor Model for Recycled Aggregate Concrete Struts, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1449-1458

in the strut (as per ACI 318-14 [5]), and f'_c is the cylinder compressive strength of concrete. Various sources in the literature recommend differing values of strut efficiency factors, with perhaps the simplest recommendations being those of ACI 318-14, wherein the nominal compressive strength of concrete in the strut is pre-multiplied by an efficiency factor varying between 0.6 and 1. Limiting concrete compressive stresses in struts specified by selected design codes is presented in Table 1.

For suitability and adaptability with respect to the changing environment, industrial wastes including construction and demolition (C&D) waste have been recognized as a possible source to substitute various ingredients of conventional cement concrete. For example, GGBFS, fly ash (FA) and other ashes [9] and silica fume to partly replace an OPC, GGBFS to replace FA and extracts of C&D to replace aggregates in new concrete [10, 11]. Utilizing recycled concrete aggregate (RCA) extracted from waste concrete for producing new concrete has some economic and environmental benefits, as the aggregates occupy 60 to 75% volume of the concrete mixture. Even with these advantages, it hasn't been extensively adopted by the construction industry, especially for structural applications. The literature review reveals that the majority of the investigations concentrated on the processing, characterization, rheology of RAC [12, 13] along with NAC [14] or on the physical and mechanical properties of concrete made with such aggregates [15]. The focus on structural application of this concrete has been more recent [16, 17]. Even the structural performance of RCA concrete under various actions has not been comprehensively investigated. Most of the reported studies focused on flexural behaviour with a few on the behaviour of shear-critical elements like corbels, beam-column joints, deep beams, pile caps, etc. Structural action in such members is governed by shear, and the internal forces can be conveniently represented by strut-and-tie models. As discussed in the preceding paragraph, struts are the compression members in STM, and the literature review indicates that bottle-shaped struts are particularly susceptible to splitting failure [18, 19]. The use of such relatively soft and porous recycled concrete aggregates in concrete is likely to raise concerns about safety and serviceability. Code provisions to design such critical struts are either empirical or not robust enough. The efficiency factor is a critical parameter for the design of bottle-shaped struts, and most of the available efficiency factor models in the literature are limited in scope and account for the effects of a very narrow range of parameters. None of the recommendations in the literature for strut-and-tie modelling is calibrated for application to recycled aggregate concrete.

TABLE 1. Comparison of code provisions on recommended strut efficiency factors

Code Name	$\beta_{s,code}$	Required Transverse or Web Reinforcement
ACI 318-14 [5]	Unreinforced bottle-shaped strut: 0.60	--
	Reinforced bottle-shaped strut: 0.75	• $\rho_T = 0.003$ (For $f'_c \leq 6000$ psi) • $\sum\left(\frac{A_{si}}{b_s s_i}\sin\alpha_i\right) \geq 0.003$ (For $f'_c > 6000$)
AASHTO [20]	$\frac{1}{(0.8+170\,\varepsilon_1)} \leq 0.85$	Orthogonal grid of reinforcement bars near each face, i.e., $\rho_H \geq 0.003$ or $\rho_V \geq 0.003$
Eurocode 2 [6]	$0.6\left(1-\frac{f_{ck}}{250}\right)$	• Partial discontinuity $T = \frac{1}{4}\left(\frac{b-a}{b}\right)F$ • Full discontinuity $T = \frac{1}{4}\left(1-0.7\frac{a}{h}\right)F$ • $T = F_{sc}\,A_{st}$
AS-3600 [8]	$\frac{1}{1.0+0.66\cot^2\alpha_s}$	Transverse reinforcement required to resist design bursting force in accordance with Clause 7.2.4 of AS-3600
Present Study	$\left(\frac{f_t}{12+0.002R}\right)$	--

It is thus obvious that, the use of the efficiency factor in the design equations of STM is a simplified approach and doesn't take into account all the factors affecting the strut capacity. More specifically, it does not consider the effect of modified properties of the concrete when alternate materials such as recycled aggregates, are used in place of conventional ingredient since the provisions for β_s are originally made for NA-concrete. Therefore, concerns have been raised about the applicability of current code provisions (such as ACI 318-14 [5], AASHTO [20], Eurocode 2 [6], and AS-3600 [8]) for RCA-concrete [21]. In the present investigation, a mathematical model for β_s is developed by performing regression analysis on a database of 123 RCA-concrete deep beam specimens extracted from the literature. This proposed model is a function of compressive stress, 'f'_c', replacement percentage of NA with RCA, 'R', and is capable of estimating the shear capacity of RCA-concrete specimens in general and the capacity of critical struts like the bottle-shaped strut in specific. Further, the modified form of the model in terms of tensile strength 'f_{ct}' is also discussed. Besides the validation of the test results of the beam specimens reported in the literature, the proposed efficiency factor model is calibrated by testing deep beams containing recycled aggregate concrete.

2. STM ANALYSIS OF SELECTED BEAM SPECIMENS

The specimens satisfying the deep beam criteria of ACI 318-14 [5] were considered for this investigation. All the selected deep beam specimens which were originally designed using empirical equations or by the sectional method are reanalyzed by applying STM. A suitable strut-and-tie model was superimposed on the geometry of selected beam specimens in order to carry out STM analysis. Further, the capacity of the critical strut was calculated, which was subsequently used to estimate the shear capacity of the beams. A total of 123 beam specimens were filtered out from the database. The selection of 123 beams was made on the basis of the qualifying condition that the beam specimens should be composed of concrete containing aggregates partly or fully replaced by recycled aggregates. The beam specimens tested and reported by Choi et al. [22], Han and Chung [23], Singh et al. [24], Fathifazl et al. [25], Kim et al. [26], Etman et al. [27], Aly et al. [28], Al-Zahraa et al. [29], Lian et al. [30], Arabiyat et al. [31] and Li et al. [32] have considered for this study. All the specimens have the limits of shear span to depth ratio of 0.54 to 2.50, a compressive stress from 16.7 to 58.60, and a replacement level ranging from 0% to 100%.

A typical beam (RAC30-H1.5) tested by Choi et al. [22] under a four-point bending test is considered to illustrate the STM analysis procedure. The beam specimen was 1840 mm long, 400 mm deep, and 200 mm wide, with an effective span of 1440 mm, as shown in Figure 1. To carry out analysis, the suitable STM is superimposed on the geometry of selected beam specimens.

Once the strut-and-tie model to describe the flow of forces in a beam was assigned, the required dimensions were easily determined. The node dimensions were assigned to find out strut width at the strut and the node interfaces at both ends, and the minimum value of the two was considered as the width of the bottle-shaped strut, w_s. Next, the effective transverse reinforcement ratio was determined from the provided web reinforcement in the form of either stirrups or an orthogonal grid. With the use of equilibrium equations, support reactions are found as usual. The theoretical capacity of the strut is estimated as $C_{Th} = \beta_s.b.w_{s,min}.f'_c$. Thereafter, the truss model was solved by applying conditions of equilibrium to determine compressive force (C_{expt}) in the critical strut using the relation $C_{expt} = F / \sin\alpha_s$. Where, F is the magnitude of the nearest associated support reaction. It should be noted that in the case of a four-point or symmetric three-point bending test, the support reaction is equal to half of the total applied load, whereas in the case of an eccentric three-point bending test, the reaction is equal to the fraction of the applied load. Finally, the efficiency factor was measured by equating the theoretical capacity and the measured capacity of the critical bottle-shaped strut as follows [Equation (2)]:

$$\beta_{s,measured} = C_{expt}/b.w_{s,min}.f'_c \tag{2}$$

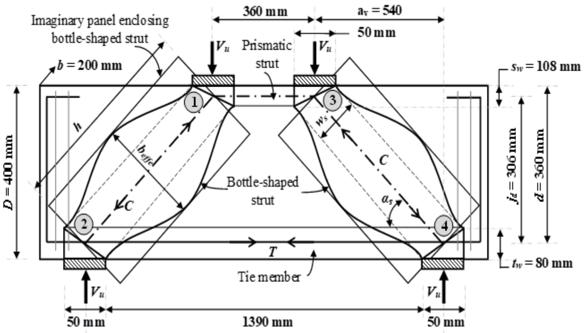
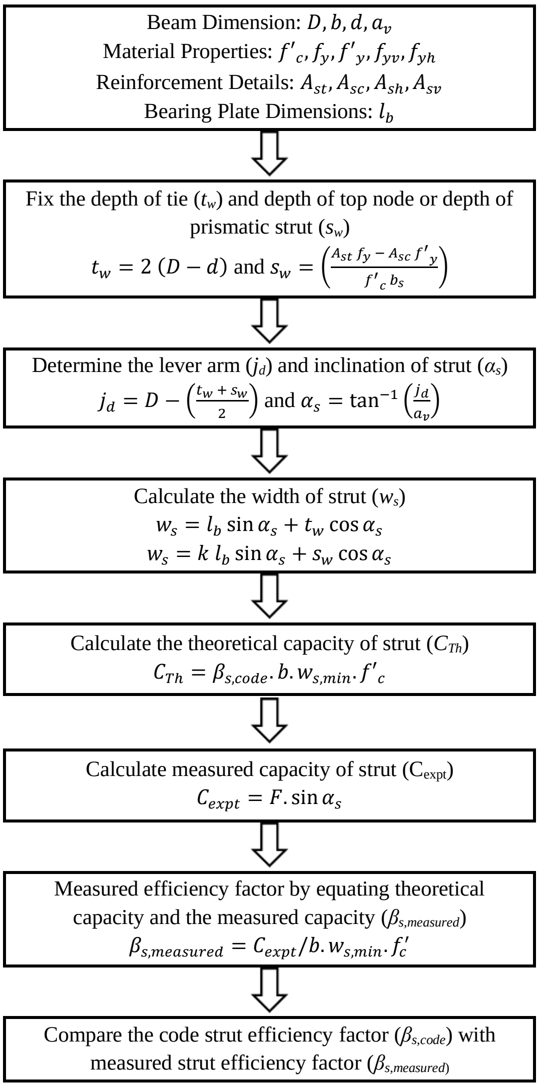


Figure 1. STM superimposed over the specimen tested by Choi et al. [22]



Beam Dimension: D, b, d, a_v
Material Properties: $f'_c, f_y, f'_y, f_{yv}, f_{yh}$
Reinforcement Details: $A_{st}, A_{sc}, A_{sh}, A_{sv}$
Bearing Plate Dimensions: l_b

Fix the depth of tie (t_w) and depth of top node or depth of prismatic strut (s_w)
$t_w = 2\,(D - d)$ and $s_w = \left(\frac{A_{st}\, f_y - A_{sc}\, f'_y}{f'_c\, b_s}\right)$

Determine the lever arm (j_d) and inclination of strut (α_s)
$j_d = D - \left(\frac{t_w + s_w}{2}\right)$ and $\alpha_s = \tan^{-1}\left(\frac{j_d}{a_v}\right)$

Calculate the width of strut (w_s)
$w_s = l_b \sin\alpha_s + t_w \cos\alpha_s$
$w_s = k\, l_b \sin\alpha_s + s_w \cos\alpha_s$

Calculate the theoretical capacity of strut (C_{Th})
$C_{Th} = \beta_{s,code}.\, b.\, w_{s,min}.\, f'_c$

Calculate measured capacity of strut (C_{expt})
$C_{expt} = F.\sin\alpha_s$

Measured efficiency factor by equating theoretical capacity and the measured capacity ($\beta_{s,measured}$)
$\beta_{s,measured} = C_{expt}/b.\, w_{s,min}.\, f'_c$

Compare the code strut efficiency factor ($\beta_{s,code}$) with measured strut efficiency factor ($\beta_{s,measured}$)

Figure 2. Flowchart to demonstrate STM analysis procedure

3. DEVELOPMENT OF PROPOSED EFFICIENCY FACTOR MODEL

In deep beams, load is primarily transferred through strut action, in which direct compression generates indirect transverse tension. It leads to a reduction in the capacity of bottle-shaped struts. Considering this fact, the following mathematical relationship for β_s is derived by regression analysis of the outcomes of STM analysis of selected specimens.

$$\beta_s = \left(\frac{0.56\sqrt{f_c'}}{12+0.002R}\right) \tag{3}$$

The numerator of Equation (3) represents the root of concrete compressive stress, and the denominator contains the relation that is the function of replacement level in percentage, *R*. Unlike flexural member design procedures, concrete tensile strength cannot be neglected in the design philosophy of shear critical members such as deep beams, especially those containing recycled aggregate concrete. Because, in the case of deep beams, usually the failure occurs due to splitting instead of crushing of concrete. Therefore, the proposed form of the equation can be more effective if the effect of split tensile strength is accommodated in the model. The value of the numerator in Equation (3) matches with the equation for the tensile strength of concrete recommended by ACI 318-14, f_t = $0.56\sqrt{f_c'}$. Thus, the direct value of split tensile strength (f_t) can be used in place of $0.56\sqrt{f_c'}$. It should be noted that, this relationship of concrete tensile and compressive strength is for NAC; however, it can be used for RAC as the effect of RCA replacement can be mitigated by the reduced compressive strength of RAC. The revised Equation (3) will take the following form [Equation (4)]:

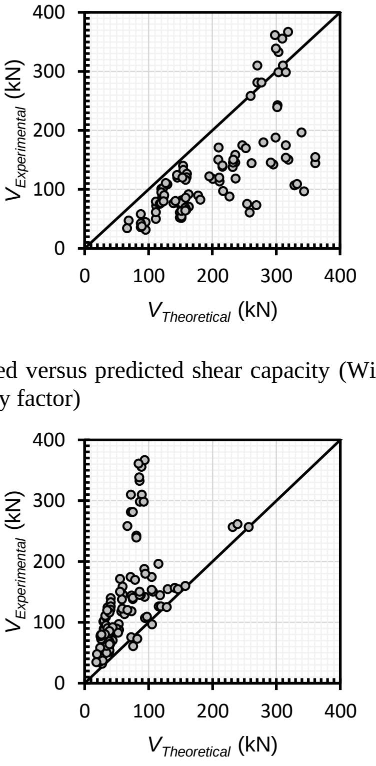
$$\beta_s = \left(\frac{f_t}{12+0.002R}\right) \tag{4}$$

The comparison of $\beta_{s,code}$ recommended by various codes along with the proposed model is compiled in Table 1.

As deliberated in the introduction, the strength of a strut is dependent on the effective compressive stress of the concrete mass that occupies the strut. It may be noted that the effective compressive stress in a bottle-shaped strut is affected by transverse stresses within the strut. Therefore, once f'_c is established, f_{cu} can be obtained by multiplying the value of β_s with the minimal of the cross-sectional areas at the two ends of the strut. Figures 3(a) and 3(b) reveal the comparison of the measured shear capacities of the selected beam specimens with and without the application of the β_s. The plot shows measured shear capacity on the ordinate and thermotical shear capacity on the abscissa. A line of 45^0 inclination is drawn to indicate the conservatism. The values of measured shear capacity lying above this line imply conservative results, whereas the values below this line indicate unsafe results. It has been observed that without the application of any efficiency factor, only 8% of the results were conservative (Figure 3(a)), on the contrary after the application of the proposed efficiency factor, around 95% of the results became conservative (Figure 3(b)). Thus, the exercise highlights the importance of the application of the strut efficiency factor in STM design procedures. The efficiency factor in any form takes care of known and unknown limitations of the STM procedures.

4. EXPERIMENTAL PROGRAM

The purpose of the beam tests was to calibrate the effectiveness of the proposed strut efficiency factor model against recycled aggregate content in concrete. The deep beam specimen was so configured that the applied load was transferred to the nearest support through a strut action. The dimensions of the beam specimens were kept constant for each replacement level of natural aggregates, as typically depicted in Figure 4 and in Table 2. The beam is proposed to be tested in 3-point bending by applying a concentrated load on the top



(a) Measured versus predicted shear capacity (Without use of efficiency factor)

(b) Measured versus predicted shear capacity (With use of efficiency factor)

Figure 3. Role of strut efficiency factor in STM Design Procedure

face of the beam at distances of 625 mm from both the supports (Figure 4). The distances were selected in such a way that the load transmitted through the bottle-shaped strut became exactly equal to the load carrying capacity of the deep beam. Therefore, the strut inclination with the adjacent tie becomes 30°. The internal force system in the deep beam could be represented using the truss models shown in Figure 4.

It can be seen that the inclined strut between nodes 3 and 4 transfers a major fraction of the applied load, *P*, to the support. Since sufficient space is available in the web of the beam for the dispersion of the compressive stress trajectories in this strut, it can be designated as a bottle-shaped strut. It is this strut which has been targeted for validation of the proposed efficiency factor model in particular and for a study of the behavior of RCA concrete bottle-shaped struts in general. It may be noted in Figure 4 that the strut inclination of 30° is close to the lower-bound strut inclination angle specified by the ACI 318-14 [5]. The design details of the deep beam specimens with concrete mix proportion are summarized in Table 2.

The control concrete mixture containing natural coarse aggregates was designed by the absolute volume method as per the provisions of IS 10262 [33] and the RCA concrete was prepared by the direct replacement of natural aggregates by weights as per a predefined percentage. Control concretes are identified in this investigation by the generic name of NCA-concrete whereas the concretes containing various fractions of recycled aggregates are identified by the generic name RCA-concrete. The nomenclature for the test specimens is defined in Table 3. Except for the substitution of the NCA fractions with the RCA, the other ingredients in these two concrete types were nominally the same. It should be noted that all the concrete ingredients conformed to the relevant Indian standards.

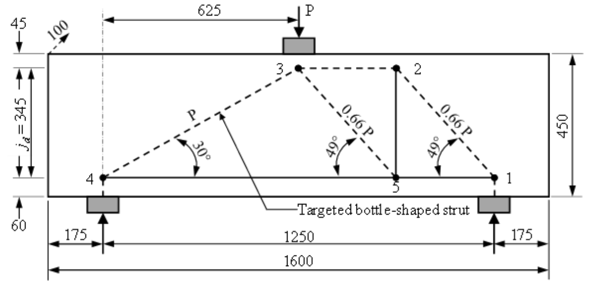


Figure 4. Truss models for the deep beam specimens (Concentrated load applied at 625 mm from nearest support)

TABLE 2. Details of deep beam specimens and concrete mixture

Specimen ID	*R*	w_s (mm)	j_d (mm)	A_{st} (mm^2)	Steel Provided
DB-R-0	0%	103	345	628	2-12Ø +2-16Ø
DB-R-50	50%	103	345	628	2-12Ø +2-16Ø
DB-R-100	100%	103	345	628	2-12Ø +2-16Ø

Concrete mix proportion (quantities in kg/m^3):

W/C Ratio	Cement	Fine Aggregate	Coarse Aggregate
0.45	370	720	1140

*All beams are 1600 mm long, 450 mm deep, and 100 mm thick

Thermo-Mechanically Treated (TMT) deformed steel bars of nominal diameters of 8 mm, 12 mm, and 16 mm were used to create the reinforcement cage as depicted in Figure 5. For obtaining the mechanical properties, all the steel bars were tested in a 1000 kN capacity tensile testing machine as per the procedure recommended in IS:1608. The longitudinal tension reinforcement in the deep beams was determined on the basis of the calculated tie forces near the beam soffits. Since the focus of this investigation was to study the effect of concrete strength and the replacement level of NA, the bottle-shaped strut region was kept free from transverse reinforcement. For the remaining region, nominal transverse reinforcement in the form of an orthogonal grid was provided. Depending upon the detailing of the reinforcement, steel cages were assembled in the laboratory, and a selection of these cages is shown in Figure 5. The reinforcement cages were placed inside steel formwork at the appropriate cover depth using concrete cover blocks, and casting was done in the laboratory.

TABLE 3. Summary of test results

Specimen ID	f'_c (MPa)	P_{cr} (*kN*)	P_u (*kN*)	w_{cr} (mm)	$\beta_{s,Measured}$	$\beta_{s,Predicted}$
DB-R-0	40.00	141	308	0.12	0.77	0.29
DB-R-50	37.50	128	294	0.12	0.76	0.28
DB-R-100	37.00	120	263	0.16	0.68	0.27

Key to specimen ID: The first two places in the nomenclature are the short form of deep beam, the third place-holder implies replacement of NA, and the last two digits indicate percentage replacement. For example, the specimen ID DB-R-50 stands for a deep beam with 50% replacement.

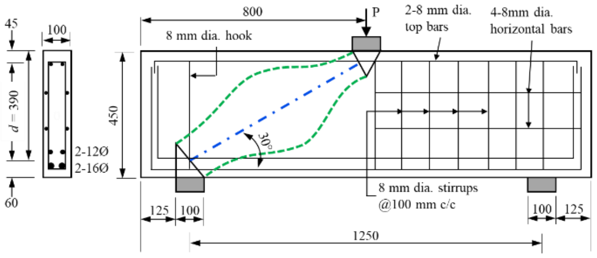


Figure 5. Detailing of reinforcement in the beams with the transversely unreinforced bottle-shaped struts

After 28 days, the beams were ready for testing. All beam specimens were subjected to 3-point bending over a simply supported span of 1250 mm. The load was applied by a 1000 kN capacity hydraulic jack, and the applied load was recorded with the help of a 1000 kN load cell. At the load point, a mild steel bearing plate of size 100 mm × 100 mm × 40 mm was used to transfer the applied load to the beam, whereas two plates of the same dimensions were used to simulate supports. A typical test setup for a deep beam test is presented in Figure 6. The loading rate was so configured that failure would occur in about 20 to 25 loading steps. The failure invariably occurred due to longitudinal splitting in the targeted bottle-shaped strut.

5. RESULTS AND DISCUSSION

In order to examine the behavior of RCA-concrete bottle-shaped strut, a series of deep beam tests were conducted. All the deep beam specimens were tested under the symmetric three-point bending test. The response of the tested specimens in terms of load at first crack (P_{cr}) and ultimate load (P_u) was recorded. Load-deformation characteristics and crack patterns were also assessed. The crack width (w_{cr}) at service load was measured. All the relevant test results are summarized in Table 3.

5. 1. Load-deformation Characteristics

Figure 7 illustrates the load-deflection relationships of the selected transversely unreinforced deep beam specimens for varying degrees of RCA replacement. Three replacement levels of NA 0%, 50%, and 100%, respectively, have been considered. The overall stiffness measured in terms of the slope of the load-deflection relationship decreased with an increase in the RCA replacement level.

To evaluate the serviceability behaviour of the bottle-shaped struts, the cracking behaviour, particularly the maximum crack widths, was monitored at every load increment. None of the service load crack width values was greater than the limiting values of 0.3 mm and 0.41 mm recommended in IS-456 [34] and ACI 318-14 [5], respectively.

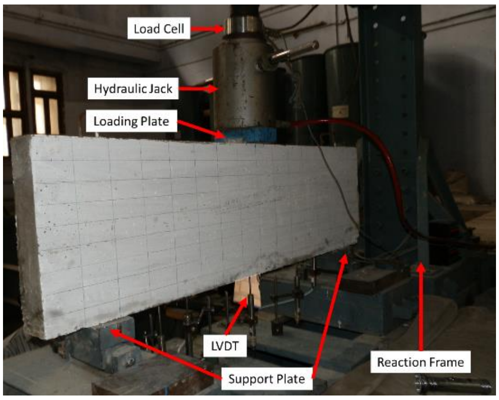


Figure 6. Typical test setup for a deep beam test

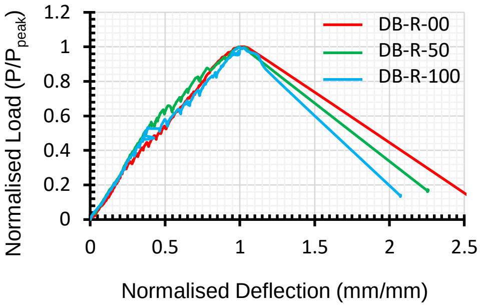


Figure 7. Load deformation relationship

Attention is drawn to the service load crack widths in the specimens, in all of which the targeted bottle-shaped strut is transversely unreinforced but the measured crack widths are all less than the limiting value of 0.3 mm. Of these three specimens, the beam DB-R-00 is made of natural aggregate concrete, whereas the specimens DB-R-50 and DB-R-100 are made of recycled aggregate concrete. One of the objectives of providing transverse reinforcement is to control cracking in the bottle-shaped struts.

The above results suggest that even in the absence of transverse reinforcement, the aim of crack control is still met. This observation may be read in the context that ACI 318-14 allows the use of transversely unreinforced bottle-shaped struts. It is emphasized here that besides controlling cracking behaviour, transverse reinforcement sustains structural capacity after splitting and imparts ductility. Hence, in line with the recommendations of Brown and Bayrak [35] transversely unreinforced RCA concrete bottle-shaped struts should not be used in practice. In Figure 8, cracking patterns of recycled concrete beams and a control beam are shown. The

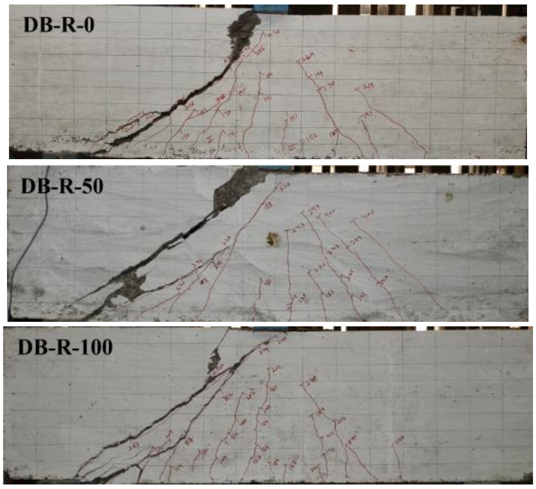


Figure 8. Comparison of crack patterns

natural concrete beam has only one prominent crack leading to failure, while the recycled concrete beam has at least two prominent cracks, indicating that recycled aggregate concrete has a relatively higher crack density.

5. 2. Appraisal of the Proposed Model The objective of the present work is to develop a simple, robust, and sensitive strut efficiency model for the concretes containing substitute materials like recycled aggregates. Another objective is to check the fit of the existing models recommended by various codes for recycled aggregate concrete. In order to meet these objectives, measured strut efficiency factors of the selected deep beam specimens are compared with the predictions of various selected models, along with the predictions of the proposed model. It is convenient to incorporate the results of deep beam tests carried out in our laboratory with the results of beam tests reported in the literature to avoid separate and repeated description. Factually, there are two sets of measured β values: a) The β values measured by processing the results of beam tests collected from the literature, and b) The experimentally investigated β values. The plots of the measured and predicted strut efficiency factors are presented in Figures 9 through 13. To differentiate outsourced and experimentally investigated values of strut efficiency factors, the experimentally measured values are indicated by squares, whereas the processed values β (of the outsourced specimens from the literature) are represented by circles.

Figure 9 depicts the comparison of measured-to-predicted efficiency factors by ACI 318-14 [5]. As can be seen in Figure 9, the ACI 318-14 gives either 0.6 or 0.75 based on the concentration of effective transverse reinforcement. Therefore, two straight vertical clusters of predicted values appear in the plot, which is practically not desired. Because, ideally, if the predictions of any model are reasonably accurate, then the scatter of the values is expected to lie above but along a 45° inclined line. Moreover, it is observed that ACI 318-14 recommendations generate a good number of unconservative results. This is mainly due to the fact that ACI 318-14 recommended efficiency factors are arbitrary values which depend on strut type. The mean and coefficient of variance (CoV) for the measured to predicted capacity were 0.95 and 0.42, respectively. The degree of conservatism, the ratio of measured to predicted strut efficiency factor, is observed to be significantly lower, i.e., 35.77 %. It should be noted that to maintain uniformity in the comparison, the reduction factors assigned for the quality of workmanship are not assigned. Application of this '$\varnothing$' factor improves the degree of conservatism; however, sometimes it leads to overly conservative estimations.

Figure 10 depicts the measured-to-predicted shear strength by AASHTO [20]. It has been observed that, the AASHTO recommendations not only have a good degree of conservatism but also a scatter that is comparatively more realistic. The only disadvantage is that AASHTO predicts overly conservative values and, also like other code provisions, is not sensitive to RCA content in the concrete. This might be a result of the overestimated value of 'ε_1'. As the AASHTO suggested model is based on the MCFT. The mean and CoV are 2.25 and 0.42, respectively.

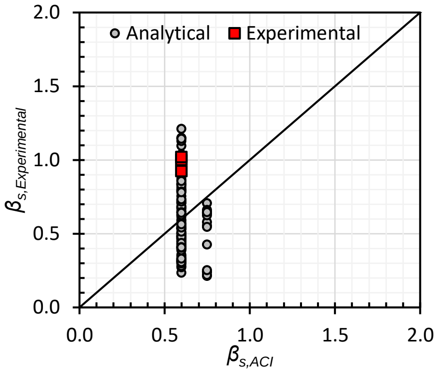


Figure 9. Comparison of measured strut efficiency factors with predictions of ACI 318-14

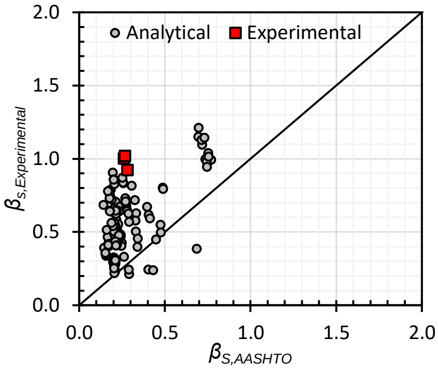


Figure 10. Comparison of measured strut efficiency factors with predictions of AASHTO

The Eurocode 2 [6] predictions are also comparatively less conservative, and the predicted values are found concentrated in one vertical cluster (Figure 11). This might be due to the fact that, the Eurocode 2 model is a single parameter model and is the function of concrete compressive strength alone. The average value of the conservatism is greater than unity, and the calculated degree of conservatism is 52.85% and CoV is 0.40.

A comparison of measured-to-predicted strut efficiency factors by AS-3600 [8] is presented in Figure 12. Like AASHTO [20], AS-3600 [8] predictions have a reasonably acceptable degree of conservatism, is calculated at 93.50%. The predictions are overly conservative and insensitive to the type of concrete. The Efficiency Factor Model adopted by AS-3600 is a

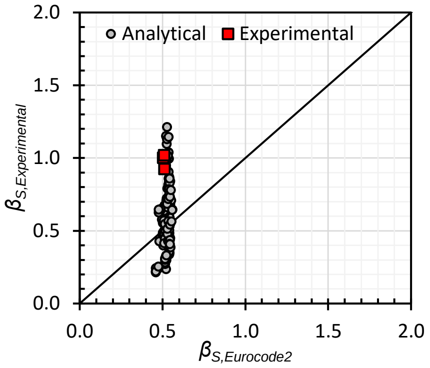


Figure 11. Comparison of measured strut efficiency factors with predictions from Eurocode 2

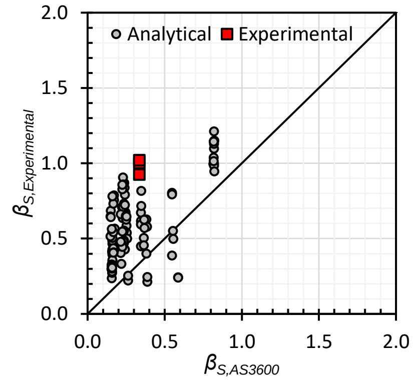


Figure 12. Comparison of measured strut efficiency factors with predictions of AS-3600

modified version of Collins and Mitchell [36] relationship. The mean value measured-to-predicted ratio was 2.33 and CoV was 0.42, respectively.

The proposed strut efficiency factor model takes into account the effect of the replacement level of RCAs, *R* and *f'*$_c$. Therefore, it becomes sensitive to the recycled aggregate contents besides the strength of the concrete. This results in comparable predictions as those of the international code provisions. With an average value of 2.35 and a CoV of 0.44, proposed model predictions are more consistent in comparison with the predictions of selected models (Figure 13). The degree of conservatism of the proposed model is about 95%, which is relatively higher than that of ACI 318-14 [5] (35.77%) and Eurocode 2 [6] (52.85%), slightly lower than that of AASHTO [20] (95.93%), and slightly greater than AS-3600 [8] (93.50%). Besides comparable predictions, the trend of the prediction is in agreement with the trend of measured efficiency factors.

An important point is to be noted that, in this analysis, neither any additional safety factors for the load or materials were applied nor factors like *Ø* factor for the quality of workmanship. This is done in order to achieve uniformity in the analysis of predictions. However, the application of suitable safety factors and reduction factors would result in a higher level of conservatism and more economical designs. Hence, the above predictions of various codes which seem unconservative may result in conservative predictions on application of partial safety factors and load factors.

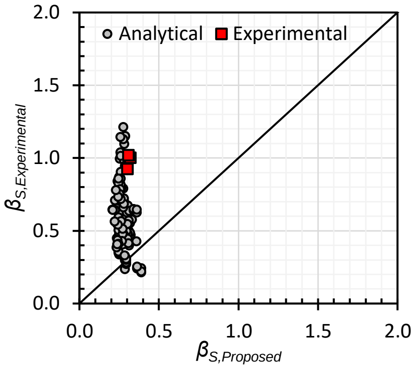


Figure 13. Comparison of measured strut efficiency factors with predictions of the proposed model

6. CONCLUSIONS

- The strut efficiency factor models recommended by various codes produce either unconservative results or overly conservative results. These models are not sensitive to the concrete containing substitute materials like recycled aggregates.
- The STM analysis reveals that the efficiency factor decreased with increased content of recycled aggregates in the concrete. Similar findings have been reported by previous studies.
- A strut efficiency factor model has been developed through regression analysis of reported test data. The proposed model accommodates the effect of concrete compressive strength and the replacement level of natural aggregates. However, accommodating the tensile strength of concrete in the proposed model makes it more rational. In the absence of measured concrete tensile strength, the relationship between compressive strength and tensile strength recommended by ACI 318-14 may be utilized.
- The proposed strut efficiency factor model has been evaluated by comparing its predictions with the measured values in the deep beam tests. Unlike the overly conservative predictions of other models, the proposed model yields moderately conservative predictions.
- The proposed model is sensitive to RCA content in the concrete. Therefore, the trends of forecasts from the proposed model are noticeably similar to the trends of measured values. On the contrary, other

models do not consider the influence of recycled aggregates on β_s. Hence, they do not exhibit similar trends, although the predictions may be conservative.

- To calibrate the efficacy of the proposed model, a series of deep beam tests were carried out. All the beam specimens were designed by STM and made up of concrete containing partly or fully replaced natural aggregates with recycled coarse aggregates. Predictions of the $\beta_{s,proposed}$ for bottle-shaped struts in deep beams made either of two concrete types were also conservative and relatively more accurate.
- The degree of conservatism of the proposed model is about 95%, which is comparable to the predictions of internationally accepted codes.

7. REFERENCES

1. Kassem, W., "Strength prediction of corbels using strut-and-tie model analysis", ***International Journal of Concrete Structures and Materials***, Vol. 9, No. 2, (2015), 255-266. https://doi.org/10.1007/s40069-015-0102-y
2. Kassem, W., "Shear strength of squat walls: A strut-and-tie model and closed-form design formula", ***Engineering Structures***, Vol. 84, (2015), 430-438. https://doi.org/10.1016/j.engstruct.2014.11.027
3. Mata-Falcón, J., Pallarés, L. and Miguel, P.F., "Proposal and experimental validation of simplified strut-and-tie models on dapped-end beams", ***Engineering Structures***, Vol. 183, (2019), 594-609. https://doi.org/10.1016/j.engstruct.2019.01.010
4. Pan, Z., Guner, S. and Vecchio, F.J., "Modeling of interior beam-column joints for nonlinear analysis of reinforced concrete frames", ***Engineering Structures***, Vol. 142, (2017), 182-191. https://doi.org/10.1016/j.engstruct.2017.03.066
5. Committee, A., "Building code requirements for structural concrete (aci 318-08) and commentary, American Concrete Institute. (2008).
6. Institution, B.S., "Eurocode 2: Design of concrete structures: Part 1-1: General rules and rules for buildings, British Standards Institution, (2004).
7. Engineers, J.S.o.C., "Jsce guidelines for concrete no. 15: Standard specifications for concrete structures—2007 "design"", (2010).
8. Chowdhury, S. and Loo, Y., "Complexities and effectiveness of australian standard for concrete structures—as 3600-2018", in EASEC16: Proceedings of The 16th East Asian-Pacific Conference on Structural Engineering and Construction, 2019, Springer. (2021), 1747-1756.
9. Doğan-Sağlamtimur, N., Bilgil, A. and Öztürk, B., Reusability of ashes for the building sector to strengthen the sustainability of waste management, in Handbook of research on supply chain management for sustainable development. 2018, IGI Global.265-281.
10. Ponnada, M.R. and Kameswari, P., "Construction and demolition waste management–a review", ***Safety***, Vol. 84, (2015), 19-46. https://doi.org/10.14257/ijast.2015.84.03
11. Nurhanim, A., "State of art reviews on physico-chemical properties of waste concrete aggregate from construction and demolition waste", ***Iranian (Iranica) Journal of Energy & Environment***, Vol. 13, No. 4, (2022), 340-348. https://doi.org/10.5829/ijee.2022.13.04.03
12. Al Martini, S., Khartabil, A. and Neithalath, N., "Rheological properties of recycled aggregate concrete incorporating supplementary cementitious materials", ***ACI Materials Journal***, Vol. 118, No. 6, (2021), 241-253. https://doi.org/10.14359/51733126
13. Singh, R.B. and Singh, B., "Rheological behaviour of different grades of self-compacting concrete containing recycled aggregates", ***Construction and Building Materials***, Vol. 161, (2018), 354-364. https://doi.org/10.1016/j.conbuildmat.2017.11.118
14. Bilgil, A., Ozturk, B. and Bilgil, H., "A numerical approach to determine viscosity-dependent segregation in fresh concrete", ***Applied Mathematics and Computation***, Vol. 162, No. 1, (2005), 225-241. https://doi.org/10.1016/j.amc.2003.12.086
15. Çakır, Ö., "Experimental analysis of properties of recycled coarse aggregate (rca) concrete with mineral additives", ***Construction and Building Materials***, Vol. 68, (2014), 17-25. https://doi.org/10.1016/j.conbuildmat.2014.06.032
16. Pawar, A.J. and Suryawanshi, S., "Comprehensive analysis of stress-strain relationships for recycled aggregate concrete", ***International Journal of Engineering, Transactions B: Applications,*** , Vol. 35, No. 11, (2022), 2102-2110. https://doi.org/10.5829/ije.2022.35.11b.05
17. Masne, N. and Suryawanshi, S., "Analytical and experimental investigation of recycled aggregate concrete beams subjected to pure torsion", ***International Journal of Engineering, Transactions A: Basics***, Vol. 35, No. 10, (2022), 1959-1966. https://doi.org/10.5829/ije.2022.35.10a.14
18. Sahoo, D.K., Singh, B. and Bhargava, P., "An appraisal of design provisions for bottle-shaped struts", ***Magazine of Concrete Research***, Vol. 64, No. 7, (2012), 647-656. https://doi.org/10.1680/macr.11.00141
19. Sahoo, D.K., Singh, B. and Bhargava, P., "Minimum reinforcement for preventing splitting failure in bottle-shaped struts", ***ACI Structural Journal***, Vol. 108, No. 2, (2011), 206. https://doi.org/10.14359/51664256
20. Specifications, A.-L.B.D., "American association of state highway and transportation officials", ***Washington, DC***, (2012).
21. Chaudhari, A.D. and Suryawanshi, S.R., An assessment of efficiency factors of recycled aggregate concrete bottle-shaped struts, in Sustainable building materials and construction: Select proceedings of icsbmc 2021. 2022, Springer.271-277.
22. Choi, H., Yi, C., Cho, H. and Kang, K., "Experimental study on the shear strength of recycled aggregate concrete beams", ***Magazine of Concrete Research***, Vol. 62, No. 2, (2010), 103-114. https://doi.org/10.1680/macr.2008.62.2.103
23. Han, B., Yun, H. and Chung, S., "Shear capacity of reinforced concrete beams made with recycled-aggregate", ***Special Publication***, Vol. 200, (2001), 503-516. https://doi.org/10.14359/10598
24. Singh, B., Sahoo, D.K. and Jacob, N.M., "Efficiency factors of recycled aggregate concrete bottle-shaped struts", ***Magazine of Concrete Research***, Vol. 65, No. 14, (2013), 878-887. http://dx.doi.org/10.1680/macr.12.00235
25. Fathifazl, G., Razaqpur, A., Isgor, O.B., Abbas, A., Fournier, B. and Foo, S., "Shear strength of reinforced recycled concrete beams without stirrups", ***Magazine of Concrete Research***, Vol. 61, No. 7, (2009), 477-490. https://doi.org/10.1680/macr.2008.61.7.477
26. Kim, S.-W., Jeong, C.-Y., Lee, J.-S. and Kim, K.-H., "Size effect in shear failure of reinforced concrete beams with recycled aggregate", ***Journal of Asian Architecture and Building***

Engineering, Vol. 12, No. 2, (2013), 323-330. https://doi.org/10.3130/jaabe.12.323

27. Etman, E.E., Afefy, H.M., Baraghith, A.T. and Khedr, S.A., "Improving the shear performance of reinforced concrete beams made of recycled coarse aggregate", ***Construction and Building Materials***, Vol. 185, (2018), 310-324. https://doi.org/10.1016/j.conbuildmat.2018.07.065

28. Aly, S.A., Ibrahim, M.A. and Khttab, M.M., "Shear behavior of reinforced concrete beams casted with recycled coarse aggregate", ***European Journal of Advances in Engineering and Technology***, Vol. 2, No. 9, (2015), 59-71.

29. Al-Zahraa, F., El-Mihilmy, M.T. and Bahaa, T., "Experimental investigation of shear strength of concrete beams with recycled concrete aggregates", ***International Journal of Materials and Structural Integrity***, Vol. 5, No. 4, (2011), 291-310. https://doi.org/10.1504/IJMSI.2011.044418

30. Lian, O.C., Wee, L.S., Masrom, M.A.a. and Hua, G.C., "Experimental study on shear behaviour of high strength reinforced recycled concrete beam", ***Pertanika Journal of Science and Technology***, Vol. 21, (2013), 601-610.

31. Arabiyat, S., Katkhuda, H. and Shatarat, N., "Influence of using two types of recycled aggregates on shear behavior of concrete beams", ***Construction and Building Materials***, Vol. 279, (2021), 122475. https://doi.org/10.1016/j.conbuildmat.2021.122475

32. Li, C., Liang, N., Zhao, M., Yao, K., Li, J. and Li, X., "Shear performance of reinforced concrete beams affected by satisfactory composite-recycled aggregates", ***Materials***, Vol. 13, No. 7, (2020), 1711. https://doi.org/10.3390/ma13071711

33. Standard, I., "Is 10262: Guidelines for concrete mix design proportioning", Indian Standard, New Delhi, (2009).

34. Visvesvarya, H., *Is 456: Plain and reinforced concrete-code of practice*. 2000, Bureau of Indian Standards.

35. Brown, M.D. and Bayrak, O., "Minimum transverse reinforcement for bottle-shaped struts", ***ACI Structural Journal***, Vol. 103, No. 6, (2006), 813. https://doi.org/10.14359/18233

36. Collins, M.P. and Mitchell, D., "Rational approach to shear design--the 1984 canadian code provisions", in Journal Proceedings. Vol. 83, (1986), 925-933.

Persian Abstract

چکیده

در روش طراحی پایه و گره (STM)، مکانیسم داخلی جریان نیروها با خرپا فرضی نشان داده می‌شود که در آن رفتار تیر توسط بار و نقاط تکیه‌گاه کنترل می‌شود. استحکام چنین پایه با ظرفیت برشی تیر عمیق از طریق عاملی به نام ضریب راندمان پایه در ارتباط است. مدل های مختلف ضریب کارایی توسط کدهای بین المللی پذیرفته شده مختلف توصیه شده است. با این حال، هیچ یک از کدها تأثیر سنگدانه های بازیافتی در بتن را در نظر نمی گیرند. اگرچه برخی از کدها نتایج محافظه کارانه ای به همراه دارند، اما این پیش بینی ها به اندازه کافی به محتوای بازیافت شده حساس نیستند. بنابراین، یک مدل ضریب کارایی حساس به بتن سنگدانه بازیافتی و کارکرد آسان بسیار مطلوب است. در این کار، نتایج منتشر شده از آزمایش‌های آزمایشگاهی بر روی نمونه‌های تیر عمیق ساخته شده از بتن متشکل از سنگدانه‌های بازیافتی، با استفاده از یک مدل پایه و اتصال مناسب برای آنالیز در نظر گرفته شد. تمامی این تیرهای عمیق در ابتدا با روش مقطعی یا تجربی طراحی شده اند. بر اساس نتایج تجزیه و تحلیل رگرسیون از STM، یک مدل عامل کارایی پیشنهاد شده است که اثر سنگدانه های بازیافتی در بتن را در نظر می گیرد. متعاقباً، نمونه‌های تیر عمیق مقیاس‌شده حاوی بتن سنگدانه‌های بازیافتی ریخته‌گری شدند و در آزمایشگاه به منظور کالیبره کردن مدل فاکتور بهره‌وری استرات پیشنهادی ریخته‌گری شدند. بازده مدل ضریب کارایی پیشنهادی با پیش‌بینی مفاد کد منتخب پذیرفته شده بین‌المللی مقایسه شد. مشخص شده است که پیش‌بینی‌های مدل عامل کارایی پیشنهادی نتایج منسجم و قابل مقایسه ای به دست آمده است.

International Journal of Engineering

Journal Homepage: www.ije.ir

A Noise-aware Deep Learning Model for Automatic Modulation Recognition in Radar Signals

M. Aslinezhad[a], A. Sezavar*[b], A. Malekijavan[a]

[a] *Department of Electrical Engineering, Shahid Sattari Aeronautical University of Science and Technology, Tehran, Iran*
[b] *Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran*

PAPER INFO

Paper history:
Received 24 March 2023
Received in revised form 08 May 2023
Accepted 13 May 2023

Keywords:
Modulation Classification
Deep Learning
Noise-aware Systems

ABSTRACT

Automatic waveform recognition has become an important task in radar systems and spread spectrum communications. Identifying the modulation of received signals helps to recognize different invader transmitters. In this paper, a noise aware model is proposed to recognize the modulation type based on time-frequency characteristics. To this end, Choi-Williams representation is used to obtain spatial 2D pattern of received signal. After that, a deep model is constructed to make signal clear from noise and extract robust and discriminative features from time-frequency pattern, based on auto-encoder and Convolutional Neural Networks (CNN). In order to reduce the effect of noise and adversarial disorders, a new database of different modulation patterns with different AWGN noises and fading Rayleigh channel is created which helps model to avoid the effects of noise on modulation recognition. Our database contains radar modulations such as Barker, LFM, Costas and Frank code which are known as frequently used modulations on wireless communication. Infact, the main novelty of this work is designing this database and proposing noise-aware model. Experimental results demonstrate that the proposed model achieves superior performance for automatic classification recognition with 99.24% of accuracy in noisy medium with minimum SNR of -5dB while the accuracy is 97.90% in SNR of -5dB and f=15 Hz of Doppler frequency. Our model outperforms 5.54% in negative and 0.4% in positive SNRs (even though with less SNR).

doi*: 10.5829/ije.2023.36.08b.06*

1. INTRODUCTION

Nowadays, digital communication plays a critical role in human life. By growing the number of transmitters in industrial mediums, i.e., Internet of Things (IoT), and with the limitation in telecommunication channels, using Cognitive Radio (CR) communications has been grown up. One of the necessary tasks for receivers is to identify the parameters of receiving unknown signals, such as kind of modulation. Therefore, Automatic Modulation Classification (AMC) can play a significant role in cognitive radio and Electronic Intelligence (ElInt). AMC is important for communication monitoring, spectrum awareness and adaptive communication [1] . AMC is necessary for both civilian and military services. One of the important applications of AMC can be sensed in Electronic Warfare (EW) in which receiver should be able to detect the modulation of unknown and adversarial signals. Beside cognitive radio, AMC is critical for radar receivers since waveform of modern signals can be changed in every pulses [2].

Researches in AMC have denoted two main categories, using Likelihood for blind classification and using feature extraction for detecting the kind of modulation. Since likelihood-based methods are time consuming with high computational complexity, feature based methods are more popular. In this way, blind modulation detection is done by extracting features from received signals and based on them, determine which modulation is used. Although there are different methods for feature extraction such as extracting hand-crafted features and extracting features based on machine learning, there still are some important challenges for AMC such as noisy mediums, adversarial attacks, multipath fading, and time varying and frequency

*Corresponding Author Email: a.sezavar@birjand.ac.ir (A. Sezavar)

Please cite this article as: M. Aslinezhad, A. Sezavar, A. Malekijavan, A Noise-aware Deep Learning Model for Automatic Modulation Recognition in Radar Signals, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1459-1467

selective channels which lead us to implement more robust and reliable systems.

In this paper, a noise-aware model is defined based on Choi-Williams transform and hybrid deep learning networks. To this end, received signal is converted to 2-D image which illustrates frequency features versus time and hybrid deep models learn to remove noises and extract robust features and classify the type of modulation. Also, a database of some important modulation with different amounts of noise is created to help models overcoming on noise effect. Block diagram of AMC by converting signal to images is illustrated in Figure 1. The novelty of the proposed method is designing arbitrary database in order for train and evaluate AMC systems. Also, a new combined noise-aware medoel is designed by combining auto-encoders and CNN which is able to overcome noise challenges. The rest of this paper is constructed as follows. Literature review of recent works on AMC is on section 2. In section 3, a brief introduction of Choi-Williams method and Convolutional Neural Networks (CNNs) are denoted following by detailed of the proposed method. Experimental results and implementation setups are shown in section 4 in which, and section 5 concludes the paper.

2. RELATED WORKS

As mentioned before, because of complexity of likelihood-based methods, feature-based models are more popular. Classical approached of feature extraction have used hand-crafted methods. Aslam et al. [2] used a combination of KNN and genetic algorithms for modulation detection of four different types of digital modulations. They have used comulants hand-crafted features in order to classify by KNN. Abdelmutalab et al. [3] used high order comulants features of received signal in order to determine the modulation by defining hierarchical polynomial classifier. Their system has achieved accurate results on two types of modulations, M-PSK and M-QAM. Saharia et al. [4] used different strong features from time, frequency and statistics domain of received signals to determine the kind of modulation. After extracting features, a Random Forest (RF) classifier was trained to identifying the modulation.

Most of recent researches on AMC have used machine learning methods especially deep learning. Several researches have used deep CNNs for extracting features from radio signals and classified them [5-9]. Since we want to use 2D inputs as images for CNNs, some resent works which converts received signals into 2D inputs are presented. Yar et al [10] used Short Time Fourier Transform (STFT) to convert raw signals to images. Before using CNN to classify input images, Hough transform was used to illustrate pulses as a single line in each image. Choi-Williams transform has been used in [11] to obtain 2D time-frequency images of modulated signals. After that, Zhang et al. [11] used CNN to classify time-frequency images and determine the kind of modulation. Although reviewed works and some other researches have achieved good results, they are limited on few number of modulations and in normal noisy channels [12]. Therefore, we need model to work in different arbitrary noise and attacks more reliable.

3. METHODOLOGY

Since we are using Choi-Williams 2D transform to obtain time-frequency images, it is needed to briefly review it and find find more about it. Also, CNNs as a powerful tool of deep learning models should be introduced. So, before going on to the proposed method, 2D transform and CNN are briefly introduced.

3. 1. Time-frequency Distribution

In order to obtain 2D images from raw signal, plotting time-frequency distribution can be useful. Although there exist some transforms which produce time-frequency analysis, Choi-Williams transform is preferred because of its advantages in removing cross-term interference. Giving raw signal as $u(t)$, Choi-Williams distribution can be obtained as follows [11, 13]:

$$CW(f,t) = \iiint_{-\infty}^{+\infty} u(s + {}^{\tau}/_{2}).u^*(s - {}^{\tau}/_{2}) k(\lambda,\tau) e^{j2\pi\lambda(s-t)} e^{-j2\pi f} d\lambda ds d\tau \tag{1}$$

$$k(\lambda,\tau) = \exp\left({}^{(\pi\lambda\tau)^2}/_{2\sigma} \right) \tag{2}$$

In which, $CW(f,t)$ denotes the time-frequency distribution and $k(\lambda,\tau)$ is a low-pass filter which helps to refuse cross-term interference and σ controls the bandwidth of the filter. By plotting $CW(f,t)$, image can be obtained and can be processed.

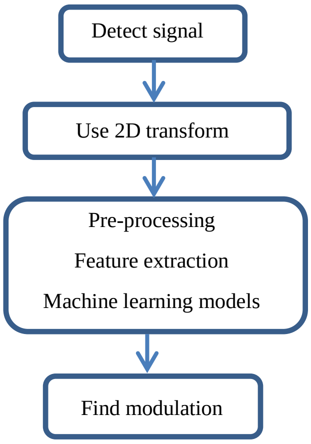


Figure 1. Block diagram of AMC using 2D transforms

3. 2. Deep Learning Deep learning is a new machine learning approach in which high-level features are extracted from input data using hierarchical layers [14]. Deep learning has demonstrated excellent data processing performance by achieving excellent accuracy in image [15-18], video[19], natural language processing [20], time series [21] and audio processing [22]. Convolutional Neural Networks (CNNs) are among the deep learning algorithms that are suitable for image processing [23, 24]. Guo et al. [14] have specifically designed CNNs for two-dimensional (2D) data such as image and video, and they also have superior image processing accuracy.

Deep learning differs from previous processing methods in that data is fed directly to the system in order to extract features, whereas in traditional processing, hand-crafted features were fed to algorithms for processing or classifying, such as artificial neural networks and other classifiers. In CNN, data is fed to the network which consists of some convolutional, pooling and fully-connected layers. During the training process, weights of convolutional kernels learn to extract meaningful features and fully-connected layers learn to classify these features to related category. Thus, the input image goes throw these hierarchical layers to extract feature and determine in which class the input belongs.

3. 3. Auto-encoder An auto-encoder is a multilayer neural network that employs encoder and decoder layers to reconstruct input [25]. In the encoder, an input image (or signal) is sent to a network where features are extracted and a tiny vector is created by downsampling. The decoder then uses supervised learning to attempt to rebuild the input by feeding it the encoded feature vector. Auto-encoders have been utilized for a variety of applications, including feature extraction and denoising in image processing [26-28].

3. 4. Implemented Method As any supervised learning model, the implemented method consists of train and test phase. To create the database, four different kinds of modulation, Barker, LFM, Costas and Frank code, are randomly created. Then, arbitrary AWGN noises with different SNRs are added to them in order to create input signal, $x(t)$ by:

$$x(t) = r(t)m_i(t) + n(t,s) \tag{3}$$

where $\mathrm{m_i(t)}$ denotes the modulated signal and $i \in$ (Barker, LFM, Costas and Frank code), $r(t)$ indicates Rayleigh fading channel described in Equation (4) and $n(t,s)$ refers to AWGN noise and s is parameters to control SNR.

$$r(t) = Ke^{j(2\pi f+\theta)} \tag{4}$$

where K is the gain of Rayleigh fading channel, f denotes Doppler frequency shift and θ is phase of path.

In order to make system noise aware and be able to overcome noises and fading effects, Artificial Distributed Signals (ADS) are created. These signals are used then to train auto-encoder to create clear transforms of signals. By adding random noises in different amounts of Doppler frequency shifts, the ADS is created in the form of Equation (3). By using Equation (1), 2D transform of each input signal is created and stored as a RGB color image. If we show the decoder and encoder performances by D(.) and E(.) respectively, the loss function for training auto-encoder is defined as follows:

$$L = \sum_i \sum_j \sqrt{(D(E(cw'(i,j)) - cw(i,j))^2} \tag{5}$$

In which, $cw'(i,j)$ is the 2D transformed of ADS and $cw(i,j)$ is the 2D transform of $\mathrm{m_i(t)}$. the training concept of auto-encoder is illustrated in Figure 2.

In training step, the train batch images are fed to CNN and during the training, until the loss function is minimum, kernel weights are updated in order to extract best features. Output of each convolution layer is calculated as follows:

$$C = Max\left(0\ , \sum_{i=1}^{K} \sum_{j=1}^{K} p(i,j) \times h(i,j)\right) \tag{6}$$

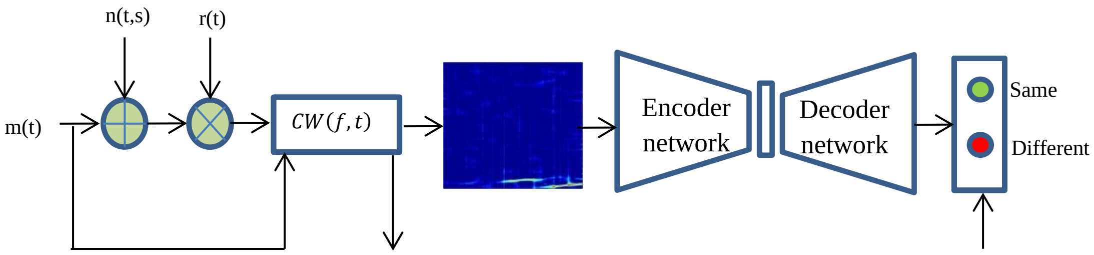


Figure 2. Training concept of auto-encoder for reconstructing main signal from ADS

In which, p is the value of pixel and h denotes the weight of filter and in order to model nonlinearity, maximum of convolution and 0 is calculated (ReLU function) and kernel size of each filter is $K \times K$. In the pooling layer, among $N \times N$ pixels, the maximum value is selected and rests of them are ignored. After some convolution and pooling, the model is followed by some fully-connected layers in which, neurons calculate a linear combination of all data in feature vector and activation function also is used to model nonlinearity. Output of each neuron is as follows:

$$f = Max\left(0\ ,\sum_{j=1}^{K} w_i \times n_i\right) \tag{7}$$

where n_i denotes a feature of is previous layer and w_i is the relevant weight to it. After training, the model is learned to extract robust features and classify them in order to distinguish type of the modulation of input signal. The structure of the implemented method is illustrated in Figure 3 and the algorithm of the proposed method is demonstrated in Algorithm 1.

Algorithm 1: proposed noise-aware deep model for modulation classification

```
Train
for $i$ in {Barker, Frank, Costas and LFM} do:
Create random $m_i(t)$
Compute $CW$ using Eq.2 and Eq.3
Initialize $K, f\ , \theta$ and $s$
$r(t) = Ke^{j(2\pi f+\theta)}$
$x(t) \leftarrow r(t)m_i(t) + n(t,s)$
Compute $CW'$ using Eq.2 and Eq.3
Train auto-encoder
Initialize $w_i$ for layers and $L$
While $L < \mathcal{E}$:
        $L^{t+1} \leftarrow \sum_i \sum_j \sqrt{(D^t(E^t(cw'(i,j)) - cw(i,j))^2}$
        $D^{t+1} \leftarrow D^t$
        $E^{t+1} \leftarrow E^t$
Train CNN
Initialize $w_i$
While max_itteration is not reached:
    For all filters and neurons in all layers do:
            $C^t = Max\left(0\ ,\sum_{i=1}^{K}\sum_{j=1}^{K} p^t(i,j) \times h^t(i,j)\right)$
            $f^t = Max\left(0\ ,\sum_{j=1}^{K} w_i{}^t \times n_i{}^t\right)$
            $t \leftarrow t + 1$
    End
End
Test
Compute $CW$ of input signal using Eq.2 and Eq.3
Compute $(D(E(CW)))$
Feed to trained CNN
Find argmax(labels)
```

4. RESULTS

In this section, before going to details of implementation and results, the dataset which is created for this paper is illustrated in subsection 4.1.

4. 1. Dataset

In order to prepare data for training CNN, four different kinds of modulation are considered, Barker, Costas, Frank code and LFM. For each kind of modulation, 120 random and different signals are created for training, with different amounts of AWGN noises with different SNR from -5dB to 5dB and 36 signals for test. Thus, we create totally 624 random noisy signals and transferred them to 624 RGB images. Some samples of created images for LFM modulation are shown in Figure 4.

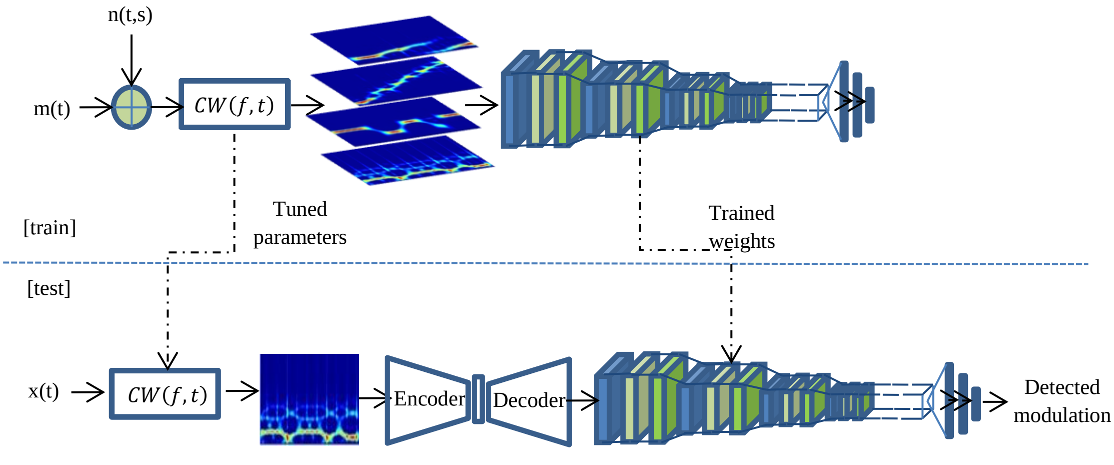


Figure 3. Diagram of the proposed method in train and test steps

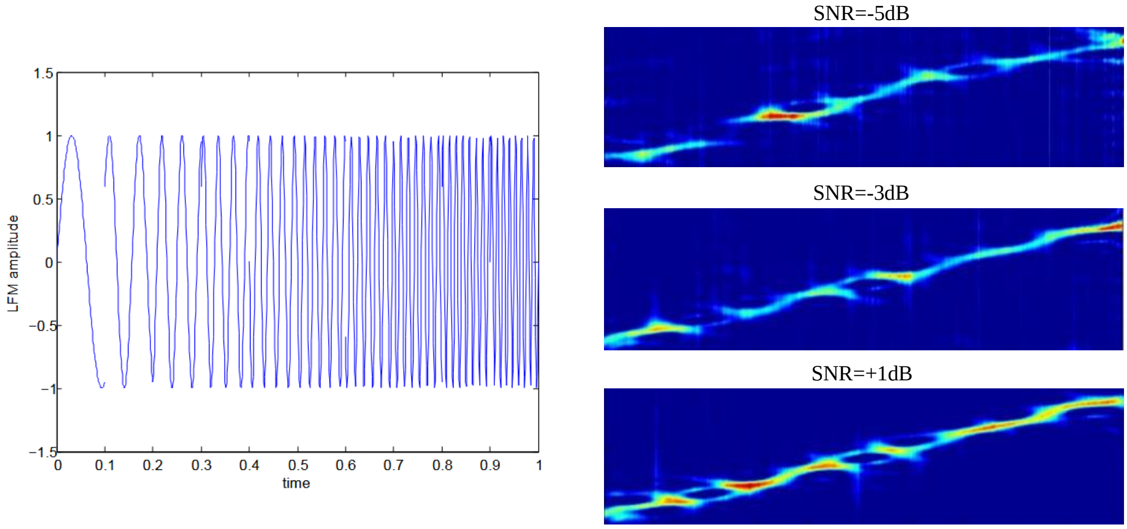


Figure 4. Samples of created images for a random LFM signal with different SNR

4. 2. Simulation Details

To go to the details of implementation, it is noticed that codes are written using python language using necessary libraries such as Tensorflow and Keras[1]. For computing 2D images, Matlab is also used. The simulation was done on 8 GB of RAM and core i-5 Intel CPU. For training auto-encoder and CNN, best hyper-parameters are obtained by tuning different amounts. The loss functions for auto-encoder and CNN was Binary and Categorical Cross-entropy, respectively; minimizing by Adam optimizer with Learning-rate of 0.005. For training, data is randomly divided to training (80%) and validation (20%) set. Because the model performs the same in training and validation data, it is understood that it can be used generally for new data with high performance. Also, by looking loss function curves of auto-encoder, it is found that the auto-encoder is trained well and is able to reconstruct input image clearly.

4. 3. Numerical Results

In order to show the performance of auto-encoder, some input noisy signals and output examples of the trained auto-encoder is illustrated in Figure 5 in which, four different samples of created ADS are shown. The first one is Barker signal with SNR=-4db in Rayleigh fading channel with Doppler frequency of 5 Hz. After using the auto-encoder, the pattern is clearly reconstructed and most parts of noises are removed as well as in other samples. It can be seen from this figure than Costas signal even with -5dB of SNR and 10 Hz of Doppler frequency is reconstructed well and clear. Results of implementing the proposed method with different SNR from 1 dB to -5 dB are illustrated under the Rayleigh fading channel with four different Doppler frequencies, 0, 5, 10 and 15 Hz. For each Doppler frequency, one diagram is considered which compares the accuracies of detection under different SNRs of white Gaussian noises.

From Figure 6, it can be found that in f=0HZ, accuracies for LFM code are upper than other and decreases from 100% to 98.6% in SNR -5dB. It can be understood that reducing 5 dB of SNR decreases just 1.4% of performance and it means that noise-aware part of model prevent noises to lack performance very much. Also, by increasing frequency to 15 Hz (which means the worthy of fading channel), accuracy of LFM falls to 98.41%. Therefore, it can be understood that the proposed model performs well even with an increase in the effect of fading Rayleigh channel. The lowest accuracy belongs to Costas modulation which is 97% in -5 dB and f=0 Hz and decreases to 96.94 in -5dB and f=15 Hz. As an ablation study, separate performance with different amount of noises and Doppler frequencies of fading channel for the proposed method, the proposed without auto-encoder and two other famous deep CNNs, are illustrated in Table 1. Based on Table 1, it can be found that although by using well-known CNNs such as VGGNet [29] and ResNet50 [24] and transfer learning, good performance can be achieved, but it will decease meaningful by decreasing SNR and increasing Doppler frequency in Rayleigh fading channel. Using the proposed method, the performance is more stable against different sittuations. In order to compare the results with the related works of AMC, accuracy of the proposed method and some related and new researches are shown in Table 2. To compare, the performance is computed in AWGN channel without fading.

As can be seen in Table 2, model proposed by Zhang et al. [11] achieved 93.7% of accuracy by combining CNN and image processing technique such as denoising and binarization on 8 different kinds of modulation.

[1] Https://keras.Io/

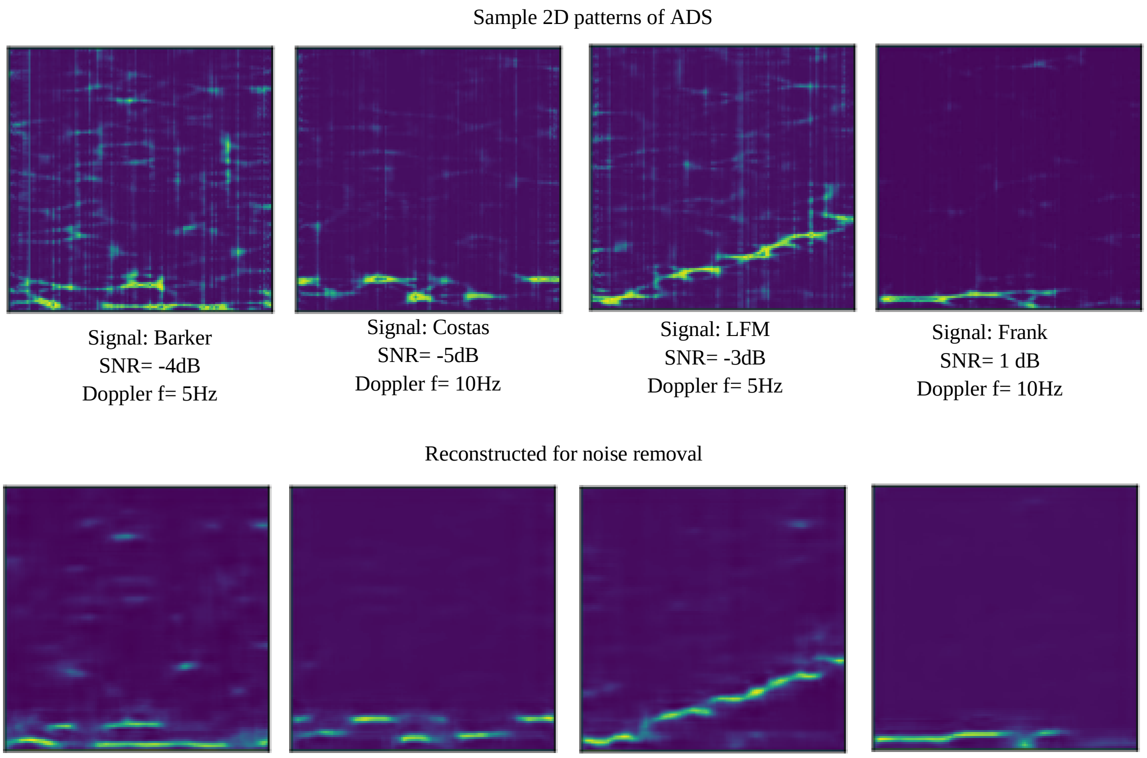
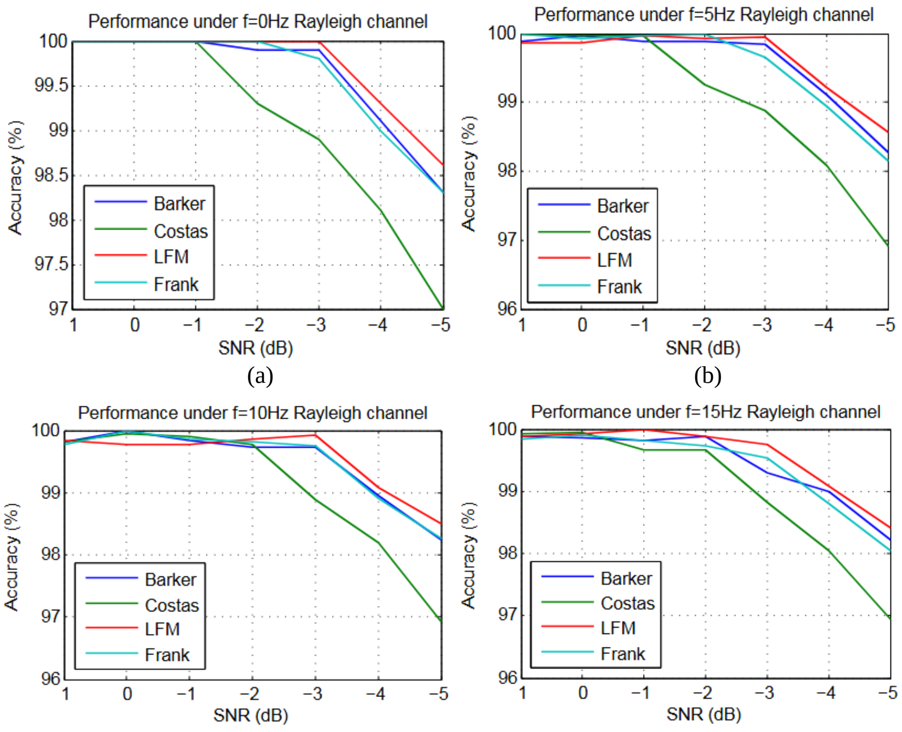


Figure 5. Samples of different ADS (first row) and their relative cleared pattern after the proposed auto-encoder (second row)



(a) (b)

(c) (d)

Figure 6. Comparison the accuracies of detection different modulations under different SNRs of white Gaussian noises with four Doppler frequencies, 0, 5, 10 and 15 Hz shown in part (a), (b), (c) and (d), respectively

TABLE 1. Numerical results and ablation for different methods under AWGN noises with SNR=0dB and -5 dB and Doppler frequencies of Rayleigh fading channel with f=5Hz, 10Hz and 15Hz

		Doppler f=5 Hz		**Doppler f=10 Hz**		**Doppler f=15 Hz**	
Model	**modulation**	**SNR=0dB**	**SNR=-5dB**	**SNR=0dB**	**SNR=-5dB**	**SNR=0dB**	**SNR=-5dB**
VGGNet	**Barker**	98.15%	90.91%	93.20%	89.80%	90.13%	81.70%
	Costas	96.73%	88.23%	91.99%	84.20%	86.07%	78.16%
	LFM	99.03%	97.03%	97.70%	88.37%	89.43%	82.84%
	Frank	98.08%	94.05%	93.86%	85.92%	88.03%	79.01%
ResNet50	**Barker**	99.01%	92.70%	94.56%	91.17%	92.47%	89.21%
	Costas	95.03%	90.42%	92.51%	88.49%	88.82%	81.40%
	LFM	98.99%	96.86%	97.51%	93.13%	92.24%	89.42%
	Frank	97.04%	92.17%	94.46%	89.18%	90.11%	84.51%
CNN (without transfer learning)	**Barker**	98.30%	91.48%	94.32%	90.06%	91.82%	87.63%
	Costas	96.23%	89.70%	91.09%	86.41%	87.19%	82.96%
	LFM	99.30%	97.72%	98.03%	94.93%	95.14%	93.02%
	Frank	98.21%	95.42%	96.06%	91.73%	93.51%	89.86%
CNN+AE (the proposed)	**Barker**	**100%**	**98.27%**	**100%**	**98.25%**	**99.86%**	**98.22%**
	Costas	**100%**	**96.91%**	**99.94%**	**96.92%**	**99.95%**	**96.94%**
	LFM	**100%**	**98.56%**	**99.97%**	**98.49%**	**99.93%**	**98.41%**
	Frank	**99.93%**	**98.16%**	**99.91%%**	**98.25%**	**99.90%**	**98.04%**

TABLE 2. Comparison between the proposed method and some state-of-the-art models for AMC in AWGN noises

Method	**SNR**	**Description**	**Accuracy**
CNNBD [11]	-2dB	CNN+binarization +denoising	93.7%
SCNN [10]	-10dB, 10dB	STFT+CNN	68.27%, 93.7%
SVMCNN [32]	2dB, +20dB	SVM+CNN	82.27%-98.52%
FCNN [33]	-10dB, +20dB	Fusuion CNN	0.09%-99.96%
3DCNN [5]	8dB, 25dB	3D CNN	98.1%, 99.6%
The proposed	-5dB, 0dB	CNN+nosie-aware training	99.24%, 100%

Combining CNN with different feature representation such as Short Term Fourier Transform (SIFT) and Support Vector Machine (SVM) leads to maximum accuracy of 93.73% and 98.52% in +10dB and +20 dB noises [10, 30]. However, between state-of-the-art models, fusion CNN [31] has achieved 99.96% of accuracy in +20dB noise and variation between accuracies are 98.1% and 99.6% in 3DCNN [5]. The proposed model achieves 100% accuracy when the SNR is 0 dB and 99.24% in the noisy environment with SNR= -5 dB which means that our method can be used generally and reliably in noisy medium.

5. CONCLUSION

Since Automatic waveform recognition is an important and challengeable task in radar systems and spread spectrum communications, this paper aims to implement a robust system for modulation classification in noisy medium. To this end, an arbitrary noisy database is created in which, different kinds of Barker, LFM, Costas and Frank code modulation in different AWGN noises are demonstrated under different Doppler frequencies of fading Rayleigh channel. Therefore, a system is implemented using Choi-Williams distribution to achieve and plot 2D features and by combining convolutional neural network and auto-encoder for training on the crated database. Experimental results showed that the proposed model outperforms new models by achieving 99.24% accuracy in minimum SNR of -5dB while the accuracy is 97.90% in SNR of -5dB and f=15 Hz of Doppler frequency. Numerical results proof that the model can be used generally on automatic modulation classification since the performance is stable in different noisy environments

6. REFERENCES

1. Peng, S., Sun, S. and Yao, Y.-D., "A survey of modulation classification using deep learning: Signal representation and data preprocessing", ***IEEE Transactions on Neural Networks and***

Learning Systems, Vol. 33, No. 12, (2021), 7020-7038. https://doi.org/10.1109/TNNLS.2021.3085433

2. Aslam, M.W., Zhu, Z. and Nandi, A.K., "Automatic modulation classification using combination of genetic programming and knn", ***IEEE Transactions on Wireless Communications***, Vol. 11, No. 8, (2012), 2742-2750. https://doi.org/10.1109/TWC.2012.060412.110460

3. Abdelmutalab, A., Assaleh, K. and El-Tarhuni, M., "Automatic modulation classification based on high order cumulants and hierarchical polynomial classifiers", ***Physical Communication***, Vol. 21, (2016), 10-18. https://doi.org/10.1109/WCNC.2016.7565127

4. Saharia, D., Boruah, M.R., Pathak, N.K. and Sarma, N., "An ensemble based modulation recognition using feature extraction", in 2021 International Conference on Intelligent Technologies (CONIT), IEEE. (2021), 1-6. https://doi.org/10.1109/CONIT51480.2021.9498547

5. Zhang, Y., Lv, X. and Min, W., "Intelligent automatic modulation classification for radar transmitting signals", in International Conference on Frontiers of Electronics, Information and Computation Technologies. (2021), 1-4. https://doi.org/10.1145/3474198.3478269

6. Hermawan, A.P., Ginanjar, R.R., Kim, D.-S. and Lee, J.-M., "Cnn-based automatic modulation classification for beyond 5g communications", ***IEEE Communications Letters***, Vol. 24, No. 5, (2020), 1038-1041. https://doi.org/10.1109/LCOMM.2020.2970922

7. Nie, J., Zhang, Y., He, Z., Chen, S., Gong, S. and Zhang, W., "Deep hierarchical network for automatic modulation classification", ***IEEE Access***, Vol. 7, (2019), 94604-94613. https://doi.org/10.1109/ACCESS.2019.2928463

8. Huynh-The, T., Hua, C.-H., Pham, Q.-V. and Kim, D.-S., "Mcnet: An efficient cnn architecture for robust automatic modulation classification", ***IEEE Communications Letters***, Vol. 24, No. 4, (2020), 811-815. https://doi.org/10.1109/LCOMM.2020.2968030

9. Chen, D., Yuan, Z., Chen, B. and Zheng, N., "Similarity learning with spatial constraints for person re-identification", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2016), 1268-1277. https://doi.org/10.1109/CVPR.2016.142

10. Yar, E., Kocamis, M.B., Orduyilmaz, A., Serin, M. and Efe, M., "A complete framework of radar pulse detection and modulation classification for cognitive ew", in 2019 27th European Signal Processing Conference (EUSIPCO), IEEE. (2019), 1-5. https://doi.org/10.23919/EUSIPCO.2019.8903045

11. Zhang, M., Diao, M. and Guo, L., "Convolutional neural networks for automatic cognitive radio waveform recognition", ***IEEE Access***, Vol. 5, (2017), 11074-11082. https://doi.org/10.1109/ACCESS.2017.2716191

12. Hessampour, K. and Latif, A., "Dimensionality reduction and improving the performance of automatic modulation classification using genetic programming (research note)", ***International Journal of Engineering, Transactions B: Applications***, Vol. 27, No. 5, (2014), 709-714. https://doi.org/10.5829/idosi.ije.2014.27.05b.05

13. Feng, Z., Liang, M. and Chu, F., "Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples", ***Mechanical Systems and Signal Processing***, Vol. 38, No. 1, (2013), 165-205. https://doi.org/10.1016/j.ymssp.2013.01.017

14. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S., "Deep learning for visual understanding: A review", ***Neurocomputing***, Vol. 187, (2016), 27-48. https://doi.org/10.1016/j.neucom.2015.09.116

15. Sezavar, A., Farsi, H. and Mohamadzadeh, S., "Content-based image retrieval by combining convolutional neural networks and sparse representation", ***Multimedia Tools and Applications***, Vol. 78, No. 15, (2019), 20895-20912. https://doi.org/10.1007/s11042-019-7321-1

16. Khan, S.U., Hussain, T., Ullah, A. and Baik, S.W., "Deep-reid: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance", ***Multimedia Tools and Applications***, (2021), 1-22. https://doi.org/10.1007/s11042-020-10145-8

17. Sezavar, A., Farsi, H. and Mohamadzadeh, S., "A modified grasshopper optimization algorithm combined with cnn for content based image retrieval", ***International Journal of Engineering, Transactions A: Basics***, Vol. 32, No. 7, (2019), 924-930. https://doi.org/10.5829/ije.2019.32.07a.04

18. Gheitasi, A., Farsi, H. and Mohamadzadeh, S., "Estimation of hand skeletal postures by using deep convolutional neural networks", ***International Journal of Engineering, Transactions A: Basics***, Vol. 33, No. 4, (2020), 552-559. https://doi.org/10.5829/ije.2020.33.04a.06

19. Nanda, A., Sa, P.K., Chauhan, D.S. and Majhi, B., "A person re-identification framework by inlier-set group modeling for video surveillance", ***Journal of Ambient Intelligence and Humanized Computing***, Vol. 10, No. 1, (2019), 13-25. https://doi.org/10.1007/s12652-017-0580-7

20. Spoorthy, G. and Sanjeevi, S.G., "Multi-criteria–recommendations using autoencoder and deep neural networks with weight optimization using firefly algorithm", ***International Journal of Engineering, Transactions A: Basics***, Vol. 36, No. 1, (2023), 130-138. https://doi.org/10.5829/ije.2023.36.01a.15

21. Pourafzal, A., Fereidunian, A. and Safarihamid, K., "Chaotic time series recognition: A deep learning model inspired by complex systems characteristics", ***International Journal of Engineering, Transactions A: Basics***, Vol. 36, No. 1, (2023), 1-9. https://doi.org/10.5829/ije.2023.36.01a.01

22. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y. and Sainath, T., "Deep learning for audio signal processing", ***IEEE Journal of Selected Topics in Signal Processing***, Vol. 13, No. 2, (2019), 206-219. https://doi.org/10.1109/JSTSP.2019.2908700

23. Krizhevsky, A., Sutskever, I. and Hinton, G.E., "Imagenet classification with deep convolutional neural networks", in Advances in neural information processing systems. (2012), 1097-1105.

24. He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2016), 770-778. https://doi.org/10.1109/CVPR.2016.90

25. Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.-A., "Extracting and composing robust features with denoising autoencoders", in Proceedings of the 25th international conference on Machine learning. (2008), 1096-1103. https://doi.org/10.1145/1390156.1390294

26. Tripathi, M., "Facial image denoising using autoencoder and unet", ***Heritage and Sustainable Development***, Vol. 3, No. 2, (2021), 89-96. https://doi.org/10.37868/hsd.v2i2.71

27. Saad, O.M. and Chen, Y., "Deep denoising autoencoder for seismic random noise attenuation", ***Geophysics***, Vol. 85, No. 4, (2020), V367-V376. https://doi.org/10.1190/geo2019-0468.1

28. Shang, Z., Sun, L., Xia, Y. and Zhang, W., "Vibration-based damage detection for bridges by deep convolutional denoising autoencoder", ***Structural Health Monitoring***, Vol. 20, No. 4, (2021), 1880-1903. https://doi.org/10.1177/1475921720942836

29. Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, (2014). https://doi.org/10.48550/arXiv.1409.1556

30. Meng, X., Shang, C., Dong, J., Fu, X. and Lang, P., "Automatic modulation classification of noise-like radar intrapulse signals

using cascade classifier", ***ETRI Journal***, Vol. 43, No. 6, (2021), 991-1003. https://doi.org/10.4218/etrij.2020-0338

31. Zheng, S., Qi, P., Chen, S. and Yang, X., "Fusion methods for cnn-based automatic modulation classification", ***IEEE Access***, Vol. 7, (2019), 66496-66504. https://doi.org/10.1109/ACCESS.2019.2918136

Persian Abstract

چکیده

امروزه تشخیص خودکار مدولاسیون به یک وظیفه مهم در سیستم های راداری و ارتباطات طیف گسترده تبدیل شده است. شناسایی مدولاسیون سیگنال های دریافتی به شناسایی فرستنده های مهاجم مختلف کمک می کند. در این مقاله، یک مدل آگاه از نویز برای تشخیص نوع مدولاسیون بر اساس ویژگی‌های زمان–فرکانس پیشنهاد شده‌است. برای این منظور، نمایش Choi-Williams برای به دست آوردن الگوی دو بعدی فضایی سیگنال دریافتی استفاده می شود. پس از آن، یک مدل عمیق برای حذف نویز از سیگنال استخراج ویژگی‌های قوی و متمایز از الگوی فرکانس زمانی، بر اساس خودرمزگزار و شبکه‌های عصبی کانولوشنی (CNN) ساخته می‌شود. به منظور کاهش تأثیر نویز و اختلالات متخاصم، یک پایگاه داده جدید از الگوهای مدولاسیون مختلف با نویزهای مختلف AWGN و کانال ریلی محوشونده ایجاد شده است که به مدل کمک می کند تا از اثرات نویز بر تشخیص مدولاسیون جلوگیری کند. پایگاه داده ما شامل مدولاسیون های راداری مانند کدBarker ، LFM، Costas و Frank است که به عنوان مدولاسیون های پرکاربرد در ارتباطات بی سیم شناخته می شوند. درواقع، نوآوری روش پیشنهادی، اولا ایجاد این پایگاه‌داده جدید و ثانیا طراحی مدل آگاه به نویز است. نتایج تجربی نشان می‌دهد که مدل پیشنهادی عملکرد برتر را برای تشخیص طبقه‌بندی خودکار با ۹۹.۲٤ درصد دقت در محیط نویزی با حداقل SNR 5-dB به دست می‌آورد در حالی که دقت در SNR 5-dB و f=15 هرتز فرکانس داپلر ۹۷.۹۰ درصد است. روش پیشنهادی باعث پیشرفت دقت به اندازه 5/54% در SNR های منفی و 0/4% ئر SNR های مثبت شده است.

International Journal of Engineering

Journal Homepage: www.ije.ir

An Incentive Mechanism for Energy Internet of Things Based on Blockchain and Stackelberg Game

H. Zhou[a], J. Gong*[b], W. Bao[b], Q. Liu[b]

[a] *Department of Electromechanical and Information Engineering, Changde Vocational Technical College, Changde, China*

[b] *School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China*

PAPER INFO

Paper history:
Received 01 November 2022
Received in revised form 18 April 2023
Accepted 19 April 2023

Keywords:
Blockchain
Energy Internet of Things
Incentive Mechanism
Stackelberg Game

ABSTRACT

In the Internet of Everything era, the Energy Internet of Things (IoT), as a typical application of IoT technology, has been extensively studied. Meanwhile, blockchain technology and energy IoT can be coordinated and complementary. The energy IoT is diversified and has a high transaction demand. it is an issue worthy of research to discuss the impact of the energy IoT environment on the performance of blockchain consensus algorithms and guarantee blockchain stability in energy IoT environment. In the research, an incentive mechanism based on Stackelberg game is proposed for the network scenario involving multiple roadside units and user nodes. The proposed strategy is analyzed through the Matlab simulation platform. The simulation results show that the proposed scheme can effectively protect the interests of blockchain users and miners. It also can improve the security and stability of the blockchain-based energy IoT system. Moreover, the numerical results not only verify the model feasibility. It also shows that when there are many blockchain miners, the model performance is fine. However, when the number of miners reaches a certain value, there will be unobvious growth. Furthermore, it is also confirmed that the wireless energy IoT environment will also create a certain impact on the game model.

***doi**: 10.5829/ije.2023.36.08b.07*

NOMENCLATURE

N	The set of miner nodes	TPS^{dag}	The number of transactions verified per second in the blockchain network
$T_s(\lambda)$	User's response time	U_r^*	The optimal total reward
$T_v(\lambda)$	The transaction verification delay	λ^*	The optimal equilibrium point
$T_w(\lambda)$	The queuing and service time	U_l^*	Benefit function
U_l	The user's benefit function	$\frac{\partial^2 U_r}{(\partial x)^2}$	The second derivative of U_r with respect to x
$f(\lambda_i)$	The satisfaction function of blockchain users	$\tau(\lambda_i)$	The verification delay of the transaction under high load.
θ	The weight factor of the response time function,	c	The computing and storage cost in each transaction
L_l	A convex function with respect to λ	$\tau(\lambda)$	The ideal response time demand
x^*	The optimal pricing strategy that can maximize U_r .		

1. INTRODUCTION

Energy IoT is a new energy internet system based on cutting-edge technologies such as 5G and artificial intelligence, combined with energy. According to the complementary mode of different energy sources, energy internet greatly promotes the linkage between electricity, fossil, and heat energy sources with the help of internet technology [1]. Meanwhile, blockchain technology and energy IoT can be coordinated and complementary in integrated development. This complement is mainly reflected in decentralization, collaborative autonomy, marketization, and smart contracts.

As a cutting-edge technology, blockchain deeply integrates a series of emerging computer technologies such as distributed data storage, P2P (peer-to-peer)

*Corresponding Author Email: junquan123gong@163.com (J. Gong)

Please cite this article as: H. Zhou, J. Gong, W. Bao, Q. Liu, An Incentive Mechanism for Energy Internet of Things Based on Blockchain and Stackelberg Game, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1468-1477

transmission, consensus mechanism, encryption algorithm, and so on. It also displays distinct application characteristics of decentralization, openness and transparency, traceability, and tamper-proofness [2]. The application value and application scenarios of blockchain technology in the field of energy IoT have been deeply discussed in a large number of studies. Zhao et al. [3] summarized and introduced the development status of blockchain energy application engineering at home and abroad. And it has provided reliable development ideas and suggestions for the engineering application of blockchain technology in China's energy field. Zhang et al. [4] comprehensively and systematically sorted out the application dimensions of blockchain technology in the energy Internet. The key role of blockchain technology in the field of energy Internet has been elaborated in detail from the perspectives of energy, information and value. Fernández-Caramés et al. [5] described the demand for blockchain technology in the IoT field and the impact of its application on the development of the modern IoT. Doshi and Varghese [6] examined how renewable energy and AI-powered IoT can be used to improve agriculture. The paper explores how to use technologies to optimize crop yield, reduce water consumption and improve the efficiency of the agricultural industry. The authors also discussed potential challenges and solutions to ensure successful implementation of smart agriculture. Wang and Liu [7] presented an energy efficient optimization method for smart-IoT data centers based on task arrival. The authors proposed a task scheduling algorithm to minimize energy consumption while ensuring system performance. The algorithm dynamically assigns tasks to different nodes based on task arrival, system load, and energy consumption. This approach is compared with existing scheduling algorithms. The results show that this method improves energy efficiency while maintaining system performance.

However, the most concerned challenge is that the current performance of the traditional blockchain cannot meet the needs of high-frequency data usage. The traditional single-chain structure results in a limited number of transactions that can be processed in a consensus cycle. This cannot meet the dynamic scalability requirements for performance of blockchain technology in the actual production. Therefore, for the scalability of blockchain, a distributed ledger based on DAG is proposed, which greatly improves the system performance under high concurrency. How to balance the response strategies of each participant to protect the interests of blockchain users, miners and the system is a problem worth studying.

Game theory is a mathematical model for the study of strategic interactions between rational decision makers [6, 7]. It can be used to analyze the strategies of nodes and the interactions between nodes. Due to the power of game theory, it is one of the new trends of future development to use game theory to solve the optimization problem in blockchain. The optimization problem, especially the CAP theory problem in current blockchain [8], is namely impossible triangle: decentralization, scalability and security. Secondly, the Stackelberg game model is generally widely used to solve the pricing problem between service providers and users [9, 10]. For wireless environments like Energy IoT, the work of end users needs to rely on the purchase of computing resources from edge computing networks. Modeling the interaction between the two using Stackelberg games is a problem worth investigating for system optimization. Nejati and Faraji [11] dealed with the issue of actuator fault detection and isolation for a helicopter unmanned aerial vehicle. The authors proposed a methodology based on the observer and residual generation technique to detect and isolate actuator faults in real-time [11]. Khosravian and Maghsoudi [12] discussed the design of an intelligent controller for station keeping, attitude control, and path tracking of a quadrotor using recursive neural networks. The authors proposed a control scheme based on the fusion of multiple recursive neural networks for precise control of the quadrotor [12]. Xiong et al. [13] discussed about cloud computing and pricing management for blockchain networks. Wei et al. [14] also investigated on application of blockchain for uncertainity in energy pricing and market pricing for the enegy sectors.

Given the basis of game theory and the problems faced in this paper, this paper proposes a Stackelberg game-based incentive mechanism based on the DAG consensus mechanism. The game model simulates the interaction between blockchain users and miners, verifying the existence of the game balance point. The simulation results show that the algorithm can effectively improve the system security and stability. Specifically, it aims to improve the system security by encouraging miners to join the blockchain network, while meeting the needs of blockchain users. The rest of the paper is organized as follows. Section 2 introduces the related problems and system models. Section 3 introduces the best solution analysis and leader analysis. In section 4, the simulation results are analyzed and the system performance is evaluated numerically. Finally, section 5 summarizes this paper. The research objective of this paper is to propose an incentive mechanism based on Stackelberg game model to simulate the transaction behavior between blockchain users and miners. The proposed scheme can effectively protect the interests of blockchain users and miners. The security and stability of the blockchain-based energy IoT system has been improved.

2. PROBLEM DESCRIPTION AND NETWORK MODEL

2. 1. System Model

Our model consists of two entities: 1. blockchain user, namely solar inverter, vehicle, etc.; 2. Blockchain consensus node, namely roadside unit with computing and storage capabilities, also known as miner, as shown in Figure 1. It is noteworthy that in DAG, miners do not need a lot of computing resources in mining, just needing to verify every collected transaction. This is referred to as mining behavior in this paper. Blockchain users deliver transactions to miner nodes through wireless channels. Wireless channels require all blockchain users in the area covered by miners' nodes to compete with each other. Miner nodes communicate with each other via wired channels, run DAG consensus algorithms, validate and store the collected transactions. This consumes computing and storage resources. Due to the selfishness of nodes themselves, this is unfair for miners. Therefore, to maintain the normal operation of the blockchain system, it is reasonable for miners to charge certain transaction fees from blockchain users. For blockchain users, the transaction verification will cause new delays, so the process from publication to confirmation of transactions in the blockchain will go through two stages: delivery and verification.

The blockchain network model considered in this study consists of multiple blockchain user clusters, each of which receives data by a miner node. Where, $N = \{1, ..., N_c\}$ represents the set of miner nodes. The number of blockchain users within the coverage area of each miner follows Poisson distribution, and the transaction arrival rate of users is $\lambda_i, i \in N$. Moreover, each user has an independent satisfaction function whose value is related to its own response time needs and the miner's pricing x of the transaction. In the blockchain-based energy IoT, the user's response time $T_s(\lambda)$ is composed of two parts. The first part is the queuing and service time in the wireless phase $T_w(\lambda) = T_q(\lambda) + T_{st}(\lambda)$, and the other part is the transaction verification delay $T_v(\lambda)$, namely:

$$T_s(\lambda) = T_w(\lambda) + T_v(\lambda) \tag{1}$$

After joining the blockchain network, the user response time is more affected by the verification delay. The delay $T_v(\lambda)$ for transactions to be validated at miner nodes is the time it takes for the cumulative weight of blockchain transactions to reach the weight threshold. Due to the directed acyclic graph property in DAG, the verification delay is proportional to the transaction generation rate λ. It means that blockchain users need to generate more transactions to meet the lower response delay requirements.

Here, in view of the queuing process in the first stage, this paper only considers the transaction verification delay under stable high load. According to the description in DAG white paper, the change process of verification delay with transaction arrival rate λ can be expressed as:

$$T_v(\lambda) = \frac{D}{0.352}\ln(4\beta L_s \lambda N_c^2 D) + \frac{W\text{-}W(T_a)}{2\beta L_s \lambda N_c^2 \omega} \tag{2}$$

Since this study only considers the block verification process during the high load phase, we need to add a restriction on the transaction generation rate, i.e.:

$$\sum_{i=1}^{N} \lambda_i \geq \frac{1}{N_c D} \tag{3}$$

where, N represents the mean value of the distributed blockchain user nodes. Meanwhile, it should be made clear that in the transaction delivery, the wireless channel capacity is limited. Therefore, the wireless channel will restrict the transaction delivery after the service intensity $\rho > 1$. Therefore, this section sets restriction $\rho \leq 1$, which can be specifically expressed as follows:

$$\lambda_i \leq \frac{m}{E[T_{st}]} \tag{4}$$

2. 2. Analysis of Stackelberg Game Model Problem

To encourage blockchain miners to share their computing resources, more miners are motivated to participate in the blockchain consensus to improve the system security. The system has the authority to require blockchain users to pay a fee for each transaction. And it allows blockchain users to have different needs for response time. Therefore, there is a non-cooperative game between blockchain users and miners. In this paper, an incentive mechanism based on Stackelberg game model is proposed to simulate the interaction between blockchain users and miner nodes. Where, the set of blockchain miners is the leader and blockchain users are the followers. Miners charge transaction fees at the expense of computing and storage resources, while blockchain

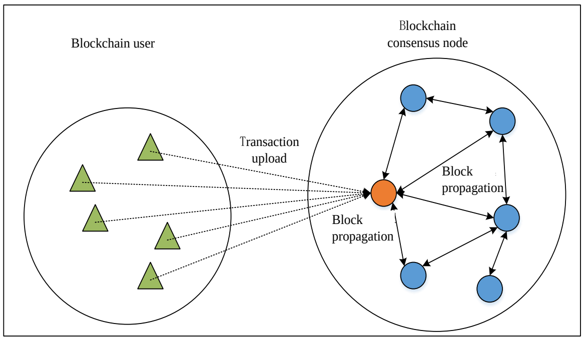


Figure 1. Game Model

users have higher demand for system response time. This paper mainly uses a game theory model to maximize the benefits of blockchain users and miner nodes. And it verifies the existence of equilibrium points in this game.

2. 2. 1. Benefit Function of Blockchain Users In terms of blockchain users, its benefit function includes satisfaction function of response time and incentive cost, namely transaction cost. The response time here represents the verification delay of transactions in the DAG network. Due to the different load of transactions arriving in the network, transactions have different verification delays. Therefore, the user's benefit function can be defined in the following equation:

$$U_l = f(\lambda_1, \lambda_2, ..., \lambda_i) - \lambda_i x \quad (5)$$

In general, logarithmic function is used to evaluate user satisfaction [11]. Therefore, in this paper, the satisfaction function of blockchain users with respect to response time is expressed as follows:

$$f(\lambda_i) = \theta \log\left[1 + g\left(\tau(\lambda_i)\right)\right] \quad (6)$$

where, θ represents the weight factor of the response time function, and $\tau(\lambda_i)$ represents the verification delay of the transaction under high load. It has been calculated previously. It can be known that $\tau(\lambda_i)$ is a function inversely proportional to the transaction rate λ_i. Let $g\left(\tau(\lambda_i)\right) = \frac{1}{\tau(\lambda_i)}$, so that can be clearly understood Equation (6).

Through the above analysis, the expression of user benefit function can be rewrite as follows:

$$U_l = \theta \log\left[1 + g\left(\tau(\lambda_i)\right)\right] - \lambda_i x \quad (7)$$

2. 2. 2. Benefit Function of Blockchain Mine For the blockchain miners, their benefit function is defined as the charged transaction fees minus the cost of computing resources consumed per transaction. Miners aim to help blockchain users verify and store valid transactions and charge transaction fees x for each transaction, thereby maximizing revenue. Mathematically, the optimization problem can be expressed as follows:

$$U_r = TPS^{dag}\left(\frac{x}{N_c} - c\right) \quad (8)$$

where, $TPS^{dag} = 2\beta L_s \lambda N_c$ represents the system throughput in the blockchain network under the wireless channel service strength $\rho \le 1$. That is, the number of transactions verified per second in the blockchain network. In Equation (8), the first term represents the average verification revenue of all blockchain miners. The second term is the computing and storage cost in each transaction c. This paper assumes that each user has the same transaction request, i.e. $\lambda_i \equiv \lambda$.

In general, the benefit functions of leaders and followers are expressed as follows:

$$\begin{aligned}
&\text{Leader}: \max_{\lambda} U_l, \\
&\text{s.t} \frac{1}{NN_c D} \le \lambda \le \frac{m}{E[T_{st}]} \\
&\text{Followers}: \max_{x} U_r, \\
&\text{s.t}\, N_c c < x < x_{\max}
\end{aligned} \quad (9)$$

3. ANALYSIS OF OPTIMAL SOLUTION

According to the Stackelberg game model proposed in section 2.2, both blockchain users and miners are rational users who want to maximize their revenues. If one party achieves the maximum revenue, it will damage the other party's revenues and eventually lead to game breakdown. Therefore, an equilibrium point must be found so that both buyers and sellers can accept it. In the model, firstly, blockchain miners fix the price of each transaction on the basis of their own cost function to gain the optimal total reward U_r^* from their own strategy space. Secondly, blockchain users choose respective response time strategy according to the pricing of miners. In this section, backward induction [15, 16] will be used to first analyze the benefit function of the following blockchain user, especially the verification delay, to obtain the optimal equilibrium point λ^* and benefit function U_l^* of the blockchain user. Then, analysis will be made on the optimal equilibrium point x^* and benefit function U_r^* of the leading blockchain miner. Finally, in the distributed environment, the optimal solution can be obtained with the help of our proposed iterative update function. Therefore, definition 1 can be obtained based on the above analysis.

Definition 1: Let the policy set of blockchain users be $R = \{\lambda_1, ..., \lambda_i\}$, and the policy set of miners be $C = \{x_1, ..., x_j\}$. When x is fixed, if λ^* meets $U_l\left(\lambda_i^*, R, x\right) \ge U_l\left(\lambda, R_{-i}, x\right)$, $\mathbb{R}_{-i}$ indicates the user policy set excluding λ_i^*. Meanwhile, when λ is fixed, if x^* meets $U_r\left(x_j^*, C, \lambda\right) \ge U_r\left(x, C_{-j}, \lambda\right)$, x_j^* $\mathbb{C}_{-j}$ represents the miner strategy set excluding. Then, the strategy $\left(\lambda^*, x^*\right)$ is the optimal equilibrium point of the non-cooperative Stackelberg game.

3. 1. Follower Analysis Through backward induction, first the benefit maximization strategy of the follower blockchain user is analyzed. For the benefit function of blockchain users, its derivative is as follows:

$$\begin{cases} \dfrac{\partial U_l}{\partial \lambda} = \dfrac{\theta}{\dfrac{W-W(T_a)}{2\beta L_s N_c^2 \omega} + \lambda} - x \\ \dfrac{\partial^2 U_l}{(\partial \lambda)^2} = -\dfrac{2\theta\beta L_s N_c^2 \omega}{\left[W-W(T_a)\right]\left(1+\dfrac{2\beta L_s \lambda N_c^2 \omega}{W-W(T_a)}\right)^2} \end{cases} \tag{10}$$

From the analysis of the above two expressions combined with the derivation in section 3, the second derivative $\frac{\partial^2 U_l}{\partial \lambda^2} < 0$ of U_l can be concluded. U_l is clearly convex function with respect to λ. Due to the constraint conditions in Equation (9), generally Lagrange multiplier method is used to solve the optimization problem. After substituting the constraint conditions into the benefit function, the following expression can be obtained.

$$L_l(\lambda, v, \vartheta) = \theta \log\left[1 + g(\tau(\lambda))\right] - \lambda x - v\left(\frac{1}{NN_cD} - \lambda\right) - \vartheta\left[\lambda - \frac{m}{E[T_{st}]}\right] \tag{11}$$

Based on this, the KKT condition can be obtained as shown in Equation (12). Where, * represents the optimal solution.

$$\begin{aligned} &\vartheta^*\left[\lambda^* - \frac{m}{E[T_{st}]}\right] = 0, \\ &v^*\left(\frac{1}{NN_cD} - \lambda^*\right) = 0, \\ &\lambda^* - \frac{m}{E[T_{st}]} \le 0, \\ &\frac{1}{NN_cD} - \lambda^* \le 0, \\ &\lambda^* > 0, v^* \ge 0, \vartheta^* \ge 0. \end{aligned} \tag{12}$$

Let $\frac{\partial L_l(\lambda, v, \vartheta)}{\partial \lambda} = 0$, then the optimal policy λ^* of blockchain users can be obtained.

$$\lambda^* = \frac{\theta}{x - v^* + \vartheta^*} - \frac{W - W(T_a)}{2\beta L_s N_c^2 \omega} \tag{13}$$

It is noteworthy that λ^* is a function of x, v^*, ϑ^*, which means that the corresponding x, v^*, ϑ^* is the information necessary to get λ^*. In addition, the instantaneous values of iteration parameters v^t, ϑ^t at time t can be calculated by solving Equations (11) and (12) simultaneously, as shown in Equation (14). t represents the index of iteration times.

$$\begin{cases} v^t = \dfrac{\dfrac{m}{E[T_{st}]} x^t - \theta \log\left\{1 + \dfrac{\left[W-W(T_a)\right]m}{2\beta L_s N_c^2 \omega E[T_{st}]}\right\}}{\dfrac{m}{E[T_{st}]} - \dfrac{1}{NN_cD}} \\ \vartheta^t = \dfrac{\dfrac{1}{NN_cD} x^t - \theta \log\left[1 + \dfrac{W-W(T_a)}{2\beta L_s NN_c^3 \omega D}\right]}{\dfrac{m}{E[T_{st}]} - \dfrac{1}{NN_cD}} \end{cases} \tag{14}$$

3. 2. Leader Analysis On the basis of the optimal strategy of the following blockchain user, the second step of backward induction method is to use the obtained optimal strategy solution of the follower and substitute it into the leader's utility function. Then the first order and second derivative analysis is used in the Stackelberg game to find the optimal strategy x^* of the leading blockchain miner.

For the blockchain consensus node, based on backward induction, the second derivative of U_r with respect to x can be expressed as follows:

$$\frac{\partial^2 U_r}{(\partial x)^2} = 2\beta L_s \left[2\frac{\partial \lambda}{\partial x} + (x - N_c c)\frac{\partial^2 \lambda}{(\partial x)^2}\right] \tag{15}$$

To prove the existence of extreme values of U_r, the concavity and convexity must be analyzed first. Therefore, to further solve the first and second derivatives of λ with respect to x, the following expression can obtained:

$$\begin{cases} \dfrac{\partial \lambda}{\partial x} = -\dfrac{\theta}{(x - \vartheta + v)^2} \\ \dfrac{\partial^2 \lambda}{(\partial x)^2} = \dfrac{2\theta}{(x - \vartheta + v)^3} \end{cases} \tag{16}$$

By analyzing the above equation, the first and second derivatives $\frac{\partial \lambda}{\partial x} < 0, \frac{\partial^2 \lambda}{(\partial x)^2} > 0$ in Equation (15) can be obtained. Finally, through the above analysis, it can be obtained that when $x_{\max} = \frac{2\theta N_c c}{(x-\vartheta+v)^3} + \frac{2\theta}{(x-\vartheta+v)^2}$, $\frac{\partial^2 U_r}{\partial x^2} < 0$ if $x \in [x_{\min}, x_{\max}]$, and the benefit function U_r of blockchain miners is a convex function with respect to x.

Therefore, when $\frac{\partial U_r}{\partial x}=2\beta L_s\left(\lambda+x\frac{\partial\lambda}{\partial x}-N_c c\right)=0$, the optimal strategy price x^* of blockchain miners can be obtained, namely:

$$x^*=\Big[\sqrt{(4\theta\lambda-4\theta N_c c)v^*-4\theta\vartheta^*\lambda+4\theta N_c\vartheta^* c+\theta^2}+(2N_c c-2\lambda)v^*+2\vartheta\lambda-2N_c\vartheta^* c-\theta\Big](2\lambda-2N_c c)^{-1} \tag{17}$$

where, x^* has a negative solution, which does not meet the conditions and will not be discussed here. Meanwhile, as can be seen from Equation (17); x^* is a closed expression related to $\lambda, v^*, \vartheta^*$. Therefore, to solve this equation, the game strategies $\lambda, v^*, \vartheta^*$ of both parties in the previous round must be obtained first.

However, in a distributed environment, since the two sides of the game are non-cooperative, neither the blockchain miner nor the user knows the optimal strategy of the other. Therefore, this paper uses the classical iterative method [17] to find the optimal solution, and this process is shown in Algorithm 1.

In Algorithm 1, if the iterative convergence condition is not met, the value calculated in this round will be used as the initial value for the next round of update, and this process will be repeated until x, λ converge.

The above analysis, on the basis of definition 1, demonstrates that the optimal solution is the unique equilibrium solution by proving 1 and 2.

Proof 1: For blockchain user, when the transaction price x is fixed, λ^* makes the user benefit function U_l globally optimal. In particular, it is proved in section 3.1 that $\frac{\partial^2 U_l}{\partial\lambda^2}<0$, $\frac{\partial^2 L_l}{\partial\lambda^2}=\frac{\partial^2 U_l}{\partial\lambda^2}<0$ under KKT, so L_l is a convex function with respect to λ, which meets the contents of Definition 1.

Proof 2: For blockchain miners, when the user gets the ideal response time demand $\tau(\lambda)$, the optimal trading strategy λ can be obtained. As proved in section 3.2, under the condition $\frac{\partial^2 U_r}{(\partial x)^2}<0$, x^* is the optimal pricing strategy that can maximize U_r.

Algorithm 1 Iterative update algorithm

```
Input: initial value $x^t, v^t, J^t$, convergence accuracy $\varepsilon$, other parameter values of energy IoT.
Output: convergent $t, x^*, \lambda^*$.
The number of initialization iterations $t$=0; the flag bit flag=flase, the initial value of $x^t$, $\Delta U_r=\left|U_r^{t+1}-U_r^t\right|>\varepsilon$, $\varepsilon$ denotes the convergence accuracy;
while (!flag)
    The blockchain user gets $x^t$ from the blockchain miner and updates it into $\lambda^t(x^t)$;
    The blockchain miner obtains the updated $\lambda^t$ from the DAG network and substitutes it into Equation (17);
    Update $v^t, \vartheta^t$ according to Equation (14);
    if ($\Delta U_r<\varepsilon$)
        flag=true;
        $x^*=x^t, \lambda^*=\lambda^t$;
    t=t+1;
endwhile;
return $t, x^*, \lambda^*$;
```

4. PERFORMANCE EVALUATION

In this paper, an incentive scheme based on Stackelberg game is proposed for the network scenario involving multiple roadside units and user nodes. The proposed strategy is analyzed through the Matlab simulation platform. The following will first explain the scenario setting of simulation verification. The specific simulation parameters are shown in Table 1.

In this section, the system performance is evaluated numerically from three aspects. First, the update process of blockchain user and miner policies with the number of iterations is examined. Second, the influence of user distribution on benefit function in the energy IoT scenario is considered. Third, as the number of blockchain miners increases, the change trend of the benefit function is analyzed.

Miners, as leaders, first have the authority to formulate pricing strategies. This is to update respective strategies for following blockchain users on the basis of miners' strategies to meet their own response time requirements. Figure 2 represents the iterative update process of transaction pricing for blockchain miners. In this figure, transaction price decreases with an increase in the number of iterations, which ultimately converges to a stable value. This is because only when the transaction price x is lower, blockchain users will choose

TABLE 1. Simulation Parameters of the Game Model

Parameter	Value (range)
DAG transaction broadcast delay *D*	1×10^{-2}s
DAG verification threshold *W*	800
DAG transaction weightω	3
Wireless transmission transaction threshold *m*	32
Algorithm convergence accuracy ε	10^{-8}
Weight factor θ	1
Mining cost in transaction *c*	10^{-2}

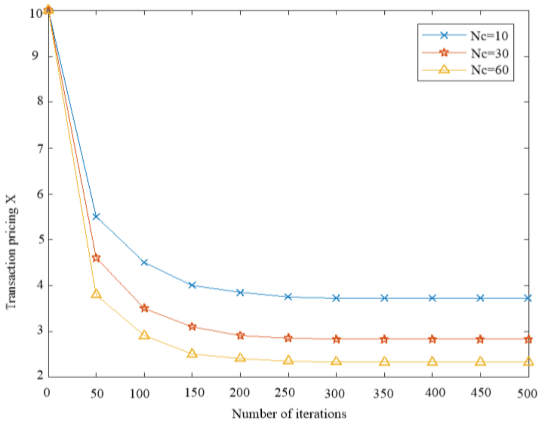


Figure 2. Update Process of Transaction Price Strategy with the Number of Iterations

to increase transaction arrival rate strategy λ. Although transaction price falls, a greater number of transactions in the network will make miner's total revenue increase. In addition, as the number of miners increases, so does the ability to collect transactions in the network. Therefore, despite the low transaction price, the miners' revenue can still be guaranteed.

Figure 3 shows the change trend of the transaction demand rate of blockchain users with the number of iterations under different number of miners. Similarly, it can be seen from the figure that with an increase in the number of iterations, the transaction demand rate of blockchain users increases and finally enters a stable state. This is because as the number of miners increases, the transaction price decreases, which exactly encourages blockchain users to demand faster transaction rates.

Where, verification delay $T_v(\lambda)$ is a function of λ, which represents the transaction verification delay of blockchain users. As can be seen from Figure 4, with an increase in the number of iterations, the value of $T_v(\lambda)$ will gradually decrease, which is consistent with the analysis result in Figure 3. Since $T_v(\lambda)$ is inversely proportional to λ, when λ increases, the user's verification delay will decrease. Consequently, the benefit function of the user is guaranteed, and eventually, $T_v(\lambda)$ will tend to a stable value.

Figures 5 and 6 show the impact of the number of miners on the benefit functions of blockchain users and miners themselves. According to the figure, as the number of miners increases, the benefit function of blockchain users and miners will also increase. This is because more miners can process more transactions per unit time. That is, the number of transactions participating in the consensus process per unit time increases. This leads to the decrease of verification delay, the improvement of blockchain users' satisfaction, and an increase in transaction demand. For blockchain miners, although the

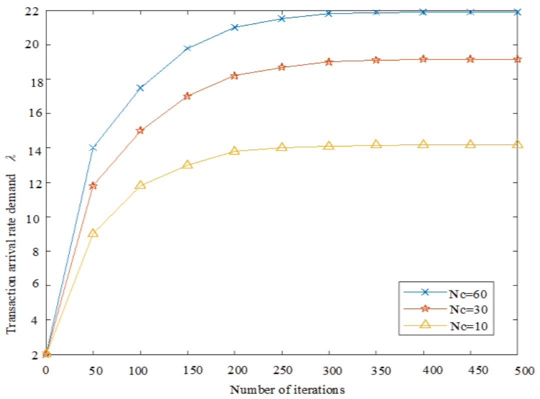


Figure 3. Update process of demand strategy for transaction arrival rate with the number of iterations

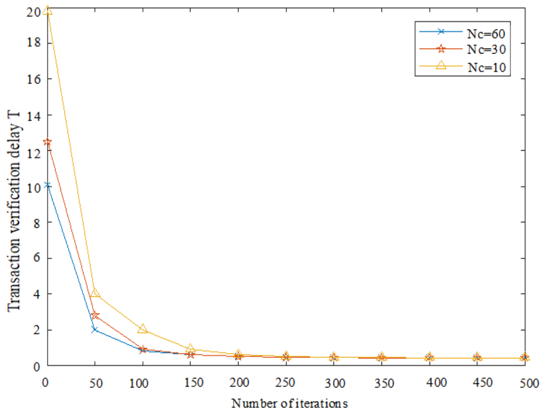


Figure 4. Update Process of Transaction Verification Delay with the Number of Iterations

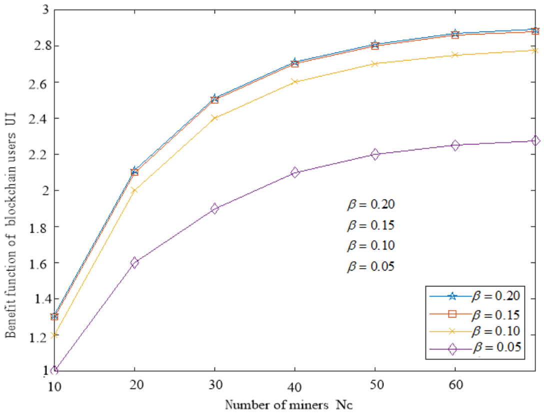


Figure 5. Benefit Function of Blockchain Users

transaction price is falling, more transactions in the system can also ensure that miners gain decent revenue.

This paper also shows distribution comparison of four groups of blockchain users in Figures 5 and 6. It can be seen that, due to the limitation of wireless environment, under greater blockchain user distribution area, that is, greater β value, the benefit function will be greater. However, despite the continuous increase in β value, the difference between the two curves β=0.15, β=0.20 in the figure is obviously less than that of β=0.05, β=0.10 . This is because the dense distribution of blockchain users will lead to the continuous decline of transaction delivery efficiency in the wireless environment. This will slow down the growth in the number of transactions in the network, reducing benefits for blockchain users and miners.

Through the above simulation, it can be concluded that the incentive mechanism proposed in this paper not only encourages miners to join the blockchain network. This increases the system stability and meets the response time requirements of blockchain users. This is the purpose of this algorithm, namely, not only guaranteeing the interests of both parties of the game, but also improving the distributed stability of the system.

6. DISCUSSION

The proposed incentive mechanism based on Stackelberg game has numerically proved to be beneficial for both blockchain users and miners. Simulation results have shown that the proposed scheme can effectively protect the interests of blockchain users and miners, and improve the security and stability of the blockchain-based energy IoT system. This conclusion is supported by the results of several studies. For example, a survey conducted by Liu et al. [8] on blockchain on the use of game theory to analyze the incentives of different participants in an energy blockchain system found that the incentive mechanism proposed in their study was able to balance the interests of energy producers, consumers, and miners. Similarly, a study by Sun et al. [18] investigated on the impact of game theory on the security of blockchain-based energy trading systems, and found that game-theoretic approaches can effectively enhance the security of energy trading systems. Moreover, a study by Dong et al. [19] on the use of game theory to optimize the performance of blockchain-based energy trading systems found that the game-theoretic approach can effectively improve the performance of blockchain-based energy trading systems. These studies all provide evidence that the proposed incentive mechanism based on Stackelberg game can protect the interests of blockchain users and miners, and improve the security and stability of the blockchain-based energy IoT system.

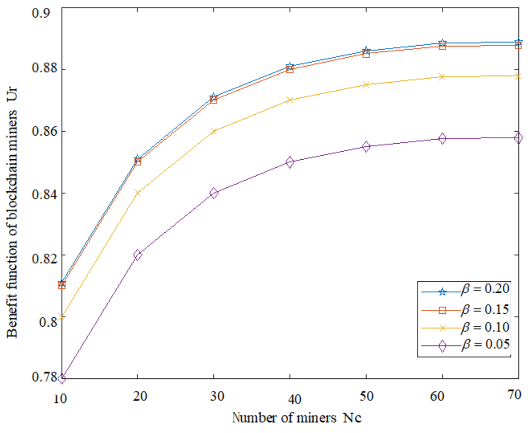


Figure 6. Benefit Function of Blockchain Miners

6. CONCLUSION

In this paper, the Stackelberg game is used to coordinate the needs of blockchain users and miners. Blockchain users can upload data to the DAG blockchain by paying a fee to blockchain miners. Miners can gain revenue by charging transaction fees. Through the game, on the one hand, the revenue of the whole blockchain miners can be guaranteed, and on the other hand, the response time demand of blockchain users can be guaranteed. The numerical results not only verify the model feasibility, but also show that when there are many blockchain miners, the model performance is fine, but when the number of miners reaches a certain value, there will be unobvious growth. Furthermore, the wireless energy IoT environment can be confirmed that it will also create a certain impact on the game model. The simulation results also show that with an increase in the number of miners, the benefit function of blockchain users and miners will also increase. This is because more miners can process more transactions per unit time. This can reduce verification delay, improve blockchain users' satisfaction, and an increase in the transaction demand. For blockchain miners, although the transaction price is falling, more transactions in the system can also ensure that miners gain decent revenue. Overall, the results of this study show that the proposed incentive scheme based on the Stackelberg game model can effectively protect the interests of blockchain users and miners, and improve the security and stability of the blockchain-based energy IoT system.

This research has several limitations. First, it only focuses on the game model between blockchain users and miners, and does not consider the impact of other factors on the system performance. Second, the simulation parameters are only applied in the energy IoT environment. There is no discussion on the application of the proposed model in other scenarios. Third, the game model in this paper only considers the response time requirements of blockchain users, and does not consider the resource utilization efficiency of blockchain miners. To further improve the system performance, there is still a lot of work to be done in the future. First, the game model should be extended to consider the resource utilization efficiency of blockchain miners. Second, the game model should consider the impact of other factors on system performance such as network latency, transaction broadcast delay, etc. Third, the application of the proposed model should be further extended to other scenarios. Finally, additional research should be done to explore other incentive mechanisms for blockchain networks.

7. FUNDINGS

The research is supported by: Basic and Advanced Research Projects of CSTC (No. cstc2019jcyj-zdxmX0008); the Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJZD-K201900605); The project of Chongqing big data application and Development Administration Bureau (No.22-30); The project of Housing and Urban-rural Development Commission of Chongqing Municipality (No.2021-0-104).

8. REFERENCES

1. Rifkin, J., "The third industrial revolution: How lateral power is transforming energy, the economy, and the world, Macmillan, (2011).
2. Nakamoto, S., "Bitcoin: A peer-to-peer electronic cash system", ***Decentralized Business Review***, (2008), 21260.
3. Zhao, Y., Peng, K., Xu, B.-y. and Liu, Y., "Status and prospect of pilot project of energy blockchain", ***Automation of Electric Power Systems***, Vol. 43, No. 7, (2019), 14-22.
4. Zhang, N., Wang, Y., Kang, C., Cheng, J. and He, D., "Blockchain technique in the energy internet: Preliminary research framework and typical applications", ***Proceedings of the CSEE***, Vol. 36, No. 15, (2016), 4011-4022.
5. Fernández-Caramés, T.M. and Fraga-Lamas, P., "A review on the use of blockchain for the internet of things", ***Ieee Access***, Vol. 6, (2018), 32979-33001. doi: 10.1109/ACCESS.2018.2842685.
6. Doshi, M. and Varghese, A., Smart agriculture using renewable energy and ai-powered iot, in Ai, edge and iot-based smart agriculture. 2022, Elsevier.205-225.
7. Wang, B. and Liu, F., "Task arrival based energy efficient optimization in smart-iot data center", ***Mathematical Biosciences and Engineering***, Vol. 18, No. 3, (2021), 2713-2732. doi: 10.3934/mbe.2021138.
8. Liu, Z., Luong, N.C., Wang, W., Niyato, D., Wang, P., Liang, Y.-C. and Kim, D.I., "A survey on blockchain: A game theoretical perspective", ***IEEE Access***, Vol. 7, (2019), 47615-47643. doi: 10.1109/ACCESS.2019.2909924.
9. Saad, W., Han, Z., Debbah, M., Hjorungnes, A. and Basar, T., "Coalitional game theory for communication networks", ***Ieee Signal Processing Magazine***, Vol. 26, No. 5, (2009), 77-97. doi: 10.1109/MSP.2009.000000.
10. Hakak, S., Khan, W.Z., Gilkar, G.A., Imran, M. and Guizani, N., "Securing smart cities through blockchain technology: Architecture, requirements, and challenges", ***IEEE Network***, Vol. 34, No. 1, (2020), 8-14. doi: 10.1109/MNET.001.1900178.
11. Nejati, Z. and Faraji, A., "Actuator fault detection and isolation for helicopter unmanned arial vehicle in the present of disturbance", ***International Journal of Engineering***, ***Transactions C: Aspects,*** Vol. 34, No. 3, (2021), 676-681. doi: 10.5829/IJE.2021.34.03C.12.
12. Khosravian, E. and Maghsoudi, H., "Design of an intelligent controller for station keeping, attitude control, and path tracking of a quadrotor using recursive neural networks", ***International Journal of Engineering, Transactions B: Applications,*** Vol. 32, No. 5, (2019), 747-758. doi: 10.5829/ije.2019.32.05b.17.
13. Xiong, Z., Feng, S., Wang, W., Niyato, D., Wang, P. and Han, Z., "Cloud/fog computing resource management and pricing for blockchain networks", ***IEEE Internet of Things Journal***, Vol. 6, No. 3, (2018), 4585-4600. doi: 10.1109/JIOT.2018.2871706.
14. Wei, W., Liu, F. and Mei, S., "Energy pricing and dispatch for smart grid retailers under demand response and market price uncertainty", ***IEEE Transactions on Smart Grid***, Vol. 6, No. 3, (2014), 1364-1374. doi: 10.1109/TSG.2014.2376522.
15. Yang, D., Xue, G., Fang, X. and Tang, J., "Incentive mechanisms for crowdsensing: Crowdsourcing with smartphones", ***IEEE/ACM Transactions on Networking***, Vol. 24, No. 3, (2015), 1732-1744. doi: 10.1109/TNET.2015.2421897.
16. Hedges, J., "Backward induction for repeated games", arXiv Preprint arXiv:1804.07074, (2018). https://doi.org/10.48550/arXiv.1804.07074
17. Cao, B., Xia, S., Han, J. and Li, Y., "A distributed game methodology for crowdsensing in uncertain wireless scenario", ***IEEE Transactions on Mobile Computing***, Vol. 19, No. 1, (2019), 15-28. doi: 10.1109/TMC.2019.2892953.
18. Sun, J., Wu, C. and Ye, J., "Blockchain-based automated container cloud security enhancement system", in 2020 IEEE international conference on smart cloud (SmartCloud), IEEE. (2020), 1-6.
19. Dong, J., Song, C., Liu, S., Yin, H., Zheng, H. and Li, Y., "Decentralized peer-to-peer energy trading strategy in energy blockchain environment: A game-theoretic approach", ***Applied Energy***, Vol. 325, (2022), 119852. https://doi.org/10.1016/j.apenergy.2022.119852

Persian Abstract

چکیده

در عصر اینترنت همه چیز، اینترنت اشیا انرژی (IoT)، به عنوان یک کاربرد معمولی فناوری اینترنت اشیا، به طور گسترده مورد مطالعه قرار گرفته است. در همین حال، فناوری بلاک چین و انرژی اینترنت اشیا می توانند هماهنگ و مکمل یکدیگر باشند. انرژی اینترنت اشیا متنوع است و تقاضای تراکنش بالایی دارد. بحث در مورد تأثیر محیط اینترنت اشیا انرژی بر عملکرد الگوریتم های اجماع بلاک چین و تضمین ثبات بلاک چین در محیط اینترنت اشیا انرژی، موضوعی است که ارزش تحقیق دارد. در این تحقیق، یک مکانیسم انگیزشی مبتنی بر بازی Stackelberg برای سناریوی شبکه شامل چندین واحد کنار جاده‌ای و گره‌های کاربر پیشنهاد شده است. استراتژی پیشنهادی از طریق پلت فرم شبیه سازی Matlab تحلیل می شود. نتایج شبیه سازی نشان می دهد که طرح پیشنهادی می تواند به طور موثر از منافع کاربران بلاک چین و ماینرها محافظت کند. همچنین می‌تواند امنیت و ثبات سیستم اینترنت اشیاء مبتنی بر بلاک چین را بهبود بخشد. علاوه بر این، نتایج عددی نه تنها امکان‌سنجی مدل را تأیید می‌کنند. همچنین نشان می دهد که وقتی ماینرهای بلاک چین زیادی وجود دارد، عملکرد مدل خوب است. با این حال، زمانی که تعداد ماینرها به مقدار مشخصی برسد، رشد نامشخصی وجود خواهد داشت. علاوه بر این، همچنین تایید شده است که محیط اینترنت اشیا انرژی بی سیم نیز تاثیر خاصی بر مدل بازی ایجاد خواهد کرد.

International Journal of Engineering

Journal Homepage: www.ije.ir

A Voice Activity Detection Algorithm Using Sparse Non-negative Matrix Factorization-based Model Learning in Spectro-Temporal Domain

S. Mavaddati*

Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran

PAPER INFO

Paper history:
Received 02 January 2023
Received in revised form 13 May 2023
Accepted 21 May 2023

Keywords:
Voice Activity Detector
Spectro-temporal Domain
Sparse Structured Principal Component Analysis
Sparse Non-negative Matrix Factorization

ABSTRACT

Voice activity detectors are presented to extract silence/speech segments of the speech signal to eliminate different background noise signals. A novel voice activity detector is proposed in this paper using spectro-temporal features extracted from the auditory model of the speech signal. After extracting the scale, rate, and frequency features from this feature space, a sparse structured principal component analysis algorithm is used to consider the basic components of these features and reduce the dimension of learning data. Then these feature vectors are employed to learn the models by the sparse non-negative matrix factorization algorithm. The model learning procedure is performed to represent each feature vector with a proper sparse rate based on the selected atoms. Voice activity detection of the input frames is performed by computing the energy of the sparse representation for each input frame over the composite model. If the calculated energy exceeds a specified threshold, it indicates that the input frame has a structure similar to the atoms of the learned models and concludes that the observed frame has voice content. The results of the proposed detector were compared with other baseline methods and classifiers in this processing field. These results in the presence of stationary, non-stationary and periodic noises were investigated and they are shown that the proposed method based on model learning with spectro-temporal features can correctly detect the silence/speech activities.

doi*: 10.5829/ije.2023.36.08b.08*

1. INTRODUCTION

One of the research fields in the speech signal processing is detection of silence/speech areas of the speech signal performed by a voice activity detector (VAD). The VAD block has an important role to eliminate the background noise from the speech signals. So far, different feature domains have been used to determine voice activities since the performance of VAD is closely related to the type of these extracted features. In these methods, an attempt is made to separate the speech frames from the silent sections of the speech signal. The energy of speech signal frames and the calculation of the zero-crossing rate (ZCR) are the most advanced features in this processing area [1]. Since various detectors have been introduced in many fields, this paper only deals with the methods presented based on the model learning technique. Ahmadi, and Joneidi [2] proposed a VAD algorithm based on the sparse representation technique using an orthogonal matching pursuit algorithm (OMP) followed by the K-singular value decomposition (K-SVD) dictionary learning method. The detection criterion of voice activity was based on the energy in the sparse representation of the input frame over the learned voice dictionary. You et al. [3] proposed a VAD algorithm based on the sparse representation technique using the Bergman iteration method and online dictionary learning. In this algorithm, the sparse power spectrum criterion was defined to calculate two types of features and decide on the label of the input frames. This criterion was achieved by averaging over the different signal segments that include the short segment average spectrum and long segment average spectrum. The labels of the different parts of the input frame are determined by calculating the energy in these frames. You et al. [4] optimized algorithm for learning speech and noise dictionaries. The

Corresponding Author Email: s.mavaddati@umz.ac.ir* (S. Mavaddati)

Please cite this article as: S. Mavaddati, A Voice Activity Detection Algorithm Using Sparse Non-negative Matrix **Factorization**-based Model Learning in Spectro-Temporal Domain, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1478-1488

goal of this optimization procedure was to reduce the coherence value between the learned dictionaries to obtain a robust VAD algorithm in the different noise conditions. The features used in this method were the modified versions of the features presented by You et al. [3] and include the long-time average energy and the long-time dynamic threshold. Also, Teng and Jia [5] designed a VAD algorithm using a non-negative sparse coding method with a noise reduction procedure. In this method, the input noise signal is first represented in the combined dictionary which contains the atoms associated with the speech and noise signals. The coefficients related to voice segments are then used as the desirable features in the conditional random field (CRF) method to model the correlation between the feature sequences and detect the speech and noise labels for each input frame. Mavaddaty et al. [6] used the spectro features of speech signal spectrograms to learn the models using the concepts of sparse representation and the K-SVD algorithm. In this work, two supervised and semi-supervised methods were presented to eliminate the background noise from the speech signal. The main part of each method was the presented voice activity detector in the wavelet packet transform domain.

The purpose of this paper is to increase the detection accuracy as much as possible based on the proposed model-based method by applying the spectro-temporal features. In this paper, scale, rate, and frequency characteristics extracted from the auditory model of the speech signal were used to learn models that show the structure of active parts of speech signals. In the following, the dimension of the mentioned features was reduced by the parse structured principal component analysis (SSPCA) algorithm and then the sparse non-negative matrix factorization (SNMF) algorithm is employed to learn the dimensionless feature sets.

In the second part of this paper, the auditory model and its extractive features are introduced. Section 3 introduces the SNMF model and the proposed VAD algorithm. In section 4, the performance of the proposed method is evaluated and the paper is concluded in section 5.

2. Spectro-Temporal Representation Using Auditory Model

As stated, the recognition process to detect speech areas of the speech signal and separate these frames from the silence frames has a great importance in many speech processing applications. In this paper, the spectro-temporal features are used to identify the speech segments of the speech signals that can be described using the auditory model. In this model, the auditory spectrum related to each speech is calculated. Then, the spectro-temporal features are extracted using this spectrogram and the auditory cortex model [7]. The features of the auditory cortex model have four dimensions: scale Ω, speech rate ω, frequency f, and time or frame number t. The auditory part of the cortical model is implemented by a time-frequency filter bank. Each filter can operate at different rates and scales to simulate the cochlear of the human ear and the first layer of the auditory brainstem. This procedure of filtering at different rates and scales is performed linearly in the spectro-temporal space by the wavelet transform function or the two-dimensional Gabor filter [7-9].

The block diagram of the auditory cortex model is shown in Figure 1. Initially, the acoustic signal enters the filter bank that consist of 128 uniformly distributed bandpass filters along the frequency-logarithmic axis that models the performance of the outer membrane of the human ear. The output of this filter bank with a time-frequency structure passes from three steps: a derivative high pass filter, a nonlinear compressor, and a low pass filter to simulate the inner portion of the human ear. In the following, the auditory spectrogram of the speech signal is obtained by the first-order derivative, half-wave rectifier, and integrator. Then, the spectro-temporal content of the auditory spectrum is achieved by a filter bank consisting of a two-dimensional Gabor filter. Then, a four-dimensional speech cortical signal including Ω scale in cycles/octave, speech velocity or rate ω in Hertz, frequency f, and the frame number of the input speech signal t is yielded.

3. THE PROPOSED VOICE ACTIVITY DETECTOR

In this section, the proposed VAD algorithm is presented using the extracted spectro-temporal features and SNMF-based model learning. The proposed method employs model learning technique to represent the structure of the input frame. Model learning in this paper is performed by the sparse non-negative matrix factorization algorithm, which is the non-negative matrix factorization (NMF) procedure that has been added to the nonlinearity constraint.

The combination of the sparse and NMF coding algorithms results in a model learning method called SNMF [10-12]. This technique results in a sparser representation than the NMF algorithm to apply the sparse constraints. In the SNMF algorithm, which is more robust than the NMF algorithm, the generalized Kolbeck-Leibler divergence method used to determine lower approximation error in the data representation. In the sparse encoding technique, each input signal frame can be represented as a linear combination of the dictionary atoms. In this procedure, it is determined which set of atoms and coefficients represent the data frame with the least approximation error. These sparse coefficients for all input signal frames constitute the H sparse coefficient matrix, which is one of the outputs of the SNMF algorithm. Many coefficients in the sparse matrix H have a zero value and indicate that each data frame can be represented only by a limited number of dictionary atoms. The sparsity or cardinality parameter determines the number of atoms in each representation procedure. The data matrix containing signal frames S can be modeled as follows by sparse coding:

$$\mathrm{S} = \mathrm{WH} \tag{1}$$

where $\mathrm{W} \in \mathrm{R}^{\mathrm{N}\times\mathrm{L}}$ is a learned model or dictionary in which the columns are atoms. The W dictionary matrix contains L columns or atoms $\{\mathrm{W}_\mathrm{l}\}_{\mathrm{l}=1}^{\mathrm{L}}$ with the unit norm $\left\|\mathrm{W}_{(:,l)}\right\|_2 = 1, \forall l = 1,\ldots,L$. Also, the K-sparse coefficient matrix H with L≫K includes the representation coefficients related to the input data matrix [13]. The sparse representation problem that consists of the approximation error and sparse constraint parts is formulated as follows [13]:

$$\mathrm{H}^* = \underset{\mathrm{H}}{\mathrm{argmin}} \left\|\mathrm{S} - \mathrm{WH}\right\|_2^2 \quad \text{s.t.} \quad \left\|\mathrm{H}\right\|_0 \le \mathrm{C} \tag{2}$$

where C represents the sparse rate or the number of non-zero coefficients in each row of the sparse matrix H. This parameter must be set correctly to avoid massive coding. If the high value is selected for this parameter, the large numbers of atoms participate in the representation of the input data frame that is improper. On the other hand, if the low value is selected for this parameter, the atoms are not enough to represent the data structure, and then the approximation error increases. The NMF algorithm performs a linear analysis on the observed data matrix $\mathrm{S} \in \mathrm{R}^{\mathrm{N}\times\mathrm{M}}$ and factorizes the input data matrix into two dictionary matrix $\mathrm{W} \in \mathrm{R}^{\mathrm{N}\times\mathrm{L}}$ and the coefficient matrix $\mathrm{H} \in \mathrm{R}^{\mathrm{L}\times\mathrm{M}}$ as $\mathrm{I} = \mathrm{WH}$ with non-negative values, which L is smaller than M and N [13]. These matrices are employed to solve the following optimization problem:

$$\begin{aligned} &\min F(W,H) = \sum_{i,j} (S_{i,j} \log(S_{i,j} [WH]_{i,j}) - S_{i,j} \\ &+ [WH]_{i,j}) \quad s.t. \quad W,H \ge 0, \sum_l \mathrm{W}_{(:,l)} = 1 \end{aligned} \tag{3}$$

The optimization of this cost function is based on the generalized Kullback-Leibler divergence method. However, solving this problem with other cost functions yields different versions of the NMF algorithm.

As stated, the SNMF algorithm will obtain a sparser representation to consider a specified constraint than the NMF algorithm [11-13]. The generalized Kullback-Leibler divergence algorithm is then used to determine the approximation error in the analysis of non-negative coefficients, which results in the following optimization problem:

$$\begin{aligned} &\min F(W,H) = \sum_{i,j} (\mathrm{S}_{i,j} \log(\mathrm{S}_{i,j} [WH]_{i,j}) - S_{i,j} \\ &+ [WH]_{i,j}) + \alpha \sum_{k,j} h_{k,j} \quad s.t. \quad W,H \ge 0, \sum_l \mathrm{W}_{(:,l)} = 1 \end{aligned} \tag{4}$$

The α parameter determines the weight coefficient of the sparsity part. The update of atoms in the W dictionary matrix is as follows:

$$\begin{aligned} &\mathrm{h}^*_{\mathrm{k,j}} = (\mathrm{h}_{\mathrm{k,j}} \sum_{\mathrm{i}} \mathrm{I}_{\mathrm{i,j}}\, \mathrm{w}_{\mathrm{i,k}} \Big/ \sum_{\mathrm{l}} \mathrm{w}_{\mathrm{i,l}} \mathrm{h}_{\mathrm{l,j}}) / (1+\alpha), \\ &\mathrm{w}^*_{\mathrm{i,k}} = (\mathrm{w}_{\mathrm{i,k}} \sum_{\mathrm{j}} \mathrm{I}_{\mathrm{i,j}}\, \mathrm{h}_{\mathrm{k,j}} \Big/ \sum_{\mathrm{l}} \mathrm{w}_{\mathrm{i,l}} \mathrm{h}_{\mathrm{l,j}}) \Big/ \sum_j h_{k,j} \;, \\ &\mathrm{w}^{**}_{\mathrm{i,k}} = (\mathrm{w}^*_{\mathrm{i,k}} \Big/ \sum_{\mathrm{i}} \mathrm{w}^*_{\mathrm{i,k}}) \end{aligned} \tag{5}$$

The NMF algorithm is obtained when the α parameter is omitted in Equation (5) [11]. Then, the dimensionality of the data matrix is reduced by the SSPCA algorithm to learn comprehensive models for the representation of the input data structure. The principal component analysis algorithm (PCA) is a commonly used statistical method to reduce data dimension and is used to convert the input data sets into a new set of the independent variables that include the maximum changes in the original data [13]. This algorithm presented to develop the SSPCA method which is used to estimate the basic components by applying a sparsity constraint [14]. The benefits of using this method include reducing computation time, extracting the components with more variance, and obtaining appropriate values for important variables of each problem. Further, by generalizing this algorithm, the SSPCA algorithm is obtained, which can extract the data with more variance using the sparsity and some structural constraints [15]. The non-convex form of the SSPCA algorithm is presented by Jenatton et al. [16] to solve the problem of structured sparse dictionary learning. The SSPCA is a robust algorithm to solve the occlusion problem using the block-coordinate descent algorithm for better data analysis.

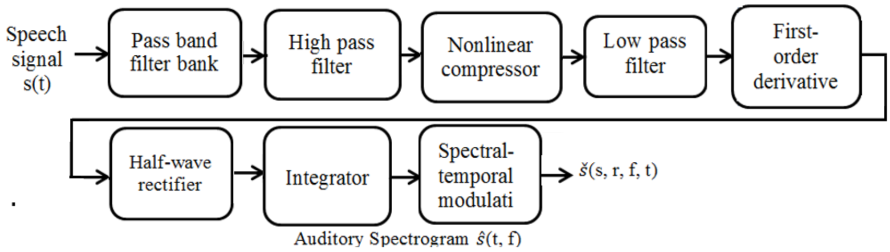


Figure 1. Block diagram of the cortical model of the speech signal

The block diagram of the proposed method to determine the labels of input speech frames using spectro-temporal properties is shown in Figure 2.

4. DETAILS OF SIMULATION

In this paper, the TIMIT dataset is used to determine the efficiency of the proposed method. This comprehensive speech dataset contains a large number of speakers and expressions that is suitable to consider the performance of a VAD algorithm [17]. The sampling rate of speech signals is set to 16kHz. The train and test scenarios contain 200 and 100 spoken expressions, respectively. In the training step, the phrases are uttered by 10 female and 10 male speakers. In the test step, the phrases uttered by 3 male and 3 female speakers were employed in the speaker-independent test. The data frame length is equal to 20 ms and the frame overlap is 50%. The parameter settings in the learning procedure are the same for all spoken data sources in the train and test scenarios.

4. 1. Simulation Results

In the proposed method, the model learning procedure using the SNMF algorithm was used to identify silence/speech speech frames. The sparsity rate for the dictionary and the coefficients matrices in the SNMF algorithm are set to 0.9 and 0.7, respectively. These parameter values are achieved based on the experimental simulations to result in a lower approximation error. Also, the number of iterations and the sparsity parameter in the SSPCA as employed in dimension reduction are 250 and 0.6, which leads to stability in the solving procedure. The performance evaluation of the algorithms is determined by the classification accuracy rate, which is calculated by the percentage of voice and silence frames that have the correct labels for the entire test data. In the first step of the proposed algorithm, 100 speech signals with silence/speech labels selected from the TIMIT dataset are used to learn the model of scale, rate, and frequency features extracted from the auditory cortex model. The auditory model of these signals is computed and then applied to the model learning after employing the SSPCA dimension reduction algorithm. Finally, the the learned models that represent the structural features of the silence/speech segments are considered in the representation of the test input signal. The sparsity parameter in this algorithm means that each input data frame can only be represented by a linear combination of a small number of learned atoms. This parameter value is determined by the cardinality rate. Input data classification in this paper is not performed by the usual classifiers such as neural networks, support vector machine or decision trees, but it is suggested to design and use a model-based classifier based on the calculated energy of the extracted features from the sparse coefficients matrix. In the proposed detection procedure, the input signal is sparsely represented by the SNMF algorithm on the combinational dictionary D= [Ds Dr Df]. This composite model D consists of the learning models related to the scale, rate, and frequency features with the same parameter values in the training step. Then, the energy of the sparse representation coefficients obtained on each dictionary is computed as:

$$\mathrm{H}_{\mathrm{S}}^{*}, \mathrm{H}_{\mathrm{R}}^{*}, \mathrm{H}_{\mathrm{F}}^{*} = \mathrm{SNMF}\left(\boldsymbol{Y},\ \left[\mathrm{D}_{\mathrm{s}}\ \mathrm{D}_{\mathrm{r}}\ \mathrm{D}_{\mathrm{f}}\right]\right) \tag{6}$$

$$\mathrm{E}_{\mathrm{s}} = {}^{1}\!/\!_{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \boldsymbol{H}^{*2}_{s,l}\ ,\ \mathrm{E}_{\mathrm{r}} = {}^{1}\!/\!_{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \boldsymbol{H}^{*2}_{r,l},\ \mathrm{E}_{\mathrm{f}} = {}^{1}\!/\!_{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \boldsymbol{H}^{*2}_{f,l} \tag{7}$$

where L represents the length of the frame and E_{s}, E_{r}, and E_{f} are the energy of the sparse representing related to scale, rate, and frequency features. $\boldsymbol{Y}$ is the observation matrix. Also, $\mathrm{H}_{\mathrm{S}}^{*}$, $\mathrm{H}_{\mathrm{R}}^{*}$ and $\mathrm{H}_{\mathrm{F}}^{*}$ are sparse coefficient matrices related to scale, rate and frequency features of the speech signal. The sum of energies is calculated and if this energy is more than half the energy of the input frame then it can result that the input frame contains the voice structure. If the difference between the calculated energy in the sparse coding procedure over the speech model and the energy of the input frame is less than a specified value of ε_1=0.04, then the average energy of the SNMF coding coefficients for one frame before and one frame after the input frame is calculated as the short-term energy. The value ε_1 has been obtained experimentally in various simulations. If the short-term energy is higher than half the energy of the input frame, the input frame has a speech label otherwise it will have a silence label.

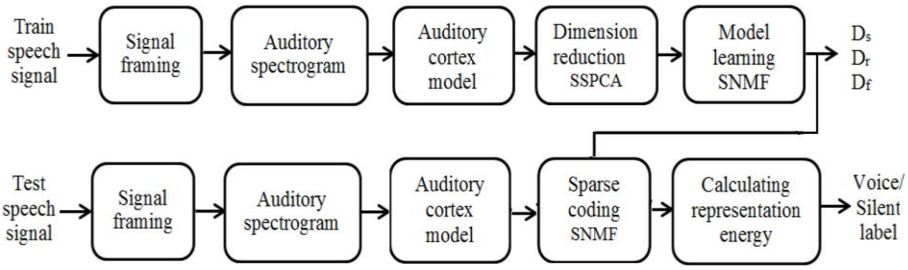
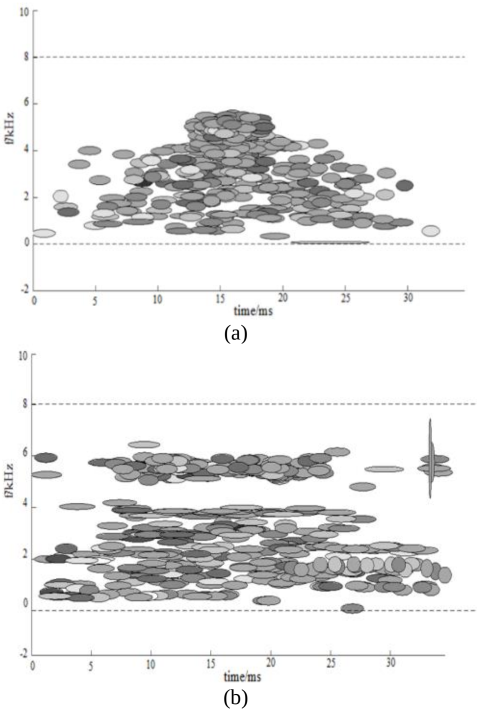


Figure 2. Block diagram of the proposed method to determine silence/speech speech frames using spectro-temporal properties

The time-frequency energy plot of the learned atoms based on the proposed SNMF-based VAD algorithm using the elliptical plots presented by Jafari and Plumbley [18] which is shown in Figure 3. This procedure determines how much of the time-frequency energy of frames is sparsely represented by the learned atoms. These plots show that the learned atoms according to the proposed method have been able to cover the entire time-frequency space of the considered speech signals. The elliptical plot of the proposed method based on the frequency features is concentrated in the center of the time axis and it does not include the entire frequency content at different times. The proposed method consists of a wide time-frequency range caused by a proper matching with the content of observation signal and the dictionary atoms.

The spectrogram plots of the atoms learned by the proposed method, the frequency features and the spectro-temporal features are shown in Figure 4. These plots show that the learned atoms according to the proposed method have the highest energy coverage in the time-frequency space and can precisely display the structure of the speech and silence frames.The parameters setting procedure was done according to the experimental simulations to have a proper decision on the input frame label. Since the input data frame with voice content has more energy in the sparse representation on the related dictionary so the energy criterion of the resulting sparse coefficients is used to determine the appropriate label. As a result, there is no need to use other classifiers, and the labeling procedure of the input frame can only be estimated using the SNMF algorithm. The results of the proposed method to detect the silence/speech frames are reported in Tables 1 and 2 for the speaker-independent and speaker-dependent detection scenarios, respectively. It is noteworthy that this paper has tried to evaluate the performance of the proposed VAD algorithm with the methods presented in the field of sparse representation technique. The results show that the proposed method has the ability to correctly identify the input area by applying the comprehensive learning models based on the structural content of the input frames. These results are slightly higher in the speaker-dependent scenario than in the speaker-independent scenario, which may be due to the overlap between the train and test data speakers.

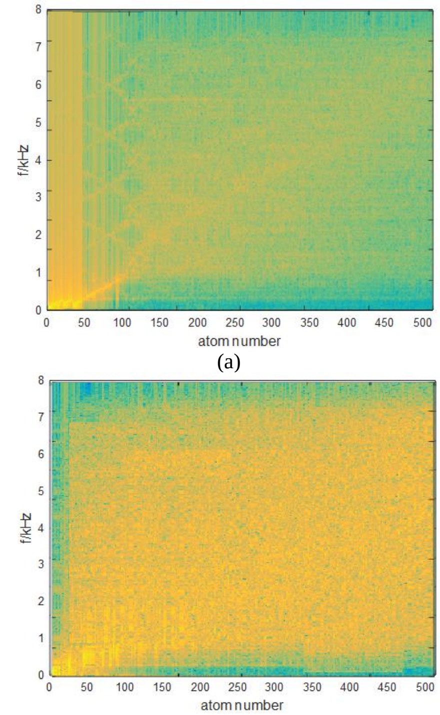
The results of the proposed algorithm were compared with the other voice activity detection methods introduced in this processing field. These methods include the algorithm presented by Sharma1 and Rajpoot [19] that utilizes a zero-crossing rate and clustering procedure and also the VAD method which uses a clustering method based on the Gaussian mixture model. Mavaddaty et al. [6] presented a VAD algorithm based on the energy of the sparse coefficient matrices extracted from the wavelet packet transform features of speech and noise signals.



(a)

(b)

Figure 3. The elliptical plots of the time-frequency energy of the atoms learned by: a) the SMF-based VAD algorithm based on frequency features. b) the proposed method based on spectro-temporal features



(a)

(b)

Figure 4. The spectrogram plot of the atoms learned by: a) the SMF-based VAD algorithm based on frequency features. b) the proposed method based on the spectro-temporal features

TABLE 1. The average accuracy of the proposed VAD algorithm in a speaker-independent scenario

Speaker	#Sentences	Voice	Silent	Average accuracy
Woman	25	97.43	98.25	97.84
Man	25	97.62	98.11	97.86

TABLE 2. The average accuracy of the proposed VAD algorithm in a speaker-dependent scenario

Speaker	#Sentences	Voice	Silent	Average accuracy
Woman	25	98.21	98.36	98.28
Man	25	97.89	98.49	98.19

The sparse coding procedure was based on a combination of orthogonal matching pursuit algorithm (OMP) and coherence criterion.

A VAD algorithm with a combination of convolutional recurrent neural network and a recurrent neural network was proposed by Wang and Zhang [20]. Also, a speech enhancement module was designed to improve the performance of VAD system in low signal-noise ratio conditions. Jordán et al. [21] introduced a VAD system to identify correctly the speech frames based on recurrent neural networks. The model defined in this paper was learned using bidirectional long short-term memory.

A comparison is also made with the method presented by Ahmadi and Joneidi [2], which is based on a sparse representation using the orthogonal matching pursuit (OMP) algorithm and K-SVD dictionary learning algorithm. As mentioned before the presented VAD algorithm employed SNMF learning method with a sparse-based statistical structure as a model learning method that has been widely used in signal processing in recent [22, 23].

These results are presented in Tables 3 and 4. The results show that the proposed method correctly identifies the voice and silent regions of the input speech signal. This success and superiority over other methods can be due to the use of appropriate learned models and the dimension reduction algorithm to eliminate the outlier data during the training step. In these simulations, the results of the speaker-dependent scenario are better than the speaker-independent test, which can be due to the similarity between the speakers in the train and test steps. The results show that employing spectro-temporal features and speech signal processing through the auditory model is a desirable approach to identify the speech frames. The combination of these two techniques has many applications as a pre-processing step in speech signal analysis. The first two rows in Tables 3 and 4 are the same since the methods proposed by Sharma1 and Rajpoot [19], they did not employ the learning-based technique and the detection procedure for them occurs in one step, not in the different scenarios.

To investigate more the performance of the proposed method, the ROC curve obtained from the results of the proposed method and other comparable methods in the speaker-independent and speaker-dependent scenarios are shown in Figures 5 and 6, respectively, which emphasize the capability of the proposed method to achieve high accuracy in detection procedure.

4. 2. Simulation Results in The Presence of Different Noise Signals

The quality of the speech signal can be significantly reduced in the presence of environmental noise signals and lead to the malfunction of hearing aids, automatic speech recognition systems, cell phones, etc. In this paper, a single-channel speech

TABLE 3. The average accuracy percentage of the proposed algorithm and the compared methods to detect the silence/speech sections of the speech signal in the speaker-independent scenario

	#Sentences	Voice	Silent	Average accuracy
Zero crossing-based method [19]	50	95.56	96.67	96.11
GMM-based method	50	97.41	97.78	97.59
Sparse representation-based method [2]	50	97.23	97.89	97.56
sparse dictionary learning-based method [6]	50	97.92	97.97	97.94
NN-based method [20]	50	97.72	97.91	97.81
CRNN-based method [21]	50	97.86	97.90	97.88
Proposed method	50	**98.05**	**98.19**	**98.12**

TABLE 4. The average accuracy percentage of the proposed algorithm and the compared methods to detect the silence/speech areas of the speech signal in the speaker-dependent scenario

	#Sentences	Voice	Silent	Average accuracy
Zero crossing-based method [19]	50	95.56	96.67	96.11
GMM-based method	50	97.41	97.78	97.59
Sparse representation-based method [2]	50	97.69	97.93	97.81
sparse dictionary learning-based method [6]	50	97.98	98.01	97.99
NN-based method [20]	50	97.81	97.99	97.90
CRNN-based method [22]	50	97.93	98.09	98.01
Proposed method	50	**98.14**	**98.24**	**98.19**

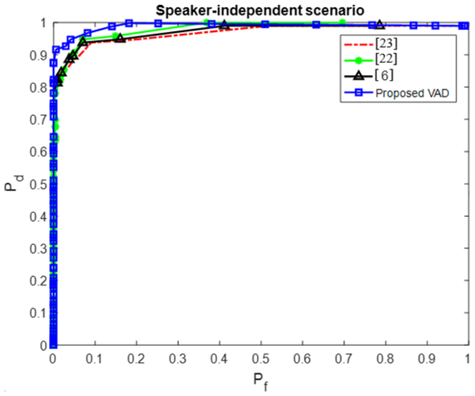


Figure 5. The ROC curve obtained from the results of the proposed method and other compared methods in the speaker-independent scenario

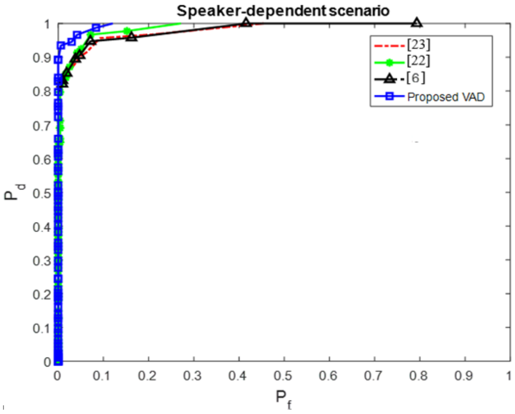


Figure 6. The ROC curve obtained from the results of the proposed method and other compared methods in the speaker-dependent scenario

processing corrupted by additive noise is considered. When the speech signal is exposed to non-stationary noise signals, the performance of the VAD algorithm decreases. This is especially for speech-like noise signals, that have fundamental overlap between the components in the spectro-temporal domain. Although the evaluation of the VAD algorithm is usually not performed in the presence of noise signals, in this paper, the performance of the proposed VAD algorithm in different noise conditions is investigated. In none of the references in which the proposed method has been compared with them, such as [2, 19-23], the performance evaluation in the presence of noise has not been done, so the results of this method have not been reviewed in the noise conditions and only the proposed method has been evaluated.

In this paper, a variety of noise signals consisting of white and babble noises from Noisex92 [24] car and train noises from Aurora2 [25] as well as piano noise from the piano society website[1] have been considered to have a proper investigation about the performance of the proposed method.

The block diagram of the proposed VAD algorithm to determine the labels of the input frames in the presence of the mentioned noise signals is shown in Figure 7. The learning procedures for speech and different noise signals were carried out with the same parameters in the SNMF coding algorithm and the dimension reduction technique.

In recent years, the use of sparse representation techniques for voice activity detector algorithms in a noisy condition has increased. An ideal VAD is used to acquire the data frames needed to learn the noise signal dictionary as reported by Sigg et al. [26]. The data frames obtained by the non-speech frames of the noisy signal are not usually enough to learn a dictionary with low approximation error. The noise dictionary learning algorithm in this approach is performed in the speech enhancement step and leads to a significant increase in the computation time. Also, Sigg et al. [27] presented a generative coherence-based dictionary learning method using the pure noise data to train noise dictionary models. The offline learning process was performed with enough noise signals. In this paper, the advantages of the SNMF technique were utilized to learn the dictionaries for scale, rate, and frequency features. The model learning procedure for the noise signals is done without any problems since adequate noise data is available. This learning process for speech and noise signals is carried

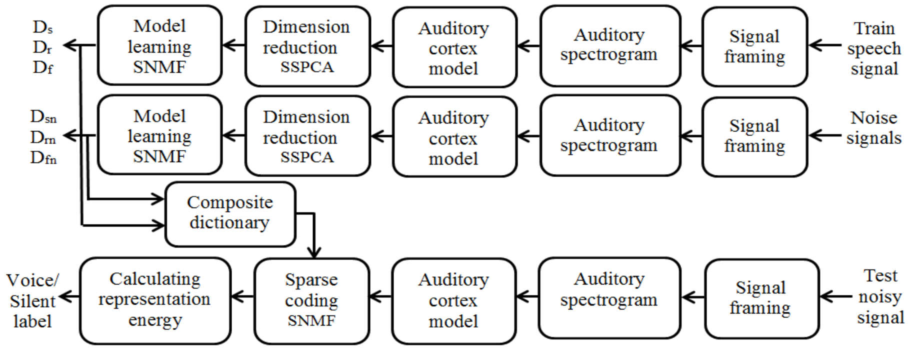


Figure 7. Block diagram of the proposed VAD based on spectro-temporal properties in the presence of noise signals

[1] http://pianosociety.com

out precisely in the same way. The sparse representation in the presence of noise is carried out over a composite dictionary that includes speech and noise models as:

$$\begin{aligned} & \mathrm{H}^*_{\mathrm{S}}, \mathrm{H}^*_{\mathrm{R}}, \mathrm{H}^*_{\mathrm{F}}, \mathrm{H}^*_{\mathrm{Sn}}, \mathrm{H}^*_{\mathrm{Rn}}, \mathrm{H}^*_{\mathrm{Fn}} = \\ & \mathrm{SNMF}\left(\boldsymbol{Y},\ \left[\mathrm{D_s\, D_r\, D_f\ \ D_{sn}\, D_{rn}\, D_{fn}}\right]\right) \end{aligned} \tag{8}$$

$$\begin{aligned} & \mathrm{E_s} = \tfrac{1}{\mathrm{L}} \sum_{\mathrm{l=1}}^{\mathrm{L}} \boldsymbol{H}^{*2}_{s,l}\ \ , \mathrm{E_r} = \tfrac{1}{\mathrm{L}} \sum_{\mathrm{l=1}}^{\mathrm{L}} \boldsymbol{H}^{*2}_{r,l}, \\ & \mathrm{E_f} = \tfrac{1}{\mathrm{L}} \sum_{\mathrm{l=1}}^{\mathrm{L}} \boldsymbol{H}^{*2}_{f,l} \\ & \mathrm{E_{sn}} = \tfrac{1}{\mathrm{L}} \sum_{\mathrm{l=1}}^{\mathrm{L}} \boldsymbol{H}^{*2}_{sn,l}\ \ , \mathrm{E_{rn}} = \tfrac{1}{\mathrm{L}} \sum_{\mathrm{l=1}}^{\mathrm{L}} \boldsymbol{H}^{*2}_{rn,l}, \\ & \mathrm{E_{fn}} = \tfrac{1}{\mathrm{L}} \sum_{\mathrm{l=1}}^{\mathrm{L}} \boldsymbol{H}^{*2}_{fn,l} \end{aligned} \tag{9}$$

where $\mathrm{E_{sn}}$, $\mathrm{E_{rn}}$, and $\mathrm{E_{fn}}$ are the energy of the sparse representation corresponding to scale, rate, and frequency features of each noise signal. Also, $\mathrm{H}^*_{\mathrm{Sn}}$, $\mathrm{H}^*_{\mathrm{Rn}}$ and $\mathrm{H}^*_{\mathrm{Fn}}$ are sparse coefficient matrices related to scale, rate, and frequency features of each noise data class.

This procedure in the train and test steps should be carried out for each noise signal. The learning and dimension reduction procedures for all kinds of noise signals are done the same as a speech signal. According to Equation (8), the input noisy frame is sparsely coded over the composite dictionary $\left[\mathrm{D_s\, D_r\, D_f\ \ D_{sn}\, D_{rn}\, D_{fn}}\right]$. In this test step, the sum of the energies calculated from the sparse coefficient matrices for speech and noise signals is considered. The total energy calculated based on speech and noise model determines the label of the input noisy frames. If this calculated energy over the speech signal model is greater then the calculated energy over the noise model, the input frame is detected as speech frame, otherwise, a noise label is assigned to this frame. Also, if the energy difference calculated on the speech and noise models is less than a certain limit of ε_2=0.08, then the total energy of the sparse coding coefficients for one frame before and one frame after the input frame is calculated over the speech and noise models to obtain the short-term energy of this representation. If the average of these calculated energies on the speech model is higher than the noise model, the speech label is assigned to the input frame, otherwise the noise label.

The average results of the proposed method to assign the proper labels in a speaker-independent scenario in the presence of various noise signals with 10dB SNR are shown in Table 5. Also, these results in a speaker-dependent scenario are reported in Table 6. For more evaluation of the performance of the proposed VAD in different conditions, the average results of the proposed VAD in the speaker-independent and speaker-dependent scenarios in the presence of various noise signals with 5dB SNR are shown in Tables 7 and 8, respectively. From the reported values in Tables 5-8, it can be concluded that the accuracy of the proposed method decreases as the SNR value decreases. Also, the accuracy of labeling to silence/speech in the presence of noise signals with stationary content such as white noise is higher than other noise signals. The accuracy in the presence of periodic piano noise signal with harmonic structure is more accurate than in other conditions. In

TABLE 5. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-independent scenario and the presence of noise signals with SNR=10dB

	#Sentences	**Voice**	**Silent**	**Average accuracy**
Without noise	100	98.05	98.19	98.12
White noise	100	97.20	97.01	97.10
Car noise	100	94.98	95.47	95.22
Piano noise	100	97.02	97.21	97.11
Babble noise	100	94.20	94.22	94.21
Train noise	100	94.59	95.23	94.91

TABLE 6. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-dependent scenario and the presence of noise signals with SNR=10dB

	#Sentences	**Voice**	**Silent**	**Average accuracy**
Without noise	100	98.14	98.24	98.19
White noise	100	97.51	97.18	97.34
Car noise	100	95.23	95.76	95.49
Piano noise	100	97.21	97.33	97.27
Babble noise	100	94.28	94.36	94.32
Train noise	100	94.76	95.32	95.04

TABLE 7. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-independent scenario and the presence of noise signals with input SNR=5dB

	#Sentences	**Voice**	**Silent**	**Average accuracy**
Without noise	100	98.05	98.19	98.12
White noise	100	94.11	94.26	94.18
Car noise	100	92.22	93.31	92.76
Piano noise	100	94.47	94.63	94.55
Babble noise	100	92.13	92.47	92.30
Train noise	100	92.65	93.06	92.85

TABLE 8. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-dependent scenario and the presence of noise signals with SNR=5dB

	#Sentences	Voice	Silent	Average accuracy
Without noise	100	98.14	98.24	98.19
White noise	100	94.73	94.41	94.57
Car noise	100	92.51	93.49	93.00
Piano noise	100	94.70	94.88	94.79
Babble noise	100	92.53	92.66	92.59
Train noise	100	92.84	93.30	93.07

addition, the results of the proposed VAD algorithm has been considered in the presence of the car noise signal that has a stationary structure. But in the presence of babble noise, which is very similar to the speech signal, accuracy is greatly reduced. It should be noted that in the speaker-dependent scenario, the results are slightly higher than in the speaker-independent scenario in different situations because there is an overlap between the speakers in the train and test steps. Therefore, it can be concluded that the best results are obtained in the high SNR value, in the presence of white and piano noise signals, and the speaker-dependent scenario. Also, the performance of labeling in the case of speech frames that consist of consonant letters such as fricatives that have a similar structure to the noise signal may be decreased. The average accuracy values in the speaker-dependent and independent scenarios evaluated at two SNR values of 10dB and 5dB are represented in Figure 8. These results are obtained for a clean speech signal case and five stationary, non-stationary and periodic noises: white, car, train, babble, and piano signals. In general, it can be concluded that the reported results emphasize that the proposed VAD has an appropriate performance in different noisy conditions.

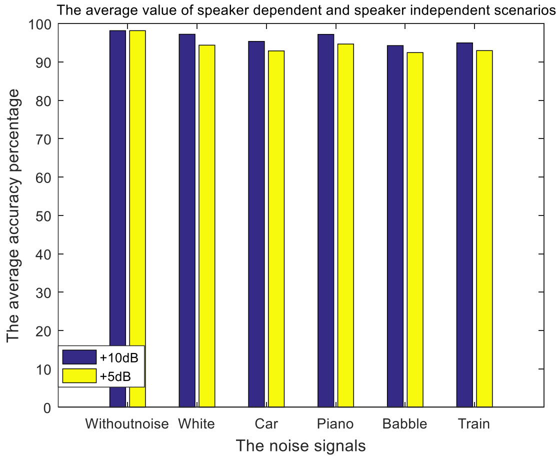


Figure 8. Performance comparison of the proposed method in terms of average accuracy in speaker-dependent and independent cases in 10dB and 5dB SNRs

5. CONCLUSION

Voice activity detection methods are very effective in the various fields of signal analysis and speech processing as a pre-processing block. In this paper, this detection procedure is performed in the space of spectro-temporal features. The features extracted from this space are used to learn comprehensive models of the input data structure. The dimension of these feature matrices is reduced by the SSPCA algorithm. Then the resulted data are used to learn models using the SNMF method which has a sparse-based statistical structure. In the following, by computing the energy derived from the representation of the input frame features on the composite model, the label of the input frame is identified. Also, these results have been examined for an extensive range of noise types including the stationary, non-stationary, and periodic noise signals in two SNR values of 5dB and 10dB. The simulation results in both speaker-independent and speaker-dependent scenarios indicate the superior performance of the proposed method compared to the other methods presented in this processing field.

6. CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7. REFERENCES

1. Park, J.-S., Yoon, J.-S., Seo, Y.-H. and Jang, G.-J., "Spectral energy based voice activity detection for real-time voice interface", ***Journal of Theoretical & Applied Information Technology***, Vol. 95, No. 17, (2017).
2. Ahmadi, P. and Joneidi, M., "A new method for voice activity detection based on sparse representation", in 2014 7th International Congress on Image and Signal Processing, IEEE. (2014), 878-882.
3. You, D., Han, J., Zheng, G. and Zheng, T., "Sparse power spectrum based robust voice activity detector", in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. (2012), 289-292.
4. You, D., Han, J., Zheng, G., Zheng, T. and Li, J., "Sparse representation with optimized learned dictionary for robust voice activity detection", ***Circuits, Systems, and Signal Processing***, Vol. 33, (2014), 2267-2291. doi: 10.1007/s00034-014-9748-y.
5. Teng, P. and Jia, Y., "Voice activity detection via noise reducing using non-negative sparse coding", ***IEEE Signal Processing Letters***, Vol. 20, No. 5, (2013), 475-478. doi: 10.1109/LSP.2013.2252615.

6. Mavaddaty, S., Ahadi, S.M. and Seyedin, S., "Speech enhancement using sparse dictionary learning in wavelet packet transform domain", ***Computer Speech & Language***, Vol. 44, (2017), 22-47. doi: 10.1016/j.csl.2017.01.009.

7. Chi, T., Ru, P. and Shamma, S.A., "Multiresolution spectrotemporal analysis of complex sounds", ***The Journal of the Acoustical Society of America***, Vol. 118, No. 2, (2005), 887-906. doi: 10.1121/1.1945807.

8. Elhilali, M., Chi, T. and Shamma, S.A., "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility", ***Speech Communication***, Vol. 41, No. 2-3, (2003), 331-348. doi: 10.1016/S0167-6393(02)00134-6.

9. Elhilali, M., Fritz, J.B., Klein, D.J., Simon, J.Z. and Shamma, S.A., "Dynamics of precise spike timing in primary auditory cortex", ***Journal of Neuroscience***, Vol. 24, No. 5, (2004), 1159-1172. doi: 10.1523/JNEUROSCI.3825-03.2004.

10. Hoyer, P.O., "Non-negative matrix factorization with sparseness constraints", ***Journal of Machine Learning Research***, Vol. 5, No. 9, (2004). doi: 10.48550/arXiv.cs/0408058.

11. Ullah, R., Islam, M.S., Ye, Z. and Asif, M., "Semi-supervised transient noise suppression using omlsa and snmf algorithms", ***Applied Acoustics***, Vol. 170, (2020), 107533. doi: 10.1016/j.apacoust.2020.107533.

12. Ullah, R., Islam, M.S., Hossain, M.I., Wahab, F.E. and Ye, Z., "Single channel speech dereverberation and separation using rpca and snmf", ***Applied Acoustics***, Vol. 167, (2020), 107406. doi: 10.1016/j.apacoust.2020.107406.

13. Jolliffe, I.T., "Principal component analysis for special types of data, Springer, (2002).

14. Zou, H., Hastie, T. and Tibshirani, R., "Sparse principal component analysis", ***Journal of Computational and Graphical Statistics***, Vol. 15, No. 2, (2006), 265-286. doi: 10.1198/106186006X113430.

15. Jenatton, R., Obozinski, G. and Bach, F., "Structured sparse principal component analysis", in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings. Vol., No. Issue, (2010), 366-373.

16. Jenatton, R., Audibert, J.-Y. and Bach, F., "Structured variable selection with sparsity-inducing norms", ***The Journal of Machine Learning Research***, Vol. 12, (2011), 2777-2824. doi: 10.48550/arXiv.0904.3523.

17. Kapadia, S., Valtchev, V. and Young, S.J., "Mmi training for continuous phoneme recognition on the timit database", in 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE. Vol. 2, (1993), 491-494.

18. Jafari, M.G. and Plumbley, M.D., "Speech denoising based on a greedy adaptive dictionary algorithm", in 2009 17th European Signal Processing Conference, IEEE. (2009), 1423-1426.

19. Sharma, P. and Rajpoot, A.K., "Automatic identification of silence, unvoiced and voiced chunks in speech", ***Journal of Computer Science & Information Technology (CS & IT)***, Vol. 3, No. 5, (2013), 87-96. doi: 10.5121/csit.2013.3509.

20. Wang, G.-B. and Zhang, W.-Q., "An rnn and crnn based approach to robust voice activity detection", in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE. (2019), 1347-1350.

21. Jordán, P.G., Bailo, I.V., Giménez, A.O., Artiaga, A.M. and Solano, E.L., "Vivovad: A voice activity detection tool based on recurrent neural networks", ***Jornada de Jóvenes Investigadores del I3A***, Vol. 7, (2019). doi: 10.26754/jji-i3a.003524.

22. Mavaddati, S., "Voice-based age and gender recognition using training generative sparse model", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 31, No. 9, (2018), 1529-1535. doi: 10.5829/ije.2018.31.09c.08.

23. Sabzalian, B. and Abolghasemi, V., "Iterative weighted non-smooth non-negative matrix factorization for face recognition", ***International Journal of Engineering, Transactions A: Basics***, Vol. 31, No. 10, (2018), 1698-1707. doi: 10.5829/ije.2018.31.10a.12.

24. Varga, A. and Steeneken, H.J., "Assessment for automatic speech recognition: Ii. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems", ***Speech Communication***, Vol. 12, No. 3, (1993), 247-251. doi: 10.1016/0167-6393(93)90095-3.

25. Hirsch, H.-G. and Pearce, D., "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW). (2000).

26. Sigg, C.D., Dikk, T. and Buhmann, J.M., "Speech enhancement with sparse coding in learned dictionaries", in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE. (2010), 4758-4761.

27. Sigg, C.D., Dikk, T. and Buhmann, J.M., "Speech enhancement using generative dictionary learning", ***IEEE Transactions on Audio, Speech, and Language Processing***, Vol. 20, No. 6, (2012), 1698-1712. doi: 10.1109/TASL.2012.2187194.

Persian Abstract

چکیده

آشکارسازهای فعالیت صوتی برای استخراج بخش‌های سکوت/صوت سیگنال گفتار برای حذف سیگنال‌های مختلف نویز پس‌زمینه ارائه شده‌اند. در این مقاله یک آشکارساز فعالیت صوتی جدید با استفاده از ویژگی‌های طیفی-زمانی استخراج شده از مدل شنیداری سیگنال گفتار پیشنهاد شده است. پس از استخراج ویژگی‌های مقیاس، نرخ و فرکانس از این فضای ویژگی، از یک الگوریتم تجزیه و تحلیل مولفه‌های اساسی ساختارمند تُنُک برای در نظر گرفتن اجزای اصلی این ویژگی‌ها و کاهش ابعاد داده‌های یادگیری استفاده می‌شود. سپس این بردارهای ویژگی برای یادگیری مدل توسط الگوریتم فاکتورسازی ماتریس تُنُک غیرمنفی استفاده می‌شوند. روش یادگیری مدل برای نشان دادن هر بردار ویژگی با نرخ تُنُک مناسب براساس اتم‌های انتخاب شده انجام می‌گیرد. تشخیص فعالیت صوتی فریم‌های ورودی با محاسبه انرژی نمایش تُنُک برای هر فریم ورودی بر روی مدل ترکیبی انجام می‌شود. اگر انرژی محاسبه شده از یک آستانه مشخص فراتر رود، نشان می‌دهد که قاب ورودی ساختاری مشابه اتم‌های مدل آموخته شده دارد و نتیجه گیری می‌شود که قاب مشاهده شده دارای محتوای صوتی است. نتایج آشکارساز پیشنهادی با سایر روش‌ها و طبقه‌بندی‌کننده‌های پایه در این زمینه از پردازش سیگنال گفتار مقایسه می‌شود. این نتایج در حضور نویزهای ایستا، غیرایستا و متناوب بررسی شده و نشان داده می‌شود که روش پیشنهادی مبتنی بر یادگیری مدل با ویژگی‌های طیفی-زمانی می‌تواند به درستی فعالیت‌های سکوت/گفتار را تشخیص دهد.

International Journal of Engineering

Journal Homepage: www.ije.ir

Joining Two Dissimilar Metals of Aluminum 5052 to Austenitic Stainless Steel 304 using Ultrasonic Friction Stir Welding

W. S. Abdullah[a], M. Shakeri*[a], M. Habibnia[b]

[a] *Faculty of Mechanical Engineering, Babol Noshirvani University of Technology, Babol, Iran*
[b] *Faculty of Mechanical Engineering, Islamic Azad University, Jouybar Branch, Jouybar, Iran*

PAPER INFO

Paper history:
Received 8 May 2023
Received in revised form 23 May 2023
Accepted 24 May 2023

Keywords:
Friction Stir Welding
Ultrasonic Vibrations
Aluminum 5052
Austenitic Stainless Steel 304

ABSTRACT

In this research, the joining of aluminum alloy 5052 to austenitic stainless steel 304 was investigated. For this purpose, friction stir welding process was used in two modes with and without ultrasonic vibrations. In order to achieve the best welding quality in terms of mechanical and metallurgical properties, welding parameters such as rotational speed, linear speed and frequency were investigated. The aim of this research is to obtain a sample with the best mechanical and metallurgical properties and the lowest residual stress. As a research innovation and the aim of measuring the values of residual stress created in the samples after the welding operation, the new method of drilling and Digital Image Correlation was used. Finally, by examining the results, it has been determined that ultrasonic vibrations have improved the mechanical and metallurgical properties about 15% to a large extent. In order to evaluate the accuracy of the results related to the residual stress, all the samples were subjected to the central drilling test by installing a strain gauge, and it was found that the error is less than 10% and obtained results were accurate and appropriate.

***doi**: 10.5829/ije.2023.36.08b.09*

1. INTRODUCTION

Friction stir welding process is a developed method of friction welding [1]. In this process, due to the friction of a small rotating tool resistant to wear and heat, the necessary heat to change the shape of the material is obtained [2]. This process is a combination of plastic deformation and severe liquefaction of the material in the welding zone. Welding parameters determine the flow pattern of the material and the temperature distribution in the joint area, and as a result, they will have a direct effect on the microstructure of this area [3]. Applying the correct arrangement of these parameters requires accurate knowledge of them, as well as checking the background of researches and conducting trial and error experiments. Habibnia et al. [4] investigated the effect of parameters of rotation speed, weld speed, penetrant depth and tool shoulder diameter on the microstructure and defects created during friction stir welding of 5050 aluminum alloy and 304 stainless steel. Boonchouytan et al. [5] investigated the bonding of 6061 aluminum alloys obtained by semi-solid forming method using friction stir welding method. Liu et al. [6] investigated the effect of welding speed parameter on the mechanical properties of 2219 aluminum alloy. El-sayed et al. [7] evaluated the temperature distribution and residual thermal stresses created by the friction stir welding process of 5083 aluminum sheets at different rotational and linear speeds by threaded and conical cylindrical pins. Hongjun et al. [8] evaluated the fatigue life in friction stir welding of different grades of aluminum. In some researches, in order to improve the conditions of friction stir welding, this process is combined with another process. Thoma et al. [9] investigated the joining of steel to aluminum using the effect of ultrasound in friction stir welding. According to the results, it was found that with the use of ultrasound, the dispersion of steel particles in the welding area has become more uniform. Benfar et al. [10]

*Corresponding Author Email: Shakeri@nit.ac.ir (M. Shakeri)

Please cite this article as: W. S. Abdullah, M. Shakeri, M. Habibnia, Joining Two Dissimilar Metals of Aluminum 5052 to Austenitic Stainless Steel 304 using Ultrasonic Friction Stir Welding, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1489-1501

investigated the effect of using ultrasound in friction stir welding on the corrosion rate. Thoma et al. [11] investigated the benefits of using ultrasound in friction stir welding of non-homogeneous steel-aluminum alloy joints. Hong et al. [12] compared the use of ultrasound in friction stir welding of dissimilar metals. It has been observed that with the use of ultrasound, the amount of intermetallic structure has decreased to a great extent, and this effect has also increased the final strength of the weld. In this research, the feasibility of joining two metals, austenitic stainless steel 304 and aluminum alloy 5052, by means of friction stir welding in a thickness of 3 mm has been investigated. The two metals have high corrosion resistance, and their connection can be used in various industries such as shipbuilding, automotive industries and other industries. By combining these two metals, the properties of both metals can be used. In places where high strength is required from steel and in cases where light weight is required, aluminum alloys can be used [4]. Today, making parts with different materials with the aim of improving efficiency is an interesting idea in the engineering industry [13]. The connection of aluminum alloy to steel has many problems due to different mechanical and thermal properties, including a large difference in melting temperature [4]. To connect these two metals, many methods such as melting and non-melting welding are used. There are many problems in melting methods due to differences in melting temperature and other properties. Among these problems, we can mention the creation of intermetallic compounds and rapid cooling, which causes the weld area to become brittle. Also, in the connection by melting methods, the chromium in steel, which is one of the reasons for its resistance to corrosion, turns into chromium carbide and reduces the corrosion resistance [14]. As an innovation in this research, the friction stir welding method using ultrasonic effect has been used to connect these two non-homogeneous aluminum alloy 5052 to austenitic stainless steel 304. The next innovation is the way of measuring the values of residual stress created in the samples after the welding operation, and this is new method of drilling and Digital Image Correlation.

2. EXPERIMENTAL WORK

In this research, the friction stir welding process has been performed on cold rolled sheets of aluminum 5052 and austenitic stainless steel 304 with a thickness of 3 mm. The samples are cut by guillotine in dimensions of 100×150 mm and are milled to make the joint edge parallel and flat. Also, before the test, the edge required for the test is cleaned with a file and the oxides are removed. According to Figure 1, after preparing the samples, pictures related to the microstructure were taken from the surfaces of the aluminum and steel samples. Fixtures are used for clamping parts and also for correct positioning. The surface of the fixture is ground so that the penetration depth remains constant in all welding points. For the tool to move correctly on the connecting line and not to deviate, two holes are created at both ends of the plate parallel to the horizontal axis of the device. A line is drawn between these two holes. Two steel and aluminum pieces are placed face to face on this line and are secured by two machined and perforated blocks that are placed on the plate. The fixture plate is connected to the table of the milling machine by means of two T-shaped numbers. To determine the chemical composition percentage of these alloys, a material analysis device (spectrometry) has been used. The chemical composition of austenitic stainless steel 304 and aluminum 5052 are shown in Tables 1 and 2.

To perform the friction stir welding process, two rotary and linear movements of the tool are needed. To provide these two movements, a vertical milling machine

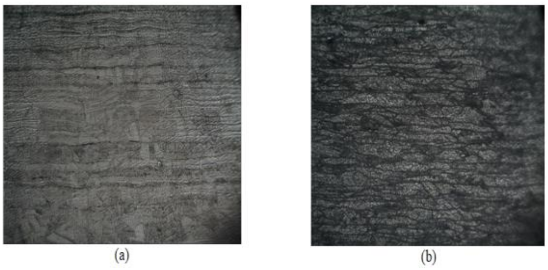


Figure 1. Microstructure of (a) steel 304 (b) aluminum 5052 at scale 40x

TABLE 1. Chemical composition of aluminum 5052

Elements (%)	Cr	Al	Other	V	Si	Mn	Mg	Fe	Cu
Aluminum 5052	0.030	0.314	8.100	0.020	18.000	0.003<	0025	1.090	0.490

TABLE 2. Chemical composition of austenitic stainless steel 304

Elements (%)	C	Other	Ni	Mo	Cr	S	P	Mn	Si	Fe
Stainless steel 304	0.030	0.314	8.100	0.020	18.000	0.003<	0.025	1.090	0.490	73.930

made by Tabriz Machinery has been used. The rotational speed range of this machine is 45-1500 rpm and the linear speed of the table is 28-900 mm/min. The friction stir welding tool has a decisive role in the quality of the connection. To obtain an acceptable connection, one of the important parameters is the tool wear factor. In the connection of these two metals, in order to minimize the amount of tool wear, tungsten carbide tools with the specifications presented in Table 3 have been used. Machining of this material by lathe and milling machine is not possible due to its high hardness. Grinding machining technique is used to make tools. The tool used to machine this type of material is CBN. The schematic drawing of the tool used in this research is shown in Figure 2.

A milling machine and a fixture are commonly used to perform the friction stir welding process. But to do the test using ultrasound, a set has been designed and built that can create vibration with the desired frequency and amplitude during the process. An ultrasound machine consists of four main parts. The first part is the power supply, which is responsible for changing the frequency from 50 Hz to 20-30 kHz. The next part is the transducer, which performs the task of converting electrical frequency to mechanical frequency. The third part is the signal amplifier and the last part is the horn, which is responsible for transmitting the vibration. All the above items should be placed on a structure and also a fixture for welding should be designed and built. A view of this collection is shown in Figure 3.

According to the vibration direction and amplitude, this set should be designed and built to be installed on the vertical milling machine. In order to investigate the effect of ultrasound on the obtained results, the results will be compared with conventional friction stir welding without vibration. Many parameters are effective in achieving better efficiency in welding quality. In this research, some welding parameters, line frequency of vibration, feed rate and rotational speed are investigated. The investigated parameters and their levels are shown in Table 4.

TABLE 3. Specifications of the tools used

Pin shape	Shoulder shape	Pin diameter (mm)	concavity angle (degree)	Pin height (mm)	Shoulder diameter (mm)
Cylinder	Cylinder	6	3	2.8	20

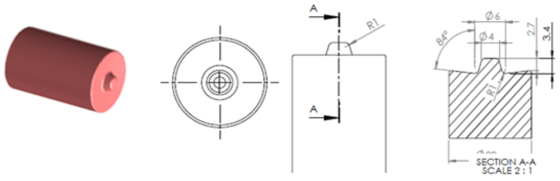


Figure 2. Friction stir welding tool

In order to investigate the effect of each parameter, full factorial design of experiment was conducted. According to the examined parameters, 27 tests should be performed, which are done in form.

3. RESULTS AND DISCUSSION

In order to investigate the effect of input parameters on the connection area, mechanical and metallurgical tests were used, each of these tests was performed according to the desired standard.

3. 1. Tensile Test

To investigate the tensile behavior of the parts resulting from the friction stir welding process, samples from the welded area were prepared and subjected to tension. Using the results of the tensile test, useful information can be obtained about the ultimate strength, percentage of elongation and toughness. In this research, tensile test samples were prepared based on ASTM E8 standard. The schematic of standard sample is shown in Figure 4 and the samples prepared for the tensile test are shown in Figure 5.

TABLE 4. The levels of the investigated parameters

Frequency (KHz)	Welding speed (mm/min)	Rotational speed (RPM)
15	28	600
20	40	800
25	60	1000

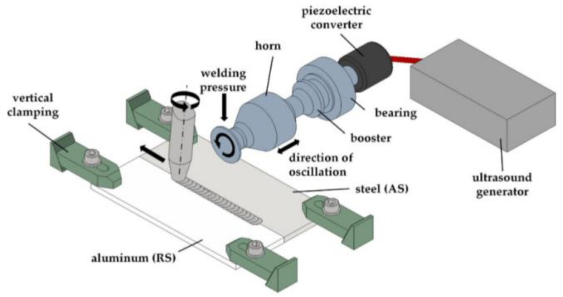


Figure 3. Ultrasonic friction stir welding equipment

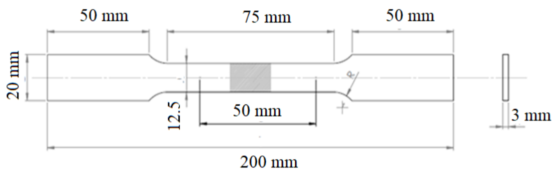


Figure 4. Dimensions of tensile test sample according to ASTM E8 standard

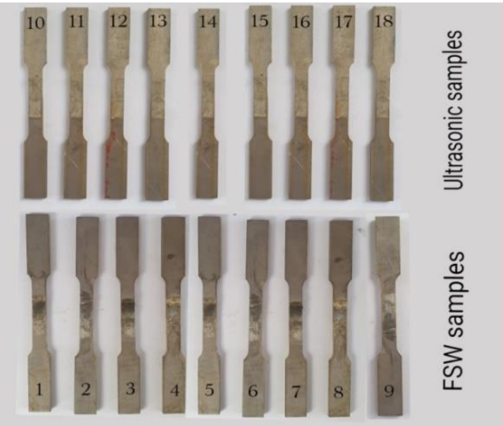


Figure 5. Prepared samples for welding

After preparing the samples, all of them were subjected to a tensile test and stretched until breaking. The samples welded by the usual friction stir welding process and with ultrasonic vibrations are shown in Figure 6. According to the tensile test results, it has been determined that the samples subjected to welding with ultrasonic vibrations have better mechanical properties than the samples with friction stir welding normally. During the friction stir welding process, due to the vibration of the work piece, the materials in the stir area get more strain than the same materials in the friction stir welding process. According to the researches, there is a direct relationship between the strain and the density of dislocations in the strained material. It is predicted that the density of dislocations in the stir zone of the vibrating friction welding part is more than its number in the other stir zone of the welded part, and therefore, during the dynamic recrystallization process, more high-angle boundaries are formed and the welding zone formed with a smaller grain size. Tensile test diagram of a sample without vibration and with vibration after the tensile test is shown in Figure 7. According to Figure 7, it has been determined that the sample welded by ultrasonic friction stir method has a higher ultimate tensile strength and elongation percentage than the friction stir welding sample. The reason for this can be the smaller grain size of the friction stir welding sample. According to the Hall-Patch relationship, strength has a direct relationship with the inverse of the square of the grain size, and the strength will increase as the grain size decreases. In fact, with the reduction of the grain size, the volume component of the grain boundaries will increase, and since the grain boundaries act as an obstacle against the movement of dislocations, the strength decreases with the reduction of the grain size. According to the mentioned cases, it can be seen that by performing ultrasonic friction stir welding and reducing the grain size in the stir zone, since the mechanical properties of the stir zone are improved. Therefore, the strength and the elongation percentage of the welded sample will also increase.

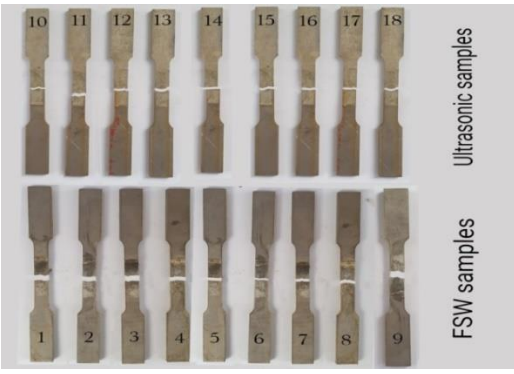


Figure 6. Welded samples after tensile test

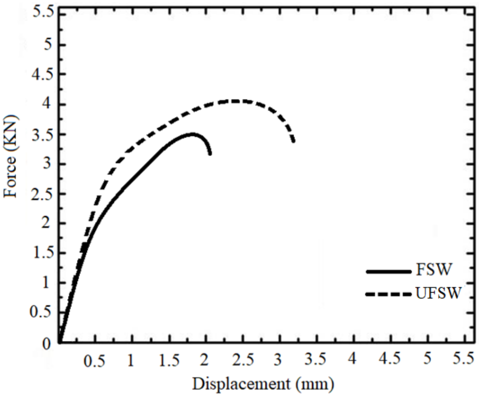


Figure 7. Force-displacement diagram of friction stir welding process and ultrasonic friction stir welding process

Although most of the conducted researches emphasize the reduction of grain size and its effect on increasing strength, there is no common opinion regarding the changes in elongation in the tensile test in terms of grain size changes. According to Figure 7, it is clear that with the reduction of the grain size in the welding area, the percentage of elongation has increased. The reason for this can be attributed to the increase in the number of boundary dislocations with the increase in the volume component of the grain boundaries and the other issue is the further prevention of grain boundaries from crack growth in fine-grained materials. Grain boundaries are the place of accumulation of geometrically essential dislocations. Since the possibility of plastic deformation increases with an increase in the number of dislocations, it is predicted that the length change due to tension will increase with the decrease in the grain size, also the results of the investigations have shown that with the decrease in the size due to the change of the fracture mechanism from grain boundary to multigrain, the elongation percentage increases. The value of welded samples maximum force using the ultrasonic friction stir method is shown in Figure 8.

In general, it can be seen that with the increase in frequency, the strength of the parts and the percentage of

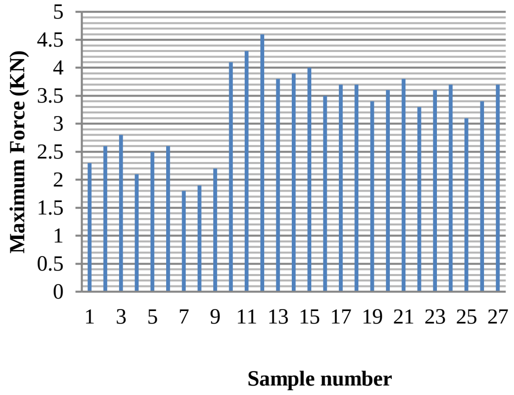


Figure 8. The values of the maximum force applied to friction stir welding samples

length increase have also increased. The reason for this can be attributed to the effect of vibration. By increasing the vibration frequency, the strain rate increases during the friction stir welding process and more dislocations are produced during welding. Since the main mechanism for fineness in the friction stir welding process is dynamic recrystallization, with an increase in the production of dislocations, more recrystallization takes place and as a result, the microstructure with smaller grains is obtained. This issue can also be interpreted using the Zener-Holoman variable. As the strain increases according to Equation (1), the value of the Zener-Holman variable (Z) increases.

$$Z=\dot{\varepsilon}.\exp(\frac{Q}{RT}) \qquad (1)$$

In Equation (1), $\dot{\varepsilon}$ is the strain rate, T is the working temperature in degrees Kelvin, Q is the activation energy, and R is the gas constant. The relationship between the parameter Z and the average size of sub-grains or dynamically crystallized grains follows the relation D-1=aLnZ-b. In such a way that D is the average size of grains, Z is the Zener-Lonman parameter, and a and b are positive numbers. By increasing the variable Z according to Equation (1), the grain size decreases and as a result, according to the Hall-Patch relation, the strength value increases.

3. 2. Microhardness Measurement

To study the effects of friction stir welding on the hardness of the samples and check the hardness changes in different areas, after welding, the sheets were cut perpendicular to the process direction and molded and mounted with a special resin solution. Due to the very low force in the microhardness test and its small effect on the welding surface, the surface of the samples should be polished well before the test. To do this, the samples are rasped by carbide rasping sheets up to 1500 series. To study the macrostructure of friction stir welding samples, the sheets are cut and mounted in the direction perpendicular to the process. The mounted samples are first rasped by carbide rasping sheets up to 5000 series and then polished by Al_2O_3 powder. An optical microscope has been used to investigate the distribution of steel particles in the disturbed area and also to see defects such as holes and cracks. In joining two metals of aluminum alloy to stainless steel, there is a possibility of creating intermetallic structures, X-ray spectroscopy is used to identify these intermetallic structures. In this method, by measuring the Bragg angle (α2) and the distance between the crystal plates, the compounds in the desired area can be obtained.

According to the investigations, the hardness value in the stir area has increased for all the samples, and its value will decrease by moving towards the two base metals. The hardness value for 304 steel base metal is around 210 HV and for 5052 aluminum around 80 HV. By moving from the side of the base metals to the stir zone, the hardness values have increased. Hardness profile for a welded sample is shown in Figure 9. Research shows that the change of hardness in friction stir welding for aluminum alloys that have the ability to be heat treated is different from the alloys without the ability to be heat treated. In friction stir welding for some heat treatable aluminum alloys, the middle region of the welded section has less hardness than other regions. Hardness values of the samples after friction stir welding are shown in Tables 5 and 6. According to the investigations, it has been determined that at a constant weld speed, the hardness value in the stir area has decreased as average of 25% with an increase in the rotational speed of the tool.

Relation between all parameters and their effects on each other is considerable. ANOVA analysis for selected factorial is shown in Table 7.

The Model F-value of 90.63 implies the model is significant. There is only a 0.01% chance that an F-value this large could occur due to noise. P-values less than 0.05 indicate model terms are significant. In this case A, B, C, AB are significant model terms. Values greater than0.10 indicate the model terms are not significant. If there are many insignificant model terms (not counting those required to support hierarchy), model reduction

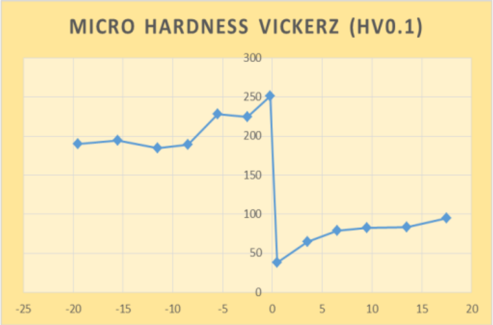


Figure 9. Micro hardness profile of sample 1

TABLE 5. Hardness values of sample 1 welded by ultrasonic friction stir method

Distance from weld center (mm)	-8 (base metal of steel area)	-5	-4	-3	-2	-1
Hardness (HV)	200	230	226	224	242	247
Distance from weld center (mm)	0	1	2	3	4	**8 (base metal of aluminum area)**
Hardness (HV)	250	45	49	62	68	81

TABLE 6. Hardness values of the welded samples by the ultrasonic friction stir method

Sample number	Hardness in aluminum zone (HV)	Hardness in steel zone (HV)
1	68	242
2	63	229
3	59	220
4	73	256
5	71	248
6	66	245
7	80	268
8	72	258
9	70	252
10	52	231
11	50	222
12	45	213
13	64	239
14	64	231
15	57	223
16	70	249
17	63	241
18	59	234
19	47	222
20	43	212
21	40	203
22	56	234
23	52	232
24	49	221
25	62	248
26	58	246
27	55	229

TABLE 7. ANOVA for selected factorial

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	2513.26	10	251.33	90.63	< 0.0001	**significant**
A-Rotational speed	1316.07	2	658.04	237.29	< 0.0001	
B-Welding speed	848.30	2	424.15	152.95	< 0.0001	
C-Frequency	291.63	2	145.81	52.58	< 0.0001	
BC	57.26	4	14.31	5.16	0.0073	
Residual	44.37	16	2.77			
Cor Total	2557.63	26				

may improve your model. Relation between parameters and their effects on hardness at different frequencies is shown in Figures 10-12.

Microstructure of the weld zone of the sample welded by ultrasonic friction stir process is shown in Figure 13. According to the images related to the microstructure, it has been determined that with the increase in the linear speed of the tool, the grain size in the welding area has decreased, and with an increase in the rotational speed of the tool, the grain size in the welding area has increased. Based on the researches, it has been determined that by increasing the rotation speed of the tool and decreasing the linear speed, more heat will be generated in the

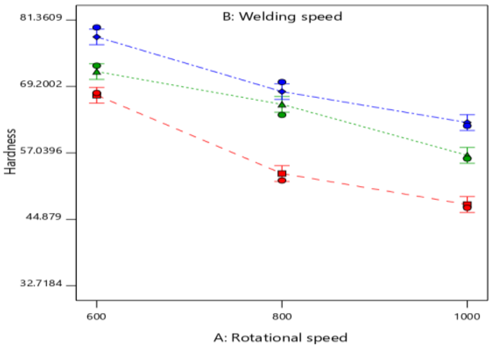


Figure 10. Relation between parameters at frequency 15KHz

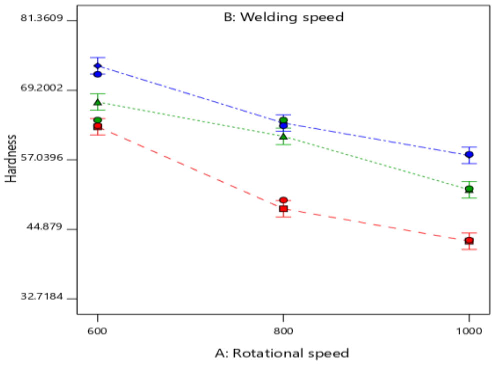


Figure 11. Relation between parameters at frequency 20KHz

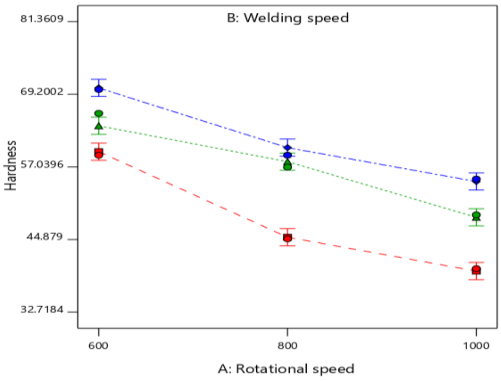


Figure 12. Relation between parameters at frequency 25KHz

welding area during the welding operation. The high heat generated in the friction stir welding process will cause grain size growth in the weld zone.

According to Figure 13, it is clear that the mixing operation between steel and aluminum has been done well and the particles are observed in fine form in each other's structure. There are no visible defects, holes, and cracks, and in the aluminum nugget, the grains have elongation, which actually have flowed.

3. 3. Residual Stress Residual stress is the stress that will still exist in the part after loading. This tension is beneficial in some parts and harmful in some parts [15].

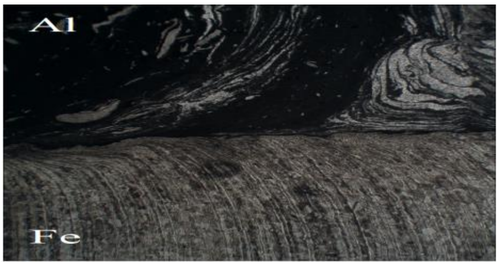


Figure 13. Microstructure of the aluminum-steel joint at scale 10x

There are different methods for measuring residual stress, and in this research, the method of Digital Image Correlation (DIC)-Central Hole Drilling is used. In the method of DIC, first, a random black and white speckle pattern is created on the surface of the part. After preparing the sample, before and after loading, two pictures of the speckle pattern of the part surface are taken, and then by analyzing these two pictures in the correlation algorithm, the field of displacement and strain can be obtained. A schematic of the DIC method equipment is shown in Figure 14. The main idea of this method is how to establish a connection between the points before and after the change of shape in the examined material. The method of DIC does this by using sub-parts of the reference photo, which are known as subsets, and determines their relative position. For each subset, displacement and strain information is calculated during the transfer to match the position of the subsets in the current condition. The final result of a network includes displacement and strain information according to the reference configuration information. In the method of DIC, the light intensity of each photo is estimated with a continuous polynomial function. Sutton et al. [16] showed in an article that the 5th degree curve shows the best results. Each time, the mapping algorithm compares the light intensity function of two subsets of two images before and after loading with dimensions of N×N pixels, and selects that subset of the photo after loading, which has the most agreement with the subset of the reference photo, as the subset of change. It considers the finding and obtains its displacement and deformations (according to Figure 15). This process is done for all the subsets of the reference image and finally the total displacement field is obtained. In order to check the degree of conformity of each pair of subsets, the correlation coefficient C is defined as Equation (2), which can be a suitable criterion for understanding the degree of conformity of two corresponding subsets [16].

$$C(R)=\frac{\sum_{i=-m}^{i=m}\sum_{j=-m}^{j=m}\left(G_r\left(X_p,Y_p\right)-G_d\left(X'_p,Y'_p\right)\right)^2}{\sum_{i=-m}^{i=m}\sum_{j=-m}^{j=m}\left(G_r\left(X_p,Y_p\right)\right)^2} \tag{2}$$

Which in Equation (2):

$X_p = x_p + i$

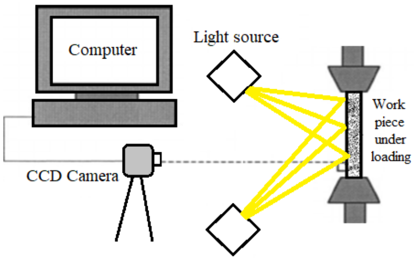


Figure 14. Schematic of the DIC process

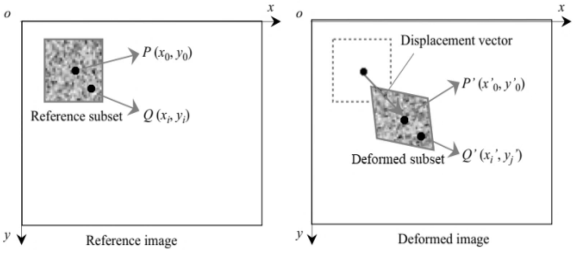


Figure 15. Reference and deformed subsets

$$Y_p = y_p + j$$

$$X_p^{'} = x_p + i + U_s(i,j)$$

$$Y_p^{'} = y_p + j + V_s(i,j)$$

And R is the unknown vector as follows:

$$R = (X, Y, U, V, \frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial y})$$

In above equations, U and V are displacement components, G_r andare continuous functions of G_d light intensity interpolation before and after loading. (x, y) and (x', y') are points coordinates in subsets of reference and deformed image that relate to each other according to Equations (3) and (4) [16].

$$X' = x + U + \frac{\partial U}{\partial x}\Delta x + \frac{\partial U}{\partial y}\Delta y \tag{3}$$

$$Y' = y + V + \frac{\partial V}{\partial x}\Delta x + \frac{\partial V}{\partial y}\Delta y \tag{4}$$

In Equations (3) and (4), Δx and Δy are horizontal and vertical distances of the point (x, y) from the subset center. In correlation relation, the amount of light intensity at each point of reference image subset is compared with the same subset in the image after loading and their difference is obtained. Then the square of their difference is divided by the square of light intensity of that point in the reference image. The obtained number is a measure of the relative error at that point. To calculate the sum of total error in a subset, the error values of the points are added together, when the correlation coefficient is zero, in fact, the error function in the whole subset is zero, and this indicates a complete match. The best solution is obtained when the coefficient C(R) in Equation (2) is minimized. In other words, interpolation functions are slightly different before and after loading anywhere. According to Equation (5), to minimize C, its gradient must be zero.

$$\nabla C = \left(\frac{\partial C}{\partial R_k}\right)_{k=1,13} \tag{5}$$

The Newton-Raphson method is used to solve the Equation (5) and obtain its roots. This method uses an approximate initial value to find the root of the equations and repeats until the error is less than a certain value. Since the correlation coefficient is a function of the displacement components and their gradients, these unknowns can be obtained by searching for a category of these components that minimize the correlation coefficient. In the correlation method algorithm, the search process for calculating the unknown displacements and displacement gradients is started with long steps. In this process, the displacement gradients are initially considered zero, and the algorithm searches for the 1-pixel steps in interest area and the pixel that minimizes the correlation coefficient is considered as the initial solution. Then, using the Newton-Raphson method, their displacements and gradients are accurately obtained with a fraction of pixel size. The results of this step are used as initial values in the Newton-Raphson algorithm for the next subset [17]. In this method, by performing general calculations, finally, the strains in different directions are calculated as Equations (6)-(8):

$$\varepsilon_{xx} = \frac{1}{2}((\frac{du}{dx})^2 + (\frac{dv}{dx})^2 + (\frac{dw}{dx})^2) + (\frac{du}{dx}) \tag{6}$$

$$\varepsilon_{yy} = \frac{1}{2}((\frac{du}{dy})^2 + (\frac{dv}{dy})^2) + (\frac{dv}{dy}) \tag{7}$$

$$\varepsilon_{zz} = \frac{1}{2}((\frac{du}{dy}) + (\frac{dv}{dx})) + \frac{1}{2}(\frac{du}{dx}\frac{du}{dy} + \frac{dv}{dx}\frac{dv}{dy}) \tag{8}$$

According to Figure 16, in the central hole drilling method, first a rosette strain gauge is attached to the surface of a piece with residual stress. In the Rosette strain gauge, the optimal strain measurement points for the strain gauges have been observed. Then a small hole with a depth slightly larger than the diameter of the hole is created in the center of the rosette strain gauge. This hole locally releases the stresses in the environment around the hole and the released strains are measured by three strain gauges on the rosette.

In this regard, Schajer and Yang [18] defined nine calibration coefficients to relate the residual stress and

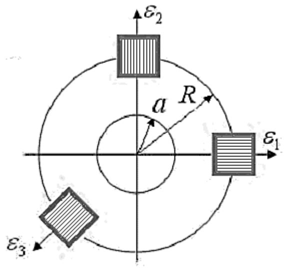


Figure 16. Strain gauge installation location in central drilling method

released strain, which can be obtained by theoretical, numerical and experimental methods. With an analytical solution, they calculated the calibration coefficient values for a suitable range of different mechanical properties of orthotropic materials. They used the following matrix relation:

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_{xy} \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \tag{9}$$

In Equation (9), the softness or calibration coefficients C_{11} to C_{33} depend on the elastic properties of the sample, the diameter and depth of the hole and the geometry of the strain gauge. In order to calculate these values, it is possible to use the analytical solution [19], referring to the standard and finite element simulation. In this research, reference is made to the standard in order to calculate the calibration coefficients.

Equations (10-18) can be used to measure the stress around the created hole. The strain values must be used to calculate the following constants as shown in Equations (10-12) [20].

$$p = \frac{\varepsilon_3 + \varepsilon_1}{2} \tag{10}$$

$$q = \frac{\varepsilon_3 - \varepsilon_1}{2} \tag{11}$$

$$t = \frac{\varepsilon_3 + \varepsilon_1 - 2\varepsilon_2}{2} \tag{12}$$

As stated, there are different methods for calculating calibration coefficients, and after calculating these coefficients, their values are incorporated in Equations (13-15) [20].

$$P = \frac{\sigma_y + \sigma_x}{2} = -\frac{Ep}{\bar{a}\left(1+\vartheta\right)} \tag{13}$$

$$Q = \frac{\sigma_y - \sigma_x}{2} = -\frac{Eq}{\bar{b}} \tag{14}$$

$$T = \tau_{xy} = -\frac{Et}{\bar{b}} \tag{15}$$

After calculating Equations (13-15), the values of plate stresses can be calculated using the Equations (16-18) [20].

$$\sigma_x = P - Q \tag{16}$$

$$\sigma_y = P + Q \tag{17}$$

$$\tau_{xy} = T \tag{18}$$

Equation (19) can also calculate the maximum and minimum stresses. The maximum tensile (or minimum compressive) principal stress, σ_{max} is at the angular position β in the clockwise direction relative to the strain gauge position 1 which is shown in the Figure 16. As the same way, minimum tensile (or maximum compressive) principal stress, σ_{min} is at the angular position β in the clockwise direction relative to the strain gauge 3 which is shown in the Figure 16. The angle β can be calculated by Equation (20) [20-22].

$$\sigma_{\max}, \sigma_{\min} = P \pm \sqrt{Q^2 + T^2} \tag{19}$$

$$\beta = \frac{1}{2} arc \tan\left(\frac{-T}{-Q}\right) \tag{20}$$

In Figure 17, the intended device for performing the drilling-DIC is presented.

As mentioned, in order to measure the amount of residual stress in the samples, the target piece is subjected to drilling operation in 10 stages and after each stage of drilling, the surface of the hole is imaged using IC Capture software. In Figure 18, 4 stages of imaging during the drilling operation are shown. After the drilling operation, all the recorded photos will be entered into the image processing software (GOM Correlate) and the

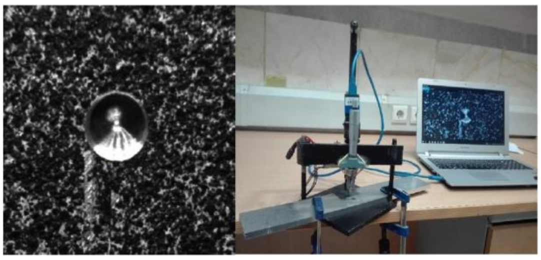


Figure 17. The device for residual stress measurement

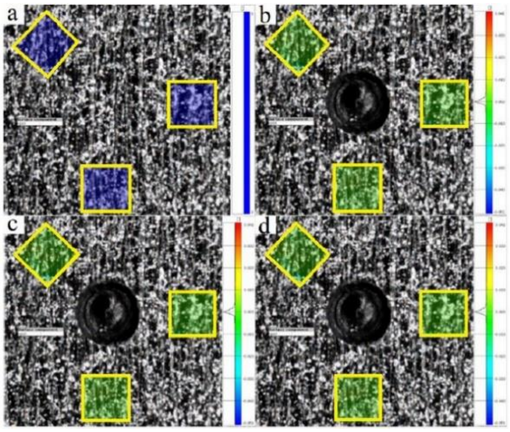


Figure 18. Stages of performing drilling operations step by step

desired mesh will be done in the software. An example of the image of the hole created on the sample and the mesh made on it are shown in Figure 19.

After meshing the sample and performing the required operations, at the end, the strain values released on the sample will be extracted numerically and contour. The contour of the strain released around the hole during the central drilling operation is shown in Figure 20. Finally, by checking the released strain values for all samples and using the required relationships, the residual stress values will be calculated. The values of the residual stress in the samples welded by the ultrasonic friction stir welding process are shown in Table 8.

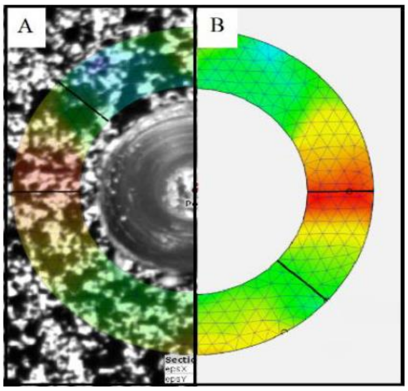


Figure 19. Meshing done on the sample image

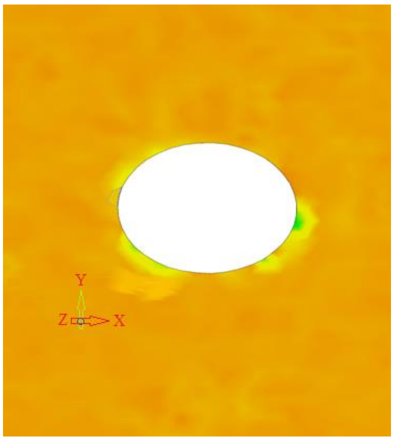


Figure 20. The contour of the strains released on the sample after the drilling operation

TABLE 8. Residual stress values for all ultrasonic friction stir welding samples

Sample number	Residual stress in aluminum zone (MPa)	Residual stress in steel zone (MPa)
1	121	323
2	113	302
3	106	293
4	144	336
5	131	329
6	120	312
7	148	353
8	141	350
9	129	339
10	160	366
11	152	350
12	141	337
13	166	380
14	164	371
15	164	366
16	169	384
17	144	379
18	138	377
19	175	395
20	164	389
21	148	377
22	181	406
23	166	398
24	152	388
25	188	425
26	159	402
27	153	373

The residual stress values depend on the thermal gradients created in the samples and their plastic deformation. According to the results, it has been determined that at a constant weld speed, the average values of the length residual stresses will increase as 30% with an increase in the rotational speed of the tool. The reason for this is the increase in the heating rate in the welding area. Also, at a constant rotational speed, the hardness in the welding area and the resistance to plastic deformation will increase with an increase in the weld speed. As a result, the average values of longitudinal residual stresses will increase. ANOVA analysis for selected factorial is shown in Table 9.

The Model F-value of 31.10 implies the model is significant. There is only a 0.01% chance that an F-value this large could occur due to noise. P-values less than 0.05 indicate model terms are significant. In this case A, B, C, BC are significant model terms. Relation between parameters and their effects on residual stress at different frequencies is shown in Figures 21-23.

At the end of the work, in order to evaluate the accuracy of the results related to the residual stress, according to Figure 24, all the samples were subjected to

TABLE 9. ANOVA for selected factorial

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	10382.59	10	1038.26	31.10	< 0.0001	**significant**
A-Rotational speed	6616.96	2	3308.48	99.10	< 0.0001	
B-Welding speed	738.74	2	369.37	11.06	0.0010	
C-Frequency	2267.19	2	1133.59	33.96	< 0.0001	
BC	759.70	4	189.93	5.69	0.0048	
Residual	534.15	16	33.38			
Cor Total	10916.74	26				

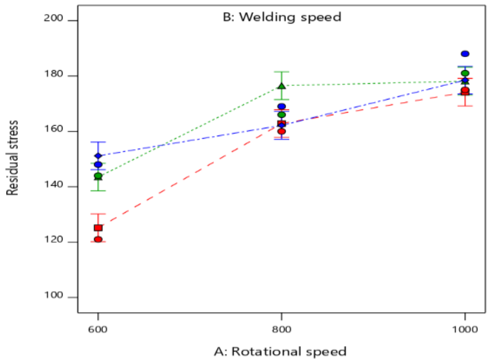


Figure 21. Relation between parameters at frequency 15KHz

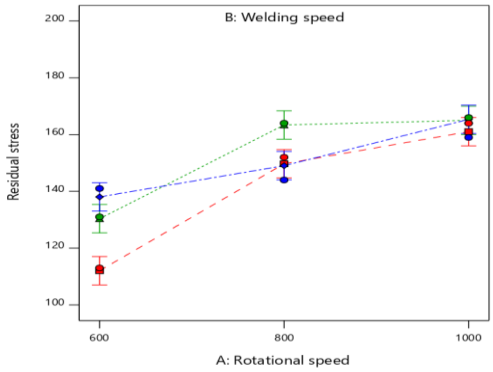


Figure 22. Relation between parameters at frequency 20KHz

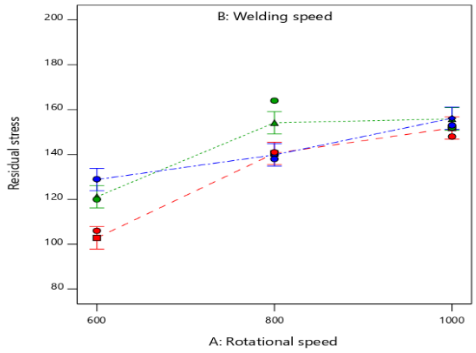


Figure 23. Relation between parameters at frequency 25KHz

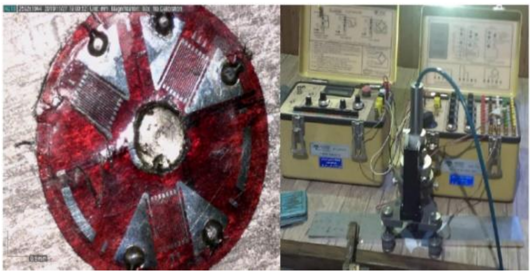


Figure 24. Central drilling test with strain gauge installation

the central drilling test by installing a strain gauge, and it was found that the error is less than 10% and obtained results were accurate and appropriate.

4. CONCLUSION

According to the experiments, the results obtained from this research can be stated as follows:

- According to the results, it has been determined that at a constant weld speed, the average values of the length residual stresses will increase as average of 30% with an increase in the rotational speed of the tool.
- According to the investigations, it has been determined that at a constant weld speed, the hardness value in the stir area has decreased as average of 25% with an increase in the rotational speed of the tool, and its value will decrease by moving towards the two base metals.
- The strength and percentage of increase in length of ultrasonic friction stir welding samples are higher than their values in the case of friction stir welding samples as 15%.
- By increasing the rotational speed and decreasing the linear speed of the tool during the ultrasonic friction stir welding process, the grain size in the welding area increased and the strength values and percentage of length increase of the welded samples decreased.

- Ultrasonic vibrations will improve mechanical properties as 15% to 25% such as strength and residual stress in welded samples and these properties are directly related to each other.

5. REFERENCES

1. Yang, Q., Mironov, S., Sato, Y. and Okamoto, K., "Material flow during friction stir spot welding", ***Materials Science and Engineering: A***, Vol. 527, No. 16-17, (2010), 4389-4398. https://doi.org/10.1016/j.msea.2010.03.082
2. Eslami, S., Farahani, B.V., Tavares, P.J. and Moreira, P., "Fatigue behaviour evaluation of dissimilar polymer joints: Friction stir welded, single and double-rivets", ***International Journal of Fatigue***, Vol. 113, (2018), 351-358. https://doi.org/10.1016/j.ijfatigue.2018.04.024
3. Susmel, L., Hattingh, D.G., James, M.N. and Tovo, R.,, "Multiaxial fatigue assessment of friction stir welded tubular joints of al 6082-t6", ***International Journal of Fatigue***, Vol. 101, (2017), 282-296. https://doi.org/10.1016/j.ijfatigue.2016.08.010
4. Habibnia, M., Shakeri, M., Nourouzi, S. and Givi, M.B., "Microstructural and mechanical properties of friction stir welded 5050 al alloy and 304 stainless steel plates", ***The International Journal of Advanced Manufacturing Technology***, Vol. 76, (2015), 819-829. https://doi.org/10.1016/j.ijfatigue.2016.08.010
5. Boonchouytan, W., Chatthong, J., Rawangwong, S. and Burapa, R., "Effect of heat treatment t6 on the friction stir welded ssm 6061 aluminum alloys", ***Energy Procedia***, Vol. 56, (2014), 172-180. https://doi.org/10.1016/j.egypro.2014.07.146
6. Liu, H.-J., Zhang, H.-j., Huang, Y.-x. and Lei, Y., "Mechanical properties of underwater friction stir welded 2219 aluminum alloy", ***Transactions of Nonferrous Metals Society of China***, Vol. 20, No. 8, (2010), 1387-1391. https://doi.org/10.1016/S1003-6326(09)60309-5
7. El-Sayed, M., Shash, A. and Abd-Rabou, M., "Finite element modeling of aluminum alloy aa5083-o friction stir welding process", ***Journal of Materials Processing Technology***, Vol. 252, (2018), 13-24. https://doi.org/10.1016/j.jmatprotec.2017.09.008
8. Li, H., Gao, J. and Li, Q., "Fatigue of friction stir welded aluminum alloy joints: A review", ***Applied Sciences***, Vol. 8, No. 12, (2018), 2626. https://doi.org/10.1016/j.jmatprotec.2017.09.008
9. Thomä, M., Gester, A., Wagner, G., Straß, B., Wolter, B., Benfer, S., Gowda, D.K. and Fürbeth, W., "Application of the hybrid process ultrasound enhanced friction stir welding on dissimilar aluminum/dual-phase steel and aluminum/magnesium joints", ***Materialwissenschaft und Werkstofftechnik***, Vol. 50, No. 8, (2019), 893-912. https://doi.org/10.1002/mawe.201900028
10. Benfer, S., Straß, B., Wagner, G. and Fürbeth, W., "Manufacturing and corrosion properties of ultrasound supported friction stir welded al/mg-hybrid joints", ***Surface and Interface Analysis***, Vol. 48, No. 8, (2016), 843-852. https://doi.org/10.1002/sia.5871
11. Thomä, M., Wagner, G., Straß, B., Wolter, B., Benfer, S. and Fürbeth, W., "Ultrasound enhanced friction stir welding (use-fsw) of hybrid aluminum/steel joints", in Friction Stir Welding and Processing X, Springer. (2019), 23-32.
12. Hong, K., Wang, Y., Zhou, J., Zhou, C. and Wang, L., "Investigation on ultrasonic assisted friction stir welding of aluminum/steel dissimilar alloys", ***High Temperature Materials and Processes***, Vol. 40, No. 1, (2021), 45-52. https://doi.org/10.1515/htmp-2021-0011
13. El-Morsy, A.-W., Ghanem, M.M. and Bahaitham, H., "Effect of friction stir welding parameters on the microstructure and mechanical properties of aa2024-t4 aluminum alloy", ***Engineering, Technology & Applied Science Research***, Vol. 8, No. 1, (2018). https://doi.org/10.48084/etasr.1704
14. Shukla, S., Komarasamy, M. and Mishra, R.S., "Grain size dependence of fatigue properties of friction stir processed ultrafine-grained al-5024 alloy", ***International Journal of Fatigue***, Vol. 109, (2018), 1-9. https://doi.org/10.1016/j.ijfatigue.2017.12.007
15. Peng, Y., Zhao, J., Chen, L.-s. and Dong, J., "Residual stress measurement combining blind-hole drilling and digital image correlation approach", ***Journal of Constructional Steel Research***, Vol. 176, (2021), 106346. https://doi.org/10.1016/j.jcsr.2020.106346
16. Sutton, M.A., Wolters, W., Peters, W., Ranson, W. and McNeill, S., "Determination of displacements using an improved digital correlation method", ***Image and Vision Computing***, Vol. 1, No. 3, (1983), 133-139. https://doi.org/10.1016/0262-8856(83)90064-1
17. Niezrecki, C., Baqersad, J. and Sabato, A., "Digital image correlation techniques for non-destructive evaluation and structural health monitoring", ***Handbook of Advanced non-Destructive Evaluation***, (2018), 46. https://doi.org/10.1007/978-3-319-26553-7_47
18. Schajer, G. and Yang, L., "Residual-stress measurement in orthotropic materials using the hole-drilling method", ***Experimental Mechanics***, Vol. 34, (1994), 324-333. https://doi.org/10.1007/BF02325147
19. Shokrieh, M.M. and Ghasemi K, A.R., "Determination of calibration factors of the hole drilling method for orthotropic composites using an exact solution", ***Journal of Composite Materials***, Vol. 41, No. 19, (2007), 2293-2311. https://doi.org/10.1177/0021998307075443
20. Standard, A., "E837-08 standard test method for determining residual stresses by the hole-drilling strain-gage method", ASMT International, West Conshohocken, PA, (2008). https://doi.org/10.1520/e0837-01
21. Khanjanzadeh, P., Amirabadi, H. and Sadri, J., "Design of broaching tool using finite element method for achieving the lowest residual tensile stress in machining of ti6al4v alloy", ***International Journal of Engineering, Transactions A: Basics***, Vol. 33, No. 4, (2020), 657-667. doi: 10.5829/IJE.2020.33.04A.17.
22. Khanjanzadeh, P., Amirabadi, H. and Sadri, J., "Experimental study on surface integrity of ti6al4v by broaching", ***International Journal of Engineering, Transactions B: Applications***, Vol. 35, No. 2, (2022), 481-492. doi: 10.5829/IJE.2022.35.02B.24.

Persian Abstract

چکیده

در این پژوهش به بررسی اتصال آلیاژ آلومینیوم ۵۰۵۲ به فولاد زنگ نزن آستنیتی ۳۰٤ پرداخته شده است. بدین منظور از فرآیند جوشکاری اصطکاکی اغتشاشی در دو حالت به همراه ارتعاشات فراصوتی و ساده استفاده شده است. به منظور دستیابی به بهترین کیفیت جوش از لحاظ خواص مکانیکی و متالورژیکی، پارامترهای جوشکاری نظیر سرعت دورانی، سرعت خطی و فرکانس مورد بررسی قرار گرفته است. هدف این تحقیق دستیابی به بهترین خواص مکانیکی و متالورژیکی با کمترین تنش پسماند در نمونه غیرهمجنس جوشکاری شده است. با هدف اندازه‌گیری مقادیر تنش پسماند ایجاد شده در نمونه‌ها پس از انجام عملیات جوشکاری از روش نوین سوراخکاری–برهمنگاری تصاویر دیجیتالی استفاده شده است. در انتها با بررسی نتایج مشخص شده است که ارتعاشات فراصوتی سبب بهبود خواص مکانیکی تا حدود ۱۵٪ و خواص متالورژیکی را هم تا حد زیادی بهبود داده است. به منظور بررسی صحت نتایج مربوط به تنش پسماند، تمامی نمونه ها با نصب کرنش سنج به روش سوراخکاری مرکزی تحت آزمایش قرار گرفتند و مشخص شد که خطای کمتر از ۱۰ درصد بوده و نتایج به دست آمده مناسب و قابل قبول بوده است.

International Journal of Engineering

Journal Homepage: www.ije.ir

A Behavioural Model for Accurate Investigation of Noisy Lorenz Chaotic Synchronization Systems

M. Nikpour*[a], M. Mobini[b], M. R. Zahabi[b]

[a] *Mazandaran Institute of Technology, Babol, Iran*
[b] *Faculty of Electrical Engineering, Babol University of Technology, Babol, Iran*

PAPER INFO

Paper history:
Received 06 May 2023
Received in revised form 30 May 2023
Accepted 31 May 2023

Keywords:
Behavioral Model
Chaotic Synchronization
Circuitry Noise Simulation

ABSTRACT

This paper presents a behavioral model for noisy Lorenz chaotic synchronization systems. This simple simulation-based model can be used for accurate noise voltage derivation of the chaotic oscillators and the investigation of chaotic synchronization systems. Moreover, the effects of circuitry noise on synchronization of Lorenz systems were analysed by using the proposed model. The performance of the synchronization system was numerically evaluated using ADS and MATLAB-SIMULINK environments. The measurement of Mean Squared Error (MSE) and Error to Noise Ratio (ENR) demonstrates that circuitry noise has a remarkable effect on the performance of chaotic Lorenz synchronization systems. For instance, the results showed that for low Signal to Noise Ratios (SNRs), i.e., $-40\ dB \leq SNR \leq 0dB$, the circuitry noise changed the ENR performance up to 1dB.

***doi**: 10.5829/ije.2023.36.08b.10*

1. INTRODUCTION

Chaotic signals can be generated by electrical circuits and nonlinear deterministic equations. It is shown that chaotic oscillators can be synchronized by linking them together with chaotic signals [1, 2]. Most of the researches on chaotic oscillators and chaotic synchronization systems are using an electrical circuit in a simulated environment [3]. In this paper we suggest a behavioral model for noisy Lorenz chaotic synchronization systems. This model allows us to study the effects of circuitry noise on the changes in the output of a Lorenz chaotic oscillator.

The chaotic waveforms can be used in a wide range of applications. They are major candidates for spread-spectrum schemes due to their wideband characteristics [4]. Moreover, numerous chaos-based modulations have been proposed for digital communications because of their robustness against noise and fading [5]. Many studies have been performed on Low Probability of Interception (LPI) features and secure communication schemes [6]. With these features, low-noise Lorenz chaotic synchronization schemes are promising for new classes of modulators. These signals meet the robustness against noise, convergence, and security requirements of the Ultra Reliable Low Latency Communications (URLLC) [7] and Industrial Internet of Things (IIoT) [8].

In this paper, internal noise is computed by using data sheets, and external noise is modelled as Additive White Gaussian Noise (AWGN) channels. In this study, we assumed that the received signal is corrupted by AWGN channels. In the context of chaotic synchronization, other destructive effects, such as multi-path fading, can be separately considered [9]. Considering AWGN channels has the following advantages:

1) Tractability: A small noise may lead to instability and synchronization error. This assumption makes the problem trackable by avoiding calculation complexity.

2) Necessity: Noise must be considered first, because it exists before any other effects of the communication channel. Other effects, such as fading channels, can be modeled in the next blocks after the noise block.

3) Generality: The performance evaluation of different communication schemes using the AWGN channel remains valid under realistic channel models, e.g., under fading channels.

*Corresponding Author Email: mhsnnikpour@yahoo.com (M. Nikpour)

Please cite this article as: M. Nikpour, M. Mobini, M.R. Zahabi, A Behavioural Model for Accurate Investigation of Noisy Lorenz Chaotic Synchronization Systems, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1502-1508

1. 1. Related Works

Most existing literatures have ignored the influence of circuitry noise. Furthermore, in a few papers, a desired internal noise is simply added to the signals that is not an accurate method. For example, Moskalenko et al. [9] discussed the general effect of noise in master-slave Colpitts oscillators. Their results showed that different values for AWGN channel can be considered and added to the system in order to determine the effects of noise on the system's performance. However, they did not determine the realistic value of noise voltage generated by Colpitts oscillators.

Sangiorgio et al. [10] have extended the analysis from a deterministic to a noisy environment, by considering both observation and structural noise. They have used recurrent neural networks for forecasting complex oscillatory time series on a multi-step horizon. Researchers in the field investigated different machine learning techniques and training approaches on dynamic systems with different degrees of complexity.

Moon et al. [11] investigated on synchronization in a set of high-dimensional generalizations of the Lorenz system obtained from the inclusion of additional Fourier modes. Numerical evidence supports that these systems exhibit self-synchronization. An example application of this phenomenon to image encryption is also provided.

Taheri et al. [12] have proposed a dynamic-free sliding mode control method to synchronize a class of unknown fractional order Laser chaotic systems. The efficacy of the proposed method is demonstrated by applying the method to a chaotic system and its practical applicability is demonstrated by using it to encrypt/decrypt color pictures.

A small noise may be effective on the stability and synchronization time of the chaotic synchronization systems, especially in weakly coupled oscillators [13]. Therefore, neglecting circuitry noise may result in some inaccurate calculations and designs. Behavioural modelling can be used for complexity reduction in circuits by modelling the effects of electrical components at the system level. In this way, electrical components replaced by some simple blocks. Considering noise causes a challenge in behavioural modelling. Order reduction methods are used for linear dynamics and some well-known methods such as sampled data models are presented for nonlinear circuits [14, 15].

1. 2. Innovative Aspects and Outlines

Innovative aspects are outlined below:

- In this paper, a simple simulation-based behavioural model is proposed for evaluation of the noisy Lorenz chaotic synchronization systems. This model can be useful for the exact noise voltage calculation of chaos-based circuits and systems. Since the generated noise is obtained from experimental measurements and published datasheets for electrical elements, this method can provide exact values of the noise, without need for physical realization of the chaotic circuits.
- The proposed model enables us to study the effects of circuitry noise on the performance of chaotic synchronization systems. Our results showed that the circuitry noise results in a remarkable error that may be vital in some applications, such as biomedical applications.

The rest of this paper is organized as follows: In section 2, we represented our behavioural model for the noisy Lorenz chaotic synchronisation system. In this section, a brief review on the role of noise in chaotic oscillators is first presented. Then, the circuitry noise measurement process is described. Finally, a filter- based method is described for noise generation in this section. Moreover, the influence of the circuitry noise is theoretically investigated in section 2. In section 3, simulation results are presented. Section 4 deals with concluding remarks.

2. BEVAVIORAL MODEL

The proposed model is shown in Figure 1. This model allows us to study the effects of circuitry noise on the performance of the chaotic synchronization systems. The block diagram of the proposed method contains three main steps.

1) The noise voltage spectrum is calculated using the ADS program. The noise of multipliers is neglected and the noise of other components is modelled by series voltage sources and a parallel current source according to the manufacturer's data-sheet on the online published user-manuals[1].

2) In the second step, the extracted noise voltage values are fed into the output of the oscillators. Thus, based on the obtained spectrum, circuitry noise can be generated using a well-known filter-based method.

3) In the final step, the generated circuitry noise is added to the chaotic synchronization system to evaluate the effect of circuitry noise on the performance of the

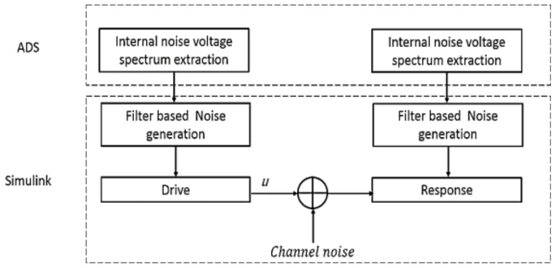

ADS
Internal noise voltage spectrum extraction
Internal noise voltage spectrum extraction
Filter based Noise generation
Filter based Noise generation
Simulink
Drive
u
Response
Channel noise

Figure 1. Block diagram of the proposed model

[1] www.ti.com/product/TL084A/datasheet, www.futurlec.com/Datasheet/Resistor

chaotic synchronization system. MATLAB SIMULINK environment provides a system level simulation for the Lorenz chaotic synchronization system.

In the next section, operation of each block is investigated and a brief review on the noise effects is presented. Finally, the effect of circuitry noise in the Lorenz chaotic synchronization systems is theoretically investigated.

2. 1. The Role of Noise in Conventional and Chaotic Oscillators In the conventional oscillators, noise can change both the amplitude and phase of a signal generated by an oscillator. The relation between phases is determined. In practical noisy oscillators the output voltage is given by [16]:

$$V_{LO\ output} = A(t)\, f[\omega_c t + \varphi(t)] \tag{1}$$

In Equation (1), A (t) and φ (t) are the amplitude and phase of the output, respectively. The frequency spectrum of an ideal and a practical oscillator is shown in Figure 2.

Unlike conventional oscillators that the phase rotation is approximately uniform, the phase variations of the chaotic oscillators have a random walk-like behaviour and instantaneous frequency depends on the amplitude. Consider the phase relations in a chaotic generator that called Poincare map [17]:

$$A_{n+1} = T(A_n),\ \frac{d\emptyset}{dt} = \omega(A_n) \equiv \omega_0 + F(A_n) \tag{2}$$

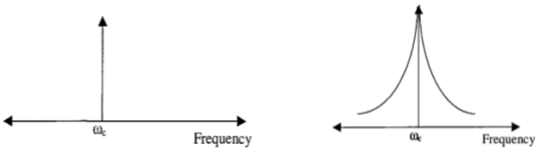
where A is the amplitude of nth state which is a discrete variable, T is the Poincare map, and ω_0 is the average frequency. The Sum of the average frequency and effective noise *F(A)* is equal to $\omega(A_n)$. As shown in Equation (2), the phase behavior in chaotic oscillators is similar to the conventional oscillators in the presence of noise. It is shown that the theoretically derivation of *F(A)* is complex. The above-mentioned characteristic of chaotic signals indicates that simulation-based methods are simpler for measurement of circuitry noise. In this paper, we calculated circuitry noise spectrum for Cumo-Oppenheim circuit using the phase noise analysis tool in an ADS simulation environment.

2. 2. Circuit Implimentation and Noise Voltage Spectrum Extraction This study is based on the circuitry noise of the Lorenz oscillators. We generate a noise for all components from manufacturer's data sheets and import them to ADS environment. Consider the Lorenz oscillator with the following differential function [18]:

$$\begin{aligned} \dot{u} &= \sigma(v - u) \\ \dot{v} &= ru - v - uw \\ \dot{w} &= uv - bw \end{aligned} \tag{3}$$

where u, v, w are the generated signals at the transmitter and σ, r, b are the bifurcation parameters. As shown in Figure 3. We used a drive-response structure to obtain a synchronized scheme.

The implementation of oscillator is shown in Figure 4. We used the Cumo-Oppenheim circuit, presented by Cuomo et al. [15]. Drive system sends a signal to the response circuit and both drive and response circuits are similar. We extract noise values of the used components from the published data sheets and input them into the simulated circuit. According to the manufacturer's data, the noise of multipliers is neglected and noise of other components is modelled by a series of voltage source and a parallel current source. The initial conditions are different in the drive and response and the parameters can change using variable resistors. The components are



(a) Ideal oscillator (b) Practical oscillator

Figure 2. Spectrum representation of an oscillator

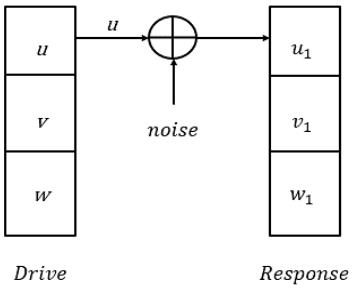


Figure 3. Drive-response synchronization system

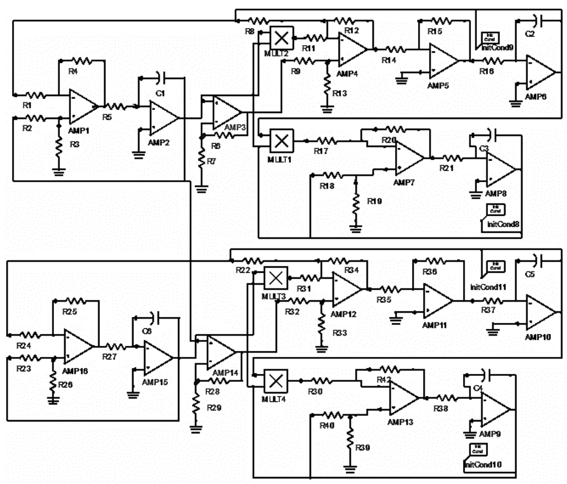


Figure 4. ADS implementation of Lorenz synchronization system

resistors, capacitors, op-amps (LF353) and multipliers (AD632AD).

2. 3. Filter-based Method For Noise Generation in SIMULINK Environment

In this step, the generated noise voltage by ADS tool is input to the chaotic oscillator for completion of the model in MATLAB SIMULINK environment. Thus, we should reconstruct the extracted noise spectrum in MATLAB SIMULINK environment. As mentioned above, the distribution of the noise voltage spectrum of chaotic oscillators is similar to the phase noise behaviour in conventional oscillators. Thus, for system level simulation, we can reconstruct the extracted noise spectrum using a filter-based method, presented by Godbole [19]. Filter-based model contains a white noise generator with power equal to the power of circuitry noise and a digital filter as shown in Figure 5. Finally, we reconstructed the noise can be added to the output of the oscillators in time domain.

The frequency response of the filter for frequency offsets $F(A_n) > 0$ can be calculated as follows:

$$H\big(F(A_n)\big) = 2 \times \sqrt{P(\omega_0 + F(A_n))}, \tag{4}$$

where P shows the power spectral density of the oscillator output. We design a Finite Impulse Response (FIR) filter by using the Filter Design and Analysis Tool (FDATOOL), and then import this filter into the SIMULINK model. Now, a system level analysis of the noisy oscillators can be performed with real circuitry noise values as shown in Figure 1. At the response side, we write the drive signal as follows:

$$\acute{u}(t) = u(t) + z(t). \tag{5}$$

where *z(t)* is total noise. Consider the total noise at the receiver side as the sum of the channel noise and circuitry noise. We assumed that *z(t)* is a zero mean Gaussian random variable with power spectral density σ_z^2. Both the parameters and initial values are assumed to be unknown in response side. We can write:

$$\begin{aligned} &u_1(0) = u(0) + e_u \\ &\sigma_1 = \sigma + \sigma\, e_\sigma \\ &v_1(0) = v(0) + e_v \\ &r_1 = r + r\, e_r \\ &w_1(0) = w(0) + e_w \\ &b_1 = b + b\, e_b \end{aligned} \tag{6}$$

where, e_u, e_v, e_w, e_σ, e_r, e_b are representation of the added errors to the parameters and initial values. The parameter errors and initial value errors are considered as random numbers distributed uniformly on [0.5, -0.5] and [1, -1], respectively. A similar assumption has been made by Kaddoum and Prunaret [20]. As described by Pecora, et al. [21], Li, et al. [22], Duong et al. [23], Zhou et al. [24], Chernoyarov et al. [25], for using the drive-response synchronization technique we can add a damping term to the response system. The damping term is shown by $\alpha(u' - u_1)$, where α is the strength of coupling.

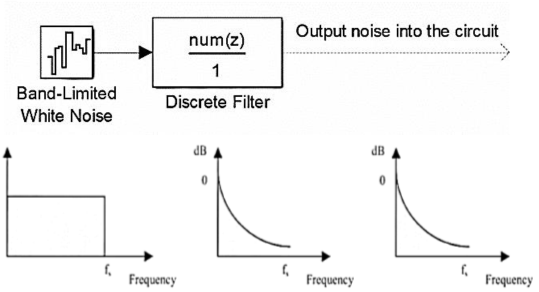


Figure 5. System level simulation of band-limited noise using filter-based method. Lower row shows spectral conditions

$$\begin{cases} \frac{du_1}{dt} = \sigma_1(v_1 - u_1 + \alpha(u' - u_1)) \\ \frac{dv_1}{dt} = -u_1 w_1 + r_1 u_1 - v_1 \\ \frac{dw_1}{dt} = u_1 v_1 - b_1 w_1 \end{cases} \tag{7}$$

A drive-response system can be considered as a communication system. On the receiver side, we have SNR= σ_u^2 / σ_z^2, where σ_u^2 is the power of drive signal at the transmitter side, and the SNR shows the signal to noise ratio. As described by Pecora, et al. [21], we can use mean square error (MSE) of synchronization as a performance metric. If the circuitry noise is modelled as mentioned above, we can write MSE values for different SNRs. Furthermore, the error to noise ratio (ENR) can be employed for synchronization performance evaluation.

$$ENR_{dB} = 10 \log_{10}\left(\frac{MSE}{\sigma_z^2}\right) \tag{8}$$

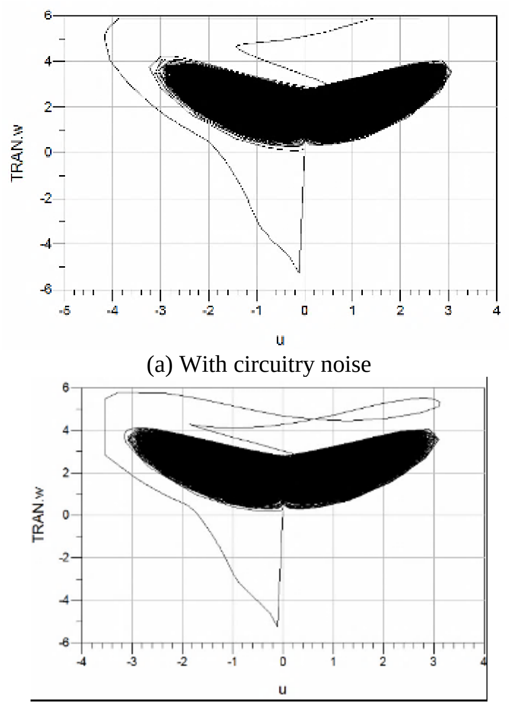
3. SIMULATIONS AND RESULTS

In this section, the simulation results of the proposed model are described.

3. 1. The Effects of Circuitry Noise on the Lorenz Oscillator

The noise of components are extracted from manufacturer's data-sheets and are imported to ADS environment. The noise of multipliers is neglected and other components is modelled by a series of voltage source and a parallel current source. We run a simulation of phase noise analysis tool in ADS. The results are shown in Figure 6, with consideration of circuit noise and without it. The bifurcation parameters and initial values of capacitors are equal for the two cases.

The difference between two trajectories indicate that noise changes the attractor. Synchronization occurs after



(a) With circuitry noise

(b) Without circuitry noise

Figure 6. Chaotic attractors projected onto uw-plane

a transient time that depends on the stability of synchronization [10]. In this case, the attractors collide with a chaotic saddle due to applied noise, and this noise create an internal crisis that generates a larger attractor. The key point is that the chaotic saddle contains an unstable steady state of the Lorenz oscillator [22]. In a weak coupling conditions a small noise can be effective on the stability, instability, and synchronization time of the system. In both cases, neglecting circuitry noise may result in some errors in calculations and applications [23].

The ADS program calculates resistors thermal noise plus the noise of other components, then plots the noise voltage spectrum. Figure 7 illustrates the noise spectrum of the output around zero value. Offset frequency is from 1 KHZ to 10 MHZ and the noise voltage is plotted both at the carrier frequency minus the offset frequency and the carrier frequency plus the offset frequency.

3. 2. The Effect of Circuitry Noise on the Synchronization System's Performance

The extracted noise voltage of the oscillator by the phase noise analysis tool of ADS, are used for investigate the circuitry noise effects, as shown in Figure 1. In other words, we have implemented a synchronization system to extract circuitry noise and used it for SIMULINK environment. The FIR filter was designed by MATLAB-FDATOOL and then, it was been imported into the SIMULINK model. It can be assumed that, at the receiver side, the parameters and initial conditions are unknown. This assumption results in that parameter error and initial value error distribute uniformly on [-0.5, 0.5] and [-1, 1], respectively.

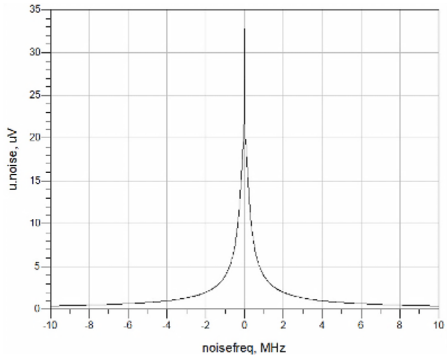
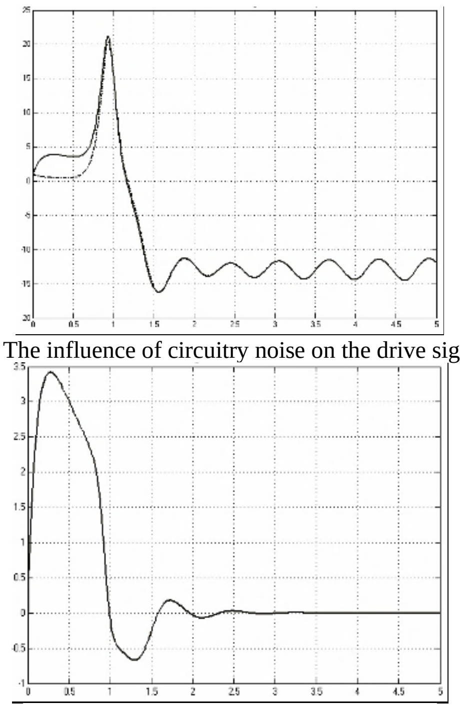


Figure 7. The noise voltage spectrum at the output of Lorenz oscillator

Here, external noise is neglected to deriving the influence of only circuitry noise. Figure 8(a) shows that the circuitry noise changes the receiver output compared to the free-noise case. In Figure 8(b), the error of the drive signal, i.e., the error that comes from neglecting circuitry noise is illustrated. In Figure 9, we can see the effect of circuit noise on u-u$_1$ trajectory plane. The circuitry noise can change the stability of the synchronization subsystems. Furthermore, the circuit noise can change the synchronization time.

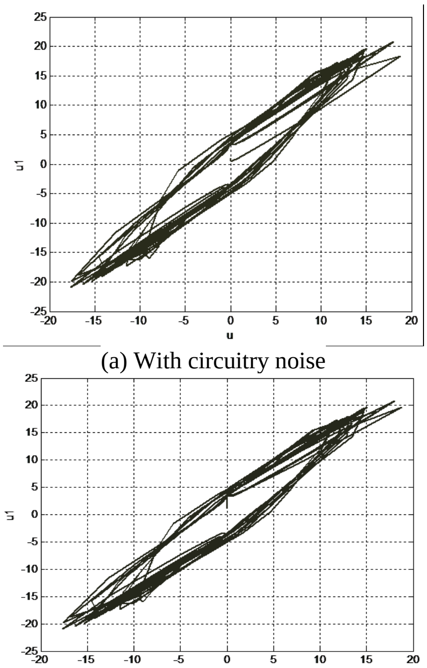


(a) The influence of circuitry noise on the drive signal

(b) Drive signal error due to circuitry noise

Figure 8. Drive signal at the receiver output

In the previous studies, the artificial noise has been used for simulations of pseudorandom generators [26]. Now, we consider our proposed model to produce practical values for simulation of the circuitry noise. As shown in Figure 10, we measured ENR for different SNRs. The effect of circuitry noise on synchronization error is visible, especially in low SNRs that it may reach to about 1 dB. This may seem a small value, but note that in weak coupling conditions the small noise can change the stability of the subsystems. As shown in Figure 10, in low SNR conditions, for example when the SNR= -40 dB, circuitry noise may be constructive and reduce the synchronization error. As a result, the calculation of MSE and ENR for system shows that with consideration of the behavioural model we can achieve exact accurate results.



(a) With circuitry noise

(b) Without circuitry noise

Figure 9. u-u1 plane of the drive signal at the transmitter

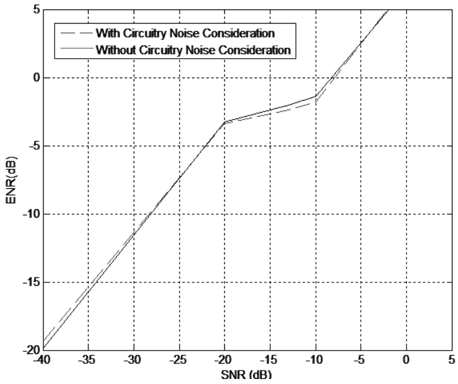


Figure 10. ENR Versus SNR for Lorenz Chaotic synchronization system

5. CONCLUSION

In this paper we present a simple model that evaluates the performance of the chaotic oscillators using a simulation-based method. This method can be used for noise voltage measurements in other chaotic circuits, including higher order differential equations or stimulus-assisted chaotic synchronization systems. Furthermore, the effect of circuitry noise on chaotic synchronization system analyzed using the proposed model. The measurement of the MSE and ENR demonstrates that circuitry noise has a remarkable effect on the performance of chaotic synchronization systems. For example, for Low-SNR conditions, i.e., $-40\,dB \leq SNR \leq 0dB,$ the circuitry noise can change the ENR performance up to 1dB. This level of error that incurred by the circuitry noise can have a huge destructive effect in some specific applications, such as rechargeable pace-makers, or may be negligible for other applications, such as communication goals. The results of this paper can be extended to a wide range of applications, from health monitoring, and chaos-based security, to human behaviour analysis and pattern studies in dynamic social networks. Therefore, the capability of the proposed model and the amount of noise effects can be explored for each of the above applications using the proposed method.

6. REFERENCES

1. Brahmbhatt, B. and Chandwani, H., "Modified second order generalized integrator-frequency locked loop grid synchronization for single phase grid tied system tuning and experimentation assessment", ***International Journal of Engineering, Transactions B: Applications***, Vol. 35, No. 2, (2022), 283-290. doi: 10.5829/ije.2022.35.02b.03.
2. Khosravi, A. and Gholipour, R., "Parameter estimation of loranz chaotic dynamic system using bees algorithm", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 26, No. 3, (2013), 257-262. doi: 10.5829/idosi.ije.2013.26.03c.05.
3. Dhanuskodi, S.N., Vijayakumar, A. and Kundu, S., "A chaotic ring oscillator based random number generator", in 2014 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST), IEEE. (2014), 160-165.
4. Cao, H., Zhang, R. and Yan, F., "Spread spectrum communication and its circuit implementation using fractional-order chaotic system via a single driving variable", ***Communications in Nonlinear Science and Numerical Simulation***, Vol. 18, No. 2, (2013), 341-350. doi: 10.1016/j.cnsns.2012.06.027.
5. Mobini, M., Kaddoum, G. and Herceg, M., "Design of a simo deep learning-based chaos shift keying (DLCSK) communication system", ***Sensors***, Vol. 22, No. 1, (2022), 333. doi: 10.3390/s22010333.
6. Aslinezhad, M., Sezavar, A. and Malekijavan, A., "A noise-aware deep learning model for automatic modulation recognition in radar signals", ***International Journal of Engineering, Transaction B: Applications***, Vol. 36, No. 8, (2023), 1459-1467. doi: 10.5829/IJE.2023.36.08B.06.
7. Mobini, M. and Kaddoum, G., "Deep chaos synchronization", ***IEEE Open Journal of the Communications Society***, Vol. 1, (2020), 1571-1582. https://doi.org/10.48550/arXiv.2104.08436

8. Khan, M.Z., Sarkar, A. and Noorwali, A., "Memristive hyperchaotic system-based complex-valued artificial neural synchronization for secured communication in industrial internet of things", ***Engineering Applications of Artificial Intelligence***, Vol. 123, (2023), 106357. http://doi.org/10.1016/j.engappai.2023.106357

9. Moskalenko, O.I., Hramov, A.E., Koronovskii, A.A. and Ovchinnikov, A.A., "Effect of noise on generalized synchronization of chaos: Theory and experiment", ***The European Physical Journal B***, Vol. 82, (2011), 69-82. doi: 10.1140/epjb/e2011-11019-1.

10. Sangiorgio, M., Dercole, F. and Guariso, G., "Forecasting of noisy chaotic systems with deep neural networks", ***Chaos, Solitons & Fractals***, Vol. 153, (2021), 111570. https://doi.org/10.1016/j.chaos.2021.111570

11. Moon, S., Baik, J.-J. and Seo, J.M., "Chaos synchronization in generalized lorenz systems and an application to image encryption", ***Communications in Nonlinear Science and Numerical Simulation***, Vol. 96, (2021), 105708. https://doi.org/10.1016/j.cnsns.2021.105708

12. Taheri, M., Chen, Y., Zhang, C., Berardehi, Z.R., Roohi, M. and Khooban, M.H., "A finite-time sliding mode control technique for synchronization chaotic fractional-order laser systems with application on encryption of color images", ***Optik***, Vol. 285, No., (2023), 170948. https://doi.org/10.1016/j.ijleo.2023.170948

13. Gharaibeh, A., "A behavioral model of a built-in current sensor for iddq testing", Texas A & M University, (2010),

14. Tamaddondar, M. and Noori, N., "Hybrid massive mimo channel model based on edge detection of interacting objects and cluster concept", ***International Journal of Engineering***, Vol. 35, No. 2, (2022), 471-480. doi: 10.5829/IJE.2022.35.02B.23.

15. Cuomo, K.M., Oppenheim, A.V. and Strogatz, S.H., "Synchronization of lorenz-based chaotic circuits with applications to communications", ***IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing***, Vol. 40, No. 10, (1993), 626-633. doi: 10.1109/82.246163.

16. Khanzadi, M.R., Kuylenstierna, D., Panahi, A., Eriksson, T. and Zirath, H., "Calculation of the performance of communication systems from measured oscillator phase noise", ***IEEE Transactions on Circuits and Systems I: Regular Papers***, Vol. 61, No. 5, (2014), 1553-1565. doi: 10.1109/TCSI.2013.2285698.

17. Schwabedal, J.T., Pikovsky, A., Kralemann, B. and Rosenblum, M., "Optimal phase description of chaotic oscillators", ***Physical Review E***, Vol. 85, No. 2, (2012), 026216. doi: 10.1103/PhysRevE.85.026216.

18. Jin, L., Zhang, Y. and Li, L., "One-to-many chaotic synchronization with application in wireless sensor network", ***IEEE Communications Letters***, Vol. 17, No. 9, (2013), 1782-1785. doi: 10.1109/LCOMM.2013.081313.131169.

19. Godbole, D., "Phase noise models for single-ended ring oscillators", (2000).

20. Kaddoum, G., Fournier-Prunaret, D., Chargé, P. and Roviras, D., "Chaos synchronization between lorenz systems using two numerical integration methods", in Proc. Conf. Nonlinear Dyn. Electron. Syst. (2014), 91-94.

21. Pecora, L.M., Carroll, T.L., Johnson, G.A., Mar, D.J. and Heagy, J.F., "Fundamentals of synchronization in chaotic systems, concepts, and applications", ***Chaos: An Interdisciplinary Journal of Nonlinear Science***, Vol. 7, No. 4, (1997), 520-543. https://doi.org/10.1103/PhysRevLett.64.821

22. Li, G., Tan, N. and Li, X., "Weak signal detection method based on the coupled lorenz system and its application in rolling bearing fault diagnosis", ***Applied Sciences***, Vol. 10, No. 12, (2020), 4086. doi: 10.3390/app10124086.

23. Duong, B.P., Kim, J.Y., Jeong, I., Im, K., Kim, C.H. and Kim, J.M., "A deep-learning-based bearing fault diagnosis using defect signature wavelet image visualization", ***Applied Sciences***, Vol. 10, No. 24, (2020), 8800. https://doi.org/10.3390/app10248800.

24. Zhou, C., Kurths, J., Kiss, I.Z. and Hudson, J.L., "Noise-enhanced phase synchronization of chaotic oscillators", ***Physical review Letters***, Vol. 89, No. 1, (2002), 014101. https://doi.org/10.1103/physrevlett.89.014101.

25. Chernoyarov, O., Glushkov, A., Lintvinenko, V., Matveev, B. and Faulgaber, A., "Digital root-mean-square signal meter", ***International Journal of Engineering, Transactions B: Applications***, Vol. 33, No. 11, (2020), 2201-2208. doi: 10.5829/IJE.2020.33.11B.11.

Persian Abstract

چکیده

این مقاله، یک مدل رفتاری برای سیستم‌های سنکرونیزاسیون آشوبی لورنز که به نویز آلوده شده ارائه می‌کند. این مدل ساده مبتنی بر شبیه‌سازی را می‌توان برای استخراج دقیق ولتاژ نویز و بررسی بیشتر سنکرونیزاسیون آشوبی دیگر استفاده کرد. علاوه بر این، اثرات نویز مدار بر سنکرونیزاسیون آشوبی لورنز با استفاده از مدل پیشنهادی تجزیه و تحلیل شده است. عملکرد سیستم سنکرونیزاسیون با استفاده از محیط‌های ADS و MATLAB-SIMULINK بررسی شده است. بررسی معیارهای اندازه‌گیری میانگین مربعات خطا (MSE) و نسبت خطا به نویز (ENR) نشان می‌دهد که نویز مدار تأثیر قابل‌توجهی بر عملکرد سیستم سنکرونیزاسیون آشوبی لورنز دارد. به عنوان مثال، نتایج نشان می دهد که برای نسبت سیگنال به نویز(SNR) پایین، یعنی در بازه صفر تا −٤٠ dB ، نویز مدار می تواند عملکرد ENR را تا ١ dB تغییر دهد.

International Journal of Engineering

Journal Homepage: www.ije.ir

Calibrating and Validation Microscopic Traffic Simulation Models VISSIM for Enhanced Highway Capacity Planning

S. M. Hafram*[a], S. Valery[b], A. H. Hasim[b]

[a] *Department of Civil Engineering, Universitas Muslim Indonesia, Makassar, Indonesia*
[b] *Department of Civil and Planning Engineering Education, Universitas Negeri Makassar, Makassar, Indonesia*

PAPER INFO

Paper history:
Received 29 March 2023
Received in revised form 19 May 2023
Accepted 21 May 2023

Keywords:
Traffic Flow
Driver Behavior
Model Calibration
Model Validation
Simulation Analysis

ABSTRACT

This research aims to calibrate and validate the VISSIM simulation model tool by comparing field data with simulation data. The ultimate goal is to evaluate traffic performance by comparing simulation results with direct observations in the field. This study uses modeling to determine a road segment's maximum flow volume. This study was conducted in Makassar, South Sulawesi, Indonesia, on Jalan Veteran Selatan. The method uses two main inputs: urban road primary capacity data from the Indonesian Highway Capacity Manual (IHCM 1997) and roadside activity data from PTV VISSIM. The GEH and MAPE have commonly used metrics for measuring the accuracy of simulation models and calibration measurements using driving behavior parameters. The research results obtained for validation measurements have met the requirements. Namely, the obtained MEPE value (7.38%) is 10% smaller than the obtained GEH value (2.032 and 3.961), which is still more than 5.00. The calibration measurements obtained the suitability of the vehicle location and intervehicle spacing in the simulation model (VISSIM) with the actual field conditions. The results obtained from using VISSIM can be reliable and helpful in designing and optimizing urban transportation systems in the future. It is essential to remember that traffic simulation with VISSIM is only a transportation decision-making and planning tool and must be combined with field observations and accurate data for adequate and efficient transportation solutions.

doi: *10.5829/ije.2023.36.08b.11*

1. INTRODUCTION

Urban development and transportation planning are closely intertwined, and transport planning is crucial to create sustainable, efficient, and livable cities [1, 2]. It involves evaluating the current transportation system, including road networks and public transportation, and developing new systems that meet the needs of urban residents. The ultimate goal of transport planning is to ensure the smooth flow of people and goods while reducing congestion, which can have many benefits, such as more efficient use of resources and less air pollution [3-5]. One of the biggest challenges in transport planning is the increasing traffic volume in cities worldwide. This leads to problems such as traffic congestion, longer travel times, and increased air pollution [6, 7]. While many efforts have been made to address this issue, such as improving road infrastructure and public transportation, these solutions are often insufficient to reduce traffic congestion effectively. Therefore, innovative and sustainable solutions are needed to tackle this challenge, including using intelligent transportation systems, encouraging alternative modes of transportation, and implementing policies that promote sustainable urban development [8, 9].

High congestion and traffic density levels often cause delays, accidents, and air pollution. Therefore, it is necessary to have the right strategy in traffic management to reduce the negative impacts. One of the practical tools in traffic management is the Microscopic Traffic Simulation Model. Simulation analysis heavily relies on software as the primary tool for facilitating the

*Corresponding Author Institutional Email: stmaryam@umi.ac.id (S. M. Hafram)

Please cite this article as: S. M. Hafram, S. Valery, A. H. Hasim, Calibrating and Validation Microscopic Traffic Simulation Models VISSIM for Enhanced Highway Capacity Planning, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1509-1519

calculation process [10]. The features of four different simulation programs: AIMSUN [11], TransModeler [12], CORSIM [13], and VISSIM were analyzed by Salgado et al. [14] and Hadi et al. [15]. Although each software package has advantages, the study ultimately chose VISSIM due to its superior vehicle routing capabilities, total output, stability, and extensive supporting documents accompanied by animations [16-19]. Traffic flow simulation can be conducted at macro and micro levels. However, Habtemichael and de Picado Santos [20] focused on transportation management and found that simulation at the micro level yields more satisfactory results compared to macro simulations. At the micro level, the simulation can better capture the impact of heterogeneous traffic and produce more comprehensive and precise results. This level of detail is crucial for evaluating traffic flow scenarios, predicting traffic patterns, and making informed traffic management and planning decisions.

Using microscopic traffic simulation models such as VISSIM has revolutionized transportation planning by providing planners with a powerful tool to evaluate various scenarios and predict the impact of infrastructure changes on traffic flow. These models use advanced algorithms to simulate the behavior of individual vehicles, considering factors such as driver behavior, traffic signals, and lane changes [21]. By analyzing the simulation results, transportation planners can identify potential issues and test different solutions before making any changes to the transportation infrastructure [22].

The level of detail provided by these models allows for a comprehensive evaluation of traffic flow in urban areas. Transportation planners can use these models to optimize the timing of traffic signals, adjust road layouts, and improve public transportation systems to reduce congestion and improve accessibility. Using microscopic traffic simulation models, transportation planners can make more informed decisions, leading to a more efficient flow of people and goods, improved safety, and reduced environmental impact [23]. VISSIM, in particular, has become a widely used and well-regarded microscopic traffic simulation software program due to its ability to predict traffic flow and congestion accurately. The software includes various customizable parameters, including vehicle types, traffic signals, and lane changes, allowing for detailed traffic flow analysis at the individual vehicle level [24]. The program also allows the simulation of various scenarios, such as changes in traffic patterns, lane configurations, or signal timings, to estimate the effect of different infrastructure changes on travel movement.

VISSIM and other traffic simulation models' accuracy depends on the calibration and validation process. This process involves adjusting the model's parameters to match real-life traffic flow data and validating the calibrated model against independent traffic data to verify the model's accuracy [25]. Calibration and validation ensure that the model accurately represents actual traffic conditions, accounting for changes in traffic volume, time of day, and weather conditions. Regularly updating and maintaining calibration and validation procedures is crucial to ensure the accuracy and reliability of simulation models, as traffic conditions can change ith respect to time [26-28]. Both calibration and validation must be periodically revised to ensure that the simulation model can still accurately replicate field conditions and produce consistent results with new observation data. This ongoing process is vital in ensuring the relevance and reliability of simulation results [29].

Calibrating a microstimulator involves two sets of parameters: driving behavior parameters and travel behavior parameters. Some examples of the former are models of acceleration, lane switching, and intersections; examples of the latter are models of origin-destination flows and route selection. However, scant information is available on calibrating traffic simulation models, with most studies focusing on one aspect typically driving behavior and assuming that the rest of the limits are already known. For example, studies conducted by Zhe et al. [30], Jha et al. [31], Daigle et al. [32], and Ratrout et al. [33] only calibrate driving behavior parameters. Route selection is a crucial element in the calibration procedure. It is commonly assumed that the flows between origin and destination have been pre-established. The estimation procedures for origin-destination flows function on the premise that the assignment matrices, which represent the impact of route selection and flow propagation, have been established or are already known. The assignment matrices are paramount in comprehending and simulating the dispersion of traffic volumes among diverse paths within a transportation system. Assuming the availability of assignment matrices, the calibration procedure can prioritize the adjustment of other parameters and variables to enhance the precision and dependability of the comprehensive model [11]. Yang and Slavin [12] took a different tack by extending the origin-destination estimation process to incorporate a route choice model, but they did so assuming that the model parameters are immutable.

The model's parameters are fine-tuned during calibration by comparing the simulated and observed traffic flows. This requires making small, incremental changes to the parameters to get simulation results as close as possible to the actual data. The calibration process is not complete until validation has been performed, as this verifies the accuracy of the model and its applicability for foreseeing the results of any future changes to the infrastructure. Predictions from the calibrated model checked against data on traffic flows that were not used during calibration. The calibration and validation process is essential to the success of traffic simulation models like VISSIM [24, 34]. Adjusting the model's parameters to correspond with observed traffic volumes is known as calibration, and checking the model's accuracy by comparing predictions to external traffic measurements is known as validation. Calibration and validation check the accuracy of traffic simulation models so that transportation infrastructure decisions can be made confidently [35].

Based on the description, the research aims to calibrate and validate using the VISSIM simulation model tool by comparing field data with simulation data. The ultimate goal of the research is to evaluate traffic performance by comparing simulation results with direct observations in the field. By evaluating these results, the research can provide recommendations to improve traffic performance in the future. VISSIM model's vehicle behavior in urban transportation systems to better understand traffic performance and predict infrastructure changes' effect on traffic movement. Therefore, by calibrating and validating, the results obtained from the VISSIM can be reliable and helpful in designing and optimizing urban transportation systems in the future use.

2. MATERIALS AND METHODS

2. 1. Research Approach

The research approach in this study involves a modeling method to define the determined movement volume a highway segment can handle. The method uses two main inputs, namely the primary capacity data from the Indonesian Highway Capacity Manual (IHCM 1997) [36] for urban roads and the number of roadside activities from the PTV VISSIM assistance program [37]. The study requires several data types to model, including road geometry, side barriers, and free-flow speed data. The side barrier data used in this study include roadside parking activities, vehicle activities entering and leaving the road segment, and slow vehicles. The study did not consider the influence of pedestrians in the modeling process.

This study employed a quantitative methodology based on the analysis and modeling of collected data. The study relies on existing data sources and software programs to perform the modeling process. The study results are presented in numerical values that indicate the maximum flow volume the road segment can handle.

2. 2. Location of Study

This study was conducted in front of the Maricaya Market, located on Jalan Veteran Selatan, Makassar sub-district, Makassar City. Maricaya Market is a traditional market located in the heart of a densely populated settlement that serves as a local trading center.

This location was chosen for research because Maricaya Market is an important land transportation area in Makassar City. Because of its strategic location, this market is a crossroads for many transportation routes, including highways, ring roads, and other major thoroughfares. This makes it a desirable location for observing and analyzing traffic patterns and interactions between vehicles and pedestrians in congested areas.

The location was chosen to analyze the impact of market activities on road volume and travel congestion. The study was conducted for one week in July-August 2022 and included observations on weekdays and holidays. On weekdays, observations were made from Monday to Friday, while on holidays, they were held on Saturday and Sunday.

Data collection was conducted in three sessions, each at different times of the day, to capture any changes in traffic conditions. Session I was held in the morning from 6.00 to 10.00 A.M. Session II in the afternoon from 12.00 to 2.00 P.M., and Session III in the evening from 4.00 to 6.00 P.M.

The study focused on peak hours, four hours in the morning and four hours in the afternoon, to capture the highest traffic volumes and travel congestion. The research used direct observation methods to collect data, including manual traffic counting, recording travel times, measuring vehicle speeds and measuring road geometry.

This study aims to collect traffic flow data consisting of four types of vehicles, namely light vehicles, heavy vehicles, motorbikes, and non-motorized vehicles, which are obtained directly from observations and measurements in the field. The road section has 2/4D divided lanes. Observations were made on this road section because it is a busy and vital area for land transportation in Makassar City.

2. 3. Data Geometric

Primary data was obtained directly from surveys of geometric road conditions. This data includes road width, number of lanes, lane width, road shoulder width, and road type. Where the observed

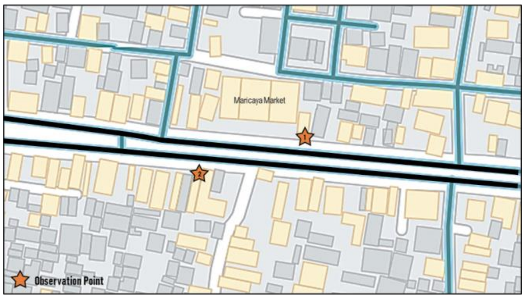


Figure 1. Test Site: Veteran Selatan Road, Makassar, South Sulawesi, Indonesia

location is at the point of the road, namely Jalan Veteran Selatan. The following is a description of the geometric conditions of the road (Table 1).

Data obtained from field observations will later be processed and analyzed to produce useful information on road capacity, traffic density, and congestion around Maricaya Market. The data from this research can assist decision-makers in traffic management in Makassar City, particularly in increasing road capacity and reducing traffic jams in busy and densely populated areas.

2. 4. Data Analysis

2. 4. 1. Traffic Volume The definition of traffic volume refers to the count of vehicles passing a particular point or line on a road cross-section. The method of traffic counting is done manually by recording vehicles in a flow that is distributed according to the type of vehicle continuously at 20-minute time intervals. The calculation of vehicle volume is determined using an equation:

$$Q = {}^{n}/_{t} \tag{1}$$

where, Q is volume of vehicles (vehicles/hour), n is the number of vehicles (vehicles) and t is observation time (hours).

2. 4. 2. Road Capacity Capacity refers to the maximum traffic volume sustained under specific conditions, including geometry, distribution of traffic directions and composition, and environmental factors, with units of PCU/hour [36]. Regarding explanations for road capacity, speed, volume, and density are related. The more vehicles on the road, the more the average speed decreases. The basic equation for determining capacity is as follows:

$$C = C_O \times FC_W \times FC_{SP} \times FC_{SF} \times FC_{CS} \tag{2}$$

Where, C is capacity (PCU/hour), Co is basic capacity for ideal conditions (PCU/hour), FCw is traffic lane width adjustment factor, FCsp is directional separation adjustment factor, FCsf is side resistance adjustment factor and FCcs is city size adjustment factor.

2. 4. 3. Degree of Saturation The degree of saturation is the traffic flow ratio (PCU/hour) to capacity (PCU/hour) and is used as a critical factor in assessing and determining the performance level of a road segment. If the Q/C Ratio exceeds 1, the traffic volume exceeds the available road capacity. This indicates the excess capacity and possible congestion. The higher the Q/C Ratio, the denser and more jammed the traffic conditions. The calculation of the degree of saturation is determined using an equation:

$$DS = {}^{Q}/_{C} \tag{3}$$

where, DS is degree of saturation, Q is traffic flow (PCU/hour) and C is capacity (PCU/hour).

TABLE 1. Road Geometric Characteristics

Road Characteristics	Observation (Existing)
Road Type	Four-lane Split or One-way Street
Type of Road Pavement	Asphalt
Road Lane Width	9 meters
Road Lane Width	3 meters
Road Shoulder Width	1 meter

2. 5. Calibration Model The purpose of calibrating driving behavior parameters is to ensure that the simulation model can accurately reproduce the field's driver behaviors. This is very important in transportation analysis and highway planning because an accurate simulation model can provide more accurate predictions about how changes in road conditions or traffic policies may affect driver behavior and traffic flow.

Calibration in VISSIM is a process of forming appropriate parameter values so that the model can replicate traffic to conditions that are as similar as possible. The method used is trial and error, which is done by comparing field observation conditions with conditions in the simulation. This simulation is accurate if the error rate between the simulation results and the observed data is relatively low. The calibration uses optimization techniques to minimize the deviation between the observed data and the simulation measurements made to match.

This calibration process is carried out by comparing empirical or field data with the simulation results of the developed mathematical model. In this case, the difference between the empirical data and the simulation results will be used to adjust the required parameter values in the model. The simulation model must first be calibrated using field data to produce accurate predictions. This can be done by collecting data from direct observations, such as measurements of speed, acceleration, head distance, and other variables related to driver and vehicle behavior on the road.

2. 6. Validation Model In VISSIM, the validation process involves comparing the results of simulations with observations to verify the accuracy of the calibration. The validation examines the traffic flow volume and the queue length. The GEH (Geoffrey E. Havers) test is a statistical method used to evaluate the accuracy of simulation models. It measures the difference between the observed and simulated values and compares it to the expected range of differences. In the following GEH [38-41], the formula has specific provisions for the resulting error values as follows:

$$GEH = \frac{\sqrt{(q\_simulated - q\_observed)^2}}{0.5 \times (q\_simulated + q\_observed)} \quad (4)$$

where q simulated is average traffic flow volume in simulation (vehicles/hour) and q_observation is traffic flow volume in the field (vehicles/hour).

The GEH test is a valuable tool for evaluating the accuracy of simulation models and can help ensure that the models are reliable and accurate for use in transportation planning and decision-making. To explain from the GEH results can be seen in Table 2.

A GEH value less than 5.00 is generally considered acceptable, indicating that the simulated values are accurate and can be used for further analysis and planning. However, a GEH value between 5.00 and 10.00 indicates a possible error or harmful data, and further investigation may be necessary. A GEH value greater than 10.00 indicates that the simulated values are significantly different from the observed values, and the model should not be used for further analysis or planning.

The Mean Absolute Percentage Error (MAPE) is a commonly used metric for measuring the accuracy of a forecast or prediction [42]. It is calculated by taking the absolute difference between the actual and predicted values, dividing that by the actual value, and multiplying by 100 to get a percentage [43]. The MAPE is then calculated as the average of these percentage errors.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{At-Ft}{At}\right| \times 100\% \quad (5)$$

where n is total data, At is the observation data and Ft is simulation model data.

MAPE is a valuable metric because it provides a simple way to evaluate the accuracy of a forecast or prediction, regardless of the scale of the data or the units of measurement. Based on Lewis [44], the range of MAPE values can be interpreted into four categories (Table 3).

TABLE 2. Description of GEH Result

GEH Range	Description
GEH < 5.00	Accepted
5.00 ≤ GEH ≤ 10.00	Caution: model error or insufficient data
GEH > 10.00	Denied

TABLE 3. Description of MAPE Result

MAPE Range	Description
≤ 10%	Simulation results are very accurate
10 – 20%	Good Simulation results
20 – 50%	Simulation results are feasible (good enough)
> 50%	Inaccurate simulation results

MAPE is a method of measuring the error or accuracy of a prediction or simulation model by comparing the difference between the actual value and the normalized predicted value in the form of a percentage.

3. RESULTS AND DISCUSSIONS

3. 1. Calibration Model

Driving Behavior must be adapted to conditions in the field so that the simulation results can represent conditions in the field. The parameter used for modeling validation with field conditions is the model traffic volume equal to the field traffic volume. If the results do not represent the conditions in the field, then a reset or calibration is required to suit the field. By calibrating the Driving Behavior parameters, the simulation model will be able to represent driver behavior and traffic volume following the conditions in the field so that the simulation results can be used to predict realistic traffic conditions. The Driving Behavior Parameters used in this study summarized in the following table:

The driving behavior Table shows several parameters with constant values in each simulation period. The interpretation of the calibration values for the parameters in Table 4 is as follows:

- Average Standstill Distance: The calibration value indicates that the vehicle has an average distance of 0.2 meter before stopping.
- Additional Part of Desired Safety Distance: Two calibration values are used, 0.5 and 1 meter. This indicates that the safe distance between the vehicle in front and behind is increased by a larger additional distance when travelling at higher speeds.
- Number of Observed Vehicles: In this simulation, the number of observed vehicles is 2.

TABLE 4. Calibration Model Validation

Parameter	Driving Behavior	
	Default	Changes
Average Standstill Distance	2 meters	0.2 meter
Add. Part of Desired Safety Distance	2 meters	0.5 meter
Add. Part of Desired Safety Distance	3 meters	1 meter
No. of Observed Vehicle	2.00	2.00
Lane Change Rule	Free Lane Selection	Free Lane Selection
Desired Lateral Position	1 meter	Any
Lateral Distance Driving	1 meter	0.15 meter
Lateral Distance Standing	1 meter	0.45 meter
Safety Distance Reduction Factor	0.6 meter	0.45 meter
Minimum Headway	0.5 second	0.5 second

- Lane Change Rule: The rule used in the simulation is free lane selection.
- Desired Lateral Position: The desired lateral position is any.
- Lateral Distance Driving: The lateral distance between one vehicle and another while driving is 0.15 meter.
- Lateral Distance Standing: The lateral distance between one vehicle and another while standing is 0.45 meter.
- Safety Distance Reduction Factor: The calibration value used for the safety distance reduction factor is 0.45 meter.
- Minimum Headway Time: The minimum headway time or the minimum time distance that must be maintained between vehicles is 0.5 second.

This shows that driver behavior can vary in traffic conditions, such as rush hour and off-peak. Therefore, to obtain accurate simulation results, it is necessary to calibrate these parameters based on traffic conditions according to the situation in the field (Figures 2 and 3).

The calibration Figures 2 and 3 show the difference in traffic flow behavior before and after calibration on the VISSIM software. The traffic in the simulation is observed to move steadily in a lane-by-lane manner with sufficient gaps between the vehicles before undergoing calibration. However, the traffic becomes more erratic after calibration, with frequent overtaking and closing gaps between vehicles.

This change indicates that the driving behavior in the VISSIM simulation model better represents real-world traffic conditions, where overtaking and chaos on the road are common occurrences. In a heterogeneous traffic context, where various vehicles with different speeds are on the same road, the calibration results show that the simulation model is acceptable and provides more accurate results. This way, the VISSIM simulation results can be used to plan and develop a more effective and efficient traffic system.

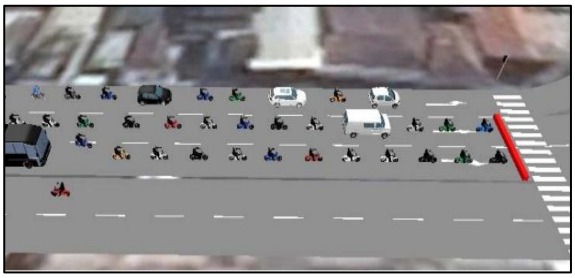


Figure 2. VISSIM Test: Before Calibration

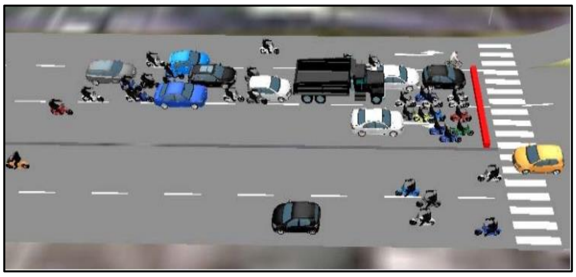


Figure 3. VISSIM Test: After Calibration

3. 2. Validation Model

Table 5 shows the validation results of the simulation models used in transportation analysis and planning. The table calculates two GEH values for two days, namely Monday and Saturday (peak hours).

Table 5 presents the validation results of the GEH test for vehicle volume per hour, comparing VISSIM and IHCM 1997. The study was conducted on two different days, namely Monday and Saturday.

The interpretation of the results shows that on Monday, there was a slight difference in vehicle volume between VISSIM and IHCM 1997, as indicated by the GEH (2.032). However, vehicle volume significantly differed on Saturday between the two models, as indicated by the higher GEH (3.961). Despite this, the GEH values for both days were below 5, indicating that the simulation model meets the desired accuracy criteria. Therefore, the simulation model is acceptable for more advanced transport planning and analysis.

The range of MAPE values (Table 6) obtained in the calibration results given is (7.38%) where these results are ≤10. This shows that the forecasting/simulation results are accurate and follow the actual field conditions. The smaller the MAPE value, the better the forecasting or simulation model's ability to predict the actual value. In this context, the MAPE values obtained indicate that the simulation model used in this study can predict actual values and is reliable for further analysis and transportation planning.

Apart from using performance evaluation metrics such as Mean Absolute Percentage Error (MAPE), validating the simulation results can also be done by comparing field conditions with simulation results. From Figures 4 and 5, it can be seen that the simulation

TABLE 5. GEH Test Validation Results (vehicle/hour)

Time	VISSIM	IHCM 1997	GEH
Monday	1707	1792	2.032
Saturday	1340	1489	3.961

TABLE 6. MAPE Test Validation Results (vehicle/hour)

Time	VISSIM	IHCM 1997	MAPE
Monday	1707	1792	2.37%
Saturday	1340	1489	5.00%
		Average	7.38%

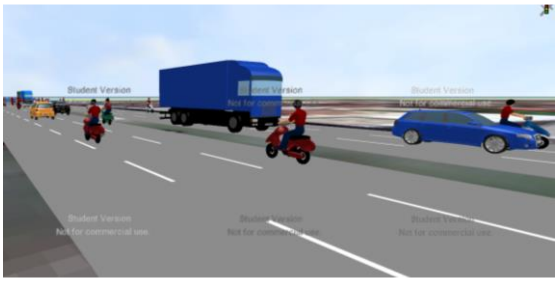


Figure 4. VISSIM Test Condition Site

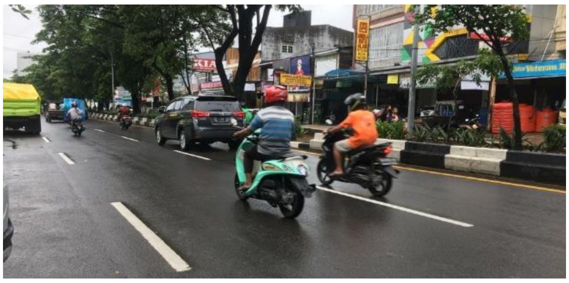


Figure 5. Existing Conditions Site

results are quite similar to the actual field conditions. This shows that the simulation model used is quite good and can represent traffic conditions in the field.

Simulation model validation is a process to check the reliability and accuracy of the model in predicting traffic behavior in the field. By doing good validation, the simulation model can be well-calibrated to be trusted in predicting traffic behavior in the field. In this case, the visualization images in Figures 4 and 5 show the suitability of the vehicle position and the distance between the vehicles in the simulation model with the actual field conditions. This proves that the simulation model has passed the validation process correctly.

By using a well-calibrated simulation model, traffic infrastructure development decisions can be taken more effectively and efficiently because the model can accurately predict traffic behavior in the field. Therefore, validation of the simulation model is essential to ensure the reliability and accuracy of the model so that decisions made based on the model can be more accurate and reduce the risk of errors in the development of traffic infrastructure.

3. 3. Comparison of Observation (IHCM 1997) and Simulation (VISSIM)

Traffic volume is one of the parameters used in validating using the Geoffrey E. Havers (GEH) formula. This aims to compare whether the simulation model is appropriate or describes the traffic conditions at the observation location. Due to the limitations of the VISSIM Software in displaying simulation results, namely for 600 seconds of simulation, the volume of vehicles compared is the volume of vehicles per hour.

Figure 6 compares simulated and observed traffic volumes on Monday and Saturday afternoons. The simulated traffic volume is calculated using the VISSIM software, while the observed traffic volume is measured directly in the field. The figure shows that the traffic volume on Monday afternoon was higher than Saturday afternoon for both simulation and observation. However, there is a difference between the simulated and observed values on the two days. On Monday afternoon, the simulation value was 1707 vehicles/hour, while the observed value was 1792 vehicles/hour. On Saturday afternoon, the simulation value was 1340 vehicles/hour, while the observed value was 1489 vehicles/hour.

This shows that even though the simulated and observed values have the same trend (i.e., the traffic volume is higher on Monday afternoon), there is a numerical difference between the two. Several factors, such as inaccuracies in observational measurements or the calibration of simulation models, can cause this difference. Therefore, it is necessary to adjust or calibrate the simulation model so that the results are more accurate and can better represent field conditions.

The VISSIM procedure utilizes predetermined parameters, such as the maximum vehicle speed, the distance between vehicles, and the red time of traffic lights. The VISSIM simulation results demonstrate the traffic service level on a road or intersection. This information can be used to evaluate the performance of existing traffic and identify areas requiring improvement or modification. The VISSIM simulation results can also be used to compare the performance of different tested scenarios. The optimal and most effective scenario for increasing traffic performance can be selected by comparing the performance of the two scenarios. The initial stage calibration and validation process must be carried out with care to obtain accurate and reliable simulation results. After that, the specified parameters can be used to run VISSIM to produce accurate and reliable service-level simulation results. The Level of Service measures road performance and traffic congestion. Average speed, travel time, number of vehicles per unit of time, road capacity, traffic density, and congestion level are used to score this system. Road service is as follows:

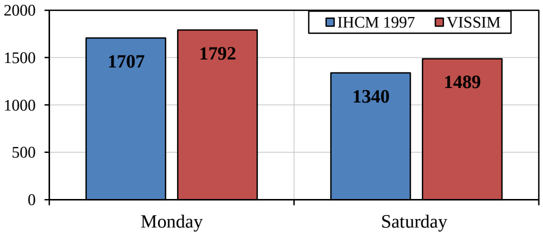


Figure 6. Comparison of Observation (IHCM 1997) and Simulation (VISSIM)

At the Level of Service, each level is denoted by a letter from A to F, with A being the best level and F being the worst level [36, 45]. At level B, traffic flow is stable with moderate vehicle volume and limited speed. The driver has sufficient freedom in choosing the speed of the vehicle. At level C, traffic flow remains stable, but the speed and movement of vehicles are controlled by traffic volume. The driver has limitations in choosing the speed of the vehicle. At level D, the traffic flow is nearly unstable, with high traffic volumes and speeds that can be tolerated but are highly influenced by flow conditions. The traffic flow is close to unstable, and almost all drivers have limited freedom in driving the vehicle.

Based on the simulation results using the VISSIM software, the level of service on Monday is D, while on Saturday, it is B (Table 7). However, there are differences in results when using the 1997 Indonesian Highway Capacity Manual (IHCM) method, which uses the degree of saturation value in categorizing service levels. Based on this calculation, the level of service on Monday afternoon is categorized as C, and on Saturday afternoon is categorized as B. This shows differences in the results of measuring the traffic service level depending on the methods and techniques used. Therefore, traffic and transportation experts need to choose the correct methods and techniques for analyzing and measuring the level of traffic services. In addition, the results of these measurements can be used to identify traffic problems and design effective solutions to improve road service levels and performance.

Several studies in Indonesia have also used VISSIM as a microscopic simulation application to evaluate the performance of a road segment. This study used VISSIM to model vehicle traffic on a road segment and evaluate traffic performance [46-53]. To compare the results of the analysis from VISSIM, the study also used the Indonesian Highway Capacity Manual (IHCM) 1997 as a comparison. The research results in several countries show that VISSIM can accurately evaluate traffic performance on a road segment. Thus, using VISSIM in traffic simulations can assist decision-makers in making more informed decisions regarding developing a better transportation system [54-58].

In the Indonesian context, which has challenges in overcoming traffic congestion, using VISSIM can assist in designing more effective and efficient transportation solutions. Using VISSIM, decision-makers can evaluate various traffic development schemes and select the most appropriate solution to address traffic problems in an area. This can help to improve the transportation system's performance in Indonesia and reduce traffic congestion, a significant problem in several big cities in Indonesia. In addition, the calibration carried out for drivers in Indonesia in this study cannot be immediately generalized to drivers in other countries. Driver behavior in each country can also vary, such as the level of discipline in following traffic rules, preparedness in dealing with emergencies, and awareness in driving. This difference may affect the driver's ability to follow the calibration model simulation results in other countries. Observations conducted in Dutch [59] and China [60] cities showed that drivers there had lower acceleration and desired speed profiles than observations in the Netherlands. Drivers in Indonesia may have experience driving on potholes or damaged roads, while drivers in other countries may not.

TABLE 7. Table of comparison of service levels resulting from IHCM 1997 and VISSIM

Time	IHCM 1997	VISSIM
Monday	D	B
Saturday	C	B

VISSIM is a traffic simulation software that models various transportation scenarios, but its accuracy can be influenced by external factors such as weather, accidents, or policy changes. Real-world traffic conditions are complex and dynamic, making it difficult to predict the impact of external factors. Therefore, it is important to exercise caution when interpreting VISSIM outcomes and to consider a range of factors, including historical data, expert insights, and real-world observations, for a comprehensive understanding of the transportation system. Traffic simulation models like VISSIM are only one tool among many for transportation planning and decision-making, and a holistic approach that considers economic, social, and environmental impacts, community needs, and priorities is necessary for effective, sustainable, and equitable transportation solutions that benefit all stakeholders.

4. CONCLUSION

Based on simulation results using VISSIM and the Indonesian Highway Capacity Manual (IHCM) 1997 method, there are differences in measuring the traffic service level. This demonstrates the importance of selecting the correct method and technique for analyzing and measuring traffic service levels. Using VISSIM as a microscopic simulation application can assist decision-makers in developing more effective and efficient transportation solutions to reduce congestion. With VISSIM, a traffic simulation is only a transportation planning and decision-making tool. The simulation results must be analyzed using data and field observations to draw more accurate and pertinent conclusions about the field situation. VISSIM must always be combined with field observations and accurate data for adequate and effective transportation solutions. VISSIM is a useful software instrument for researchers and transportation planners to evaluate road network performance, develop scenarios, and make better decisions to improve road network safety and performance.

Future research on VISSIM calibration may employ more precise techniques like drone technology to collect traffic data. Then, machine learning techniques can be used to predict simulation model parameters using historical data more accurately. The ultimate objective of this study is to improve the simulation model's ability to reflect actual traffic conditions. This research's findings can be utilized more effectively for infrastructure planning and decision-making that is more effective and efficient.

5. REFERENCES

1. Steiner, F. R., Butler, K., and American Planning Association. Planning and urban design standards. John Wiley & Sons. 2012.
2. Cox, P. Moving people: Sustainable transport development. Bloomsbury Publishing. 2010.
3. Vasconcellos, E. A. "Urban Transport Environment and Equity: The case for developing countries." Routledge. 2014.
4. Banister, D. Unsustainable Transport: City Transport in the New Century. Taylor & Francis. 2005.
5. Elmorssy, M., and Tezcan, H. O. "Application of discrete 3-level nested logit model in travel demand forecasting as an alternative to traditional 4-step model." ***International Journal of Engineering, Transactions A: Basics***, Vol. 32, No. 10, (2019), 1416-1428. https://doi.org/10.5829/ije.2019.32.10a.11
6. Lu, J., Li, B., Li, H., and Al-Barakani, A. "Expansion of city scale, traffic modes, traffic congestion, and air pollution." ***Cities***, Vol. 108, (2021), 102974. https://doi.org/10.1016/j.cities. 2020.102974
7. Armah, F. A., Yawson, D. O., and Pappoe, A. A. N. M. "A systems dynamics approach to explore traffic congestion and air pollution link in the city of Accra, Ghana." ***Sustainability***, Vol. 2, No. 1, (2010), 252-265. https://doi.org/10.3390/su2010252
8. Dimitriou, H. T., and Gakenheimer, R. Urban transport in the developing world: A handbook of policy and practice. Massachusetts: Edward Elgar Publishing. 2011.
9. Beaudoin, J., Farzin, Y. H., and Lawell, C.-Y. C. L. "Public transit investment and sustainable transportation: A review of studies of transit's impact on traffic congestion and air quality." ***Research in Transportation Economics***, Vol. 52, , (2015), 15-22. https://doi.org/10.1016/j.retrec.2015.10.004
10. Lei, Y., Hu, J., Fu, Y., Liu, Z., and Yan, B. "Simulation and Experimental Study of Vibration and Noise of Pure Electric Bus Transmission based on Finite Element and Boundary Element Methods." ***International Journal of Engineering, Transactions A: Basics***, Vol. 32, No. 7, (2019), 1023-1030. https://doi.org/10.5829/ije.2019.32.07 a.16
11. Transportation Simulation Systems. AIMSUN 5.0 Microsimulator User Manual. Barcelona: Transportation Simulation Systems. 2005.
12. Yang, Q., and Slavin, H. High fidelity, wide area traffic simulation model. Boston, USA: Caliper Corporation. 2002.
13. US Department of Transportation. FHWA CORSIM User Manual. McLean, Virginia: US Department of Transportation. 1998.
14. Salgado, D., Jolovic, D., Martin, P. T., and Aldrete, R. M. "Traffic microsimulation models assessment–a case study of international land port of entry." In ***Procedia Computer Science***, Vol. 83, 441-448, Elsevier. https://doi.org/10.1016/j.procs. 2016.04.207
15. Hadi, M., Sinha, P., and Wang, A. "Modeling reductions in freeway capacity due to incidents in microscopic simulation models." ***Transportation Research Record***, Vol. 1999, No. 1, (2007), 62-68. https://doi.org/10.3141/1999-07
16. Lin, D., Yang, X., and Gao, C. "VISSIM-based simulation analysis on road network of CBD in Beijing, China." ***Procedia-Social and Behavioral Sciences***, Vol. 96, (2013), 461-472. https://doi.org/j.sbspro.2013.08.054
17. Bloomberg, L., and Dale, J. "Comparison of VISSIM and CORSIM traffic simulation models on a congested network." ***Transportation Research Record***, Vol. 1727, No. 1, (2000), 52-60. https://doi.org/10.3141/1727-07
18. Sun, D. J., Zhang, L., and Chen, F. "Comparative study on simulation performances of CORSIM and VISSIM for urban street network." ***Simulation Modelling Practice and Theory***, Vol. 37, (2013), 18-29. https://doi.org/10.1016/j.simpat. 2013.05.007
19. Saidallah, M., El Fergougui, A., and Elalaoui, A. E. "A comparative study of urban road traffic simulators." In MATEC Web of Conferences, Vol. 81, EDP Sciences, 5002. https://doi.org/10.1051/matecconf/20168105002
20. Habtemichael, F. G., and de Picado Santos, L. "The need for transition from macroscopic to microscopic traffic management schemes to improve safety and mobility." ***Procedia-Social and Behavioral Sciences***, Vol. 48, (2012), 3018-3029. https://doi.org/10.1016/j.sbspro.2012.06.1269
21. Waddell, P. "UrbanSim: Modeling urban development for land use, transportation, and environmental planning." ***Journal of the American Planning Association***, Vol. 68, No. 3, (2002), 297-314. https://doi.org/10.1080/01944360208976274
22. Litman, T. "Developing indicators for comprehensive and sustainable transport planning." ***Transportation Research Record***, Vol. 2017, No. 1, (2007), 10-15. https://doi.org/ 10.3141/2017-02
23. Suh, W., Henclewood, D., Greenwood, A., Guin, A., Guensler, R., Hunter, M. P., and Fujimoto, R. "Modeling pedestrian crossing activities in an urban environment using microscopic traffic simulation." ***Simulation***, Vol. 89, No. 2, (2013), 213-224. https://doi.org/10.1177/0037549712469843
24. Park, B., and Schneeberger, J. D. "Microscopic simulation model calibration and validation: case study of VISSIM simulation model for a coordinated actuated signal system." ***Transportation Research Record***, Vol. 1856, No. 1, (2003), 185-192. https://doi.org/10.3141/1856-20
25. Mohan, R., Eldhose, S., and Manoharan, G. "Network-level heterogeneous traffic flow modelling in VISSIM." ***Transportation in Developing Economies***, Vol. 7, (2021), 1-17. https://doi.org/10.1007/s40890-021-00119-2
26. Barceló, J., and Casas, J. "Dynamic network simulation with AIMSUN." Simulation Approaches in Transportation Analysis: Recent Advances and Challenges, (2005), 57-98. https://doi.org/10.1007/0-387-24109-4_3
27. Park, B., and Qi, H. "Development and Evaluation of a Procedure for the Calibration of Simulation Models." ***Transportation Research Record***, Vol. 1934, No. 1, (2005), 208-217. https://doi.org/10.1177/0361198105193400122

28. Thacker, B. H., Doebling, S. W., Hemez, F. M., Anderson, M. C., Pepin, J. E., and Rodriguez, E. A. "Concepts of model verification and validation," (2004). https://doi.org/10.2172/835920

29. Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulos, H. N., and Wen, Y. "Calibration of microscopic traffic simulation models: Methods and application." ***Transportation Research Record***, Vol. 1999, No. 1, (2007), 198-207. https://doi.org/ 10.3141/1999-21

30. Zhe, L., Hao, L., and Ke, Z. "Calibration and validation of PARAMICS for freeway using toll data." In 2009 12th International IEEE Conference on Intelligent Transportation Systems, 1-6. https://doi.org/10.1109/ITSC.2009.5309678

31. Jha, M., Gopalan, G., Garms, A., Mahanti, B. P., Toledo, T., and Ben-Akiva, M. E. "Development and Calibration of a Large-Scale Microscopic Traffic Simulation Model." ***Transportation Research Record***, Vol. 1876, No. 1, (2004), 121-131. https://doi.org/10.3141/1876-13

32. Daigle, G., Thomas, M., and Vasudevan, M. "Field applications of CORSIM: I-40 freeway design evaluation, Oklahoma city, OK." In ***1***998 Winter Simulation Conference. Proceedings (Cat. No. 98CH36274), Vol. 2, IEEE, 1161-1167. https://doi.org/10.1109/WSC.1998.745974

33. Ratrout, N. T., Rahman, S. M., and Reza, I. "Calibration of paramics model: Application of artificial intelligence-based approach." ***Arabian Journal for Science and Engineering***, Vol. 40, (2015), 3459-3468. https://doi.org/10.1007/s13369-015-1816-5

34. Mathew, T. V, and Radhakrishnan, P. "Calibration of microsimulation models for nonlane-based heterogeneous traffic at signalized intersections." ***Journal of Urban Planning and Development***, Vol. 136, No. 1, (2010), 59-66. https://doi.org/10.1061/(ASCE)0733-9488(2010)136:1(59)

35. Stevanovic, J., Stevanovic, A., Martin, P. T., and Bauer, T. "Stochastic optimization of traffic control and transit priority settings in VISSIM." ***Transportation Research Part C: Emerging Technologies***, Vol. 16, No. 3, (2008), 332-349. https://doi.org/10.1016/j.trc.2008.01.002

36. Directorate General of Highways. Indonesian Highway Capacity Manual (IHCM). Jakarta, Indonesia: Department of Public Works. 1997.

37. PTV AG. PTV VISSIM 9.0 User Manual. Karlsruhe: PTV AG. 2016.

38. Highways Agency. Design Manual for Roads and Bridges. London: Her Majesty's Stationery Office. 1996.

39. Dowling, R., Skabardonis, A., and Alexiadis, V. *Traffic analysis toolbox, volume III: Guidelines for applying traffic microsimulation modeling software*. United States: Federal Highway Administration, Office of Operations. 2004.

40. Dowling, R. G. California Department of Transportation: Guidelines for Applying Traffic Microsimulation Modeling Software. California: Dowling Associates. 2002.

41. Smith, J., and Blewitt, R. Traffic modelling guidelines: TfL traffic manager and network performance best practice version 3.0. (3rd ed.). London: Transport for London. 2010.

42. Bowerman, B., O'Connell, R., and Koehler, A. Forecasting: methods and applications. Belmont, California: Thomson Brooks/Cole. 2004.

43. Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. Forecasting methods and applications (3rd ed.). Wiley India Pvt. Limited. 2008.

44. Lewis, C. International and Business Forecasting Methods. Oxfordshire: Butterworths. 1982.

45. National Research Council (U.S) Transportation Research Board. HCM 2010: Highway Capacity Manual. Washington, D.C.: Transportation Research Board. 2010.

46. Cremer, M., and Ludwig, J. "A fast simulation model for traffic flow on the basis of boolean operations." ***Mathematics and Computers in Simulation***, Vol. 28, No. 4, (1986), 297-303. https://doi.org/10.1016/0378-4754(86)90051-0

47. Marlianny, T. "Overview of Side Friction Factors for Evaluation of Capacity Calculation at Indonesian Highway Capacity Manual." In ***P***roceedings of the Second International Conference of Construction, Infrastructure, and Materials, Springer, 373-383. https://doi.org/10.1007/978-981-16-7949-0_33

48. Della, R. H., Arliansyah, J., and Artiansyah, R. "Traffic performance analysis of u-turn and fly over u-turn scenario; a case study at Soekarno Hatta Road, Palembang, Indonesia." ***Procedia Engineering***, Vol. 125, (2015), 461-466. https://doi.org/10.1016/ j.proeng.2015.11.123

49. Mahmudah, N., and Akbar, R. "Analysis of congestion cost at signalized intersection using Vissim 9 (Case study at Demak Ijo Intersection, Sleman)." In MATEC Web of Conferences (Vol. 181, p. 6001). EDP Sciences. https://doi.org/10.1051/matecconf/ 201818106001

50. Sandhyavitri, A., Maulana, A., Ikhsan, M., Putra, A. I., Husaini, R. R., and Restuhadi, F. "Simulation modelling of traffic flows in the central business district using PTV vissim in Pekanbaru, Indonesia." In ***Journal of Physics: Conference Series*** (Vol. 2049, p. 12096). IOP Publishing. https://doi.org/10.1088/1742-6596/2049/1/012096

51. Muchlisin, M., Wijayanti, F. A., and Amanda, N. "Traffic Detection Program using Image Processing and the 1997 Indonesian Highway Capacity Manual (MKJI)." In IOP Conference Series: Materials Science and Engineering (Vol. 1144, p. 12098). IOP Publishing. https://doi.org/10.1088/1757-899X/1144/1/012098

52. Munawar, A., Irawan, M. Z., and Fitrada, A. G. "Developing Indonesian Highway Capacity Manual Based on Microsimulation Model (A Case of Urban Roads)." In ***The World Congress on Engineering,*** Springer, 153-163. https://doi.org/10.1007/ 978-981-13-0746-1_12

53. Munawar, A. "Speed and capacity for urban roads, Indonesian experience." ***Procedia-Social and Behavioral Sciences***, Vol. 16, (2011), 382-387. https://doi.org/10.1016/j.sbspro.2011.04.459

54. Brockfeld, E., and Wagner, P. "Calibration and validation of microscopic traffic flow models." In Traffic and Granular Flow'03, Springer, 67-72. https://doi.org/10.1007/3-540-28091-X_6

55. Punzo, V., and Simonelli, F. "Analysis and comparison of microscopic traffic flow models with real traffic microscopic data." ***Transportation Research Record***, Vol. 1934, No. 1, (2005), 53-63. https://doi.org/10.1177/0361198105193400106

56. Ciuffo, B., Punzo, V., and Torrieri, V. "Comparison of simulation-based and model-based calibrations of traffic-flow microsimulation models." ***Transportation Research Record***, Vol. 2088, No. 1, (2008), 36-44. https://doi.org/10.3141/2088-05

57. Liu, B., Mehrara Molan, A., Pande, A., Howard, J., Alexander, S., and Luo, Z. "Microscopic Traffic Simulation as a Decision support system for road diet and tactical urbanism strategies." ***Sustainability***, Vol. 13, No. 14, (2021), 8076. https://doi.org/10.3390/su13148076

58. Lin, D., Yang, X., and Gao, C. "VISSIM-based simulation analysis on road network of CBD in Beijing, China." ***Procedia-Social and Behavioral Sciences***, Vol. 96, (2013), 461-472. https://doi.org/10.1016/j.sbspro.2013.08.054

59. Jie, L., van Zuylen, H. J., Chen, Y., and Lu, R. "Comparison of driver behaviour and saturation flow in China and the

Netherlands.” ***IET Intelligent Transport Systems***, Vol. 6, No. 3, (2012), 318-327. https://doi.org/10.1049/iet-its.2010.0203

60. Xu, J., Lin, W., Wang, X., and Shao, Y.-M. “Acceleration and deceleration calibration of operating speed prediction models for two-lane mountain highways.” ***Journal of Transportation Engineering, Part A: Systems***, Vol. 143, No. 7, (2017), 4017024. https://doi.org/10.1061/JTEPBS.0000050

Persian Abstract

چکیده

هدف این تحقیق کالیبراسیون و اعتبارسنجی ابزار مدل شبیه سازی VISSIM با مقایسه داده های میدانی با داده های شبیه سازی است. هدف نهایی ارزیابی عملکرد ترافیک با مقایسه نتایج شبیه سازی با مشاهدات مستقیم در میدان است. این مطالعه از مدل سازی برای تعیین حداکثر حجم جریان یک بخش جاده استفاده می کند. این مطالعه در ماکاسار، سولاوسی جنوبی، اندونزی، در جالان کهنه سرباز سلاتان انجام شد. این روش از دو ورودی اصلی استفاده می کند: داده های ظرفیت اولیه جاده های شهری از راهنمای ظرفیت بزرگراه اندونزی (IHCM 1997) و داده های فعالیت کنار جاده از PTV VISSIM. GEH و MAPE معمولاً از معیارهایی برای اندازه گیری دقت مدل های شبیه سازی و اندازه گیری های کالیبراسیون با استفاده از پارامترهای رفتار رانندگی استفاده می کنند. نتایج تحقیق به دست آمده برای اندازه گیری اعتبار، الزامات را برآورده کرده است. یعنی مقدار MEPE بدست آمده 10% (7.38%) کوچکتر از مقدار GEH بدست آمده (2.032 و 3.961) است که همچنان بیش از 5.00 است. اندازه گیری های کالیبراسیون مناسب بودن مکان وسیله نقلیه و فاصله بین خودرو را در مدل شبیه سازی (VISSIM) با شرایط میدان واقعی به دست آورد. نتایج به دست آمده از استفاده از VISSIM می تواند در طراحی و بهینه سازی سیستم های حمل و نقل شهری در آینده قابل اعتماد و کمک کننده باشد. مدل شبیه سازی ترافیک یک ابزار تصمیم گیری و برنامه ریزی است که باید با مشاهدات میدانی و داده های دقیق ترکیب شود تا راه حل های حمل ونقل کارآمد ایجاد کند.

International Journal of Engineering

Journal Homepage: www.ije.ir

Energy Management of an Integrated PV/Battery/Electric Vehicles Energy System Interfaced by a Multi-port Converter

P. Tarassodi[a], J. Adabi*[a], M. Rezanejad[b]

[a] *Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran*
[b] *Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran*

PAPER INFO

Paper history:
Received 7 April 2023
Received in revised form 22May 2023
Accepted 24 May 2023

Keywords:
Energy Mannagement
Multiport Converter
Electric Vehice
Photovoltaic System

ABSTRACT

Integrated energy systems, including renewable energy sources (RES) and battery energy storage (BES), have high potentialities to deal with issues caused by the high penetration of electric vehicles (EVs) in power systems. The full realization of the benefits of such systems depends on implementation of an energy management system (EMS) in order to monitor power sharing between different components of the system. In this paper, an EMS is proposed for a multi-port converter as an integrated PV/BES/EV energy system. It takes into account the EV mileage, BES dis/charge cycles and financial benefits, and schedule for the optimal dis/charge of batteries, and also involves EVs in V2X programs. In this approach, the potential of EVs as a portable energy storage can be employed in providing ancillary services to the power grid. The obvious advantages of the proposed EMS performance have been specified by simulation and comparison with the benchmark method. According to the obtained results, for a specific period of time, a better interaction has been established between the average achievement of the final SOC and the financial profit of the integrated energy system under the proposed EMS. According to the proposed method, for a 10% reduction in the final SOC compared to the benchmark method, the minimum financial benefit is about 0.2607 pounds (received from the grid), equivalent to 0.2082 pounds (paid to the grid) in the benchmark method.

doi*: 10.5829/ije.2023.36.08b.12*

1. INTRODUCTION

In past decades, the tendency to use RES in power systems has increased as a promising solution to deal with environmental, technical and economic challenges. Concerns regarding the reliability of their operation have also increased. One solution is the integration of RES with BES, which result in financial benefits from the sale of excess power [1]. Such systems also play a significant role in supporting the power grid. For instance, with the increasing penetration of EV, providing charging points based on integrated energy systems is considered a suitable solution to avoid problems caused by the increase in power grid load demand. Such that it provides a clean charging point, without applying an additional load on the power grid. Integrated energy systems are created by combining different sources such as RES, BES, EV and Grid with power electronic interfaces that exchange power under the supervision of a control system to achieve technical and financial objectives.

So far, various researches have investigated several aspects of the implementation of integrated energy systems and their performance. An EMS strategy based on particle swarm optimization (PSO) and fuzzy controller for a microgrid consisting of distributed generation resources and energy storage system consisting of batteries and supercapacitors are proposed by Sepehrzad et al. [2]. A family of multifunctional multi-port converters suitable for PV-based EV charging stations and grid-connected BES is introduced and simulated by Gohari et al. [3]. Active and reactive exchange power control is carried out by two-loop PI control scheme in this structure. Moreover, a rule-based EMS is presented to improve system reliability at a

*Corresponding Author Email: j.adabi@nit.ac.ir (J. Adabi)

Please cite this article as: P. Tarassodi, J. Adabi, M. Rezanejad, Energy Management of an Integrated PV/Battery/Electric Vehicles Energy System Interfaced by a Multi-port Converter, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1520-1531

reduced cost. A dis/charging scheduling algorithm based on a chance constrained programming method for an EV charging station including BES and PV is proposed by Li et al. [4]. In the proposed algorithm, EV dis/charging is influenced by PV uncertainty, dis/ charging priorities and electricity price, and it's able to reduce energy costs by 50%. A practical test of a smart charging controller based on PSO-ANN-Fuzzy is conducted by Ali et al. [5], which considers user needs, energy tariff, grid conditions (for example, voltage or frequency), renewable (PV) output, and battery's state of health (SOH) status. The intended structure has RES and BES and is connected to the three-phase power grid from one side. A topology for a multi-port DC/DC/AC converter, suitable for use in BES-based hybrid microgrids is proposed by Zhang et al. [6]. Also, a detailed description of the control system and decentralized power management of the proposed structure has been discussed. A multi-mode hierarchical power management strategy for a smart home as a nine-port energy router is proposed by Wang et al. [7]. Nine-ports have been provided through separate conventional converters. The proposed strategy is based on droop and three-mode switching relationships to ensure the power balance of the entire system, which sends the current references to the decentralized controller. In the controller module, the power sharing of scattered productions is realized in order to support the voltage and frequency of the power system. A smart charging approach for off-grid electric chargers in home applications including PV, BES, and EV is proposed by Gholinejad et al. [8]. The optimal charging profile of BES and EV is carried out through Bellman-Ford-Moor algorithm, in which the financial benefits of EV and home owners have been fulfilled by considering their comfort level. A new interface topology is created by Savrun et al. [9] by combining several existing topologies to integrate two EVs with home DC microgrids. The proposed power management algorithm in this reference is rule-based that determines the state of power flow between the elements of this structure. Its control system is also based on voltage loops that stabilize voltage of the ports on their reference values. A new multiport converter for use in systems with storage is proposed by Yi et al. [10]. In this topology, two active switches are used to provide two bidirectional ports and one unidirectional DC/DC port. A new structure for connecting BESs, EVs and RES is investigated by Engelhardt et al. [11], in which a set of strings is employed instead of the power electronic interface. The proposed EMS for this structure assigns an appropriate string for two fast charging points of EVs according to the PV and SOC production power of BESs. An EMS for a grid-connected charging station is proposed by Mumtaz et al. [12], which is capable of simultaneous scheduling of dis/charging of five different EVs. The proposed algorithm makes it possible to implement V2X and X2V scenarios including 7 different operational modes. The adaptive PID control schematic has been employed to control converters of this system.

To sum up, Table 1 compares the major features in existing researches with those of presented in this article where the relevant researches can be categorized into three general types: power electronics interface topology [3, 6], control systems [9, 12] and EMS strategies [2, 4, 5, 7, 8, 10, 11]. In general, it has been observed that in some of these papers, EMS strategies and control schematics have not been verified through experimental tests. In some studies, the description of power electronic converters as integral components of integrated energy systems has not been addressed. Also, the researches that focused on topology are different from each other in terms of the number of bidirectional ports and simultaneous availability of DC and AC ports. In addition, in some works, control of voltage and current of sources has been neglected, while without an efficient control system, power sharing between different sources will not be possible. Battery dis/charging plans are often aimed at power balance, and the lack of short and long-term EMS is clearly felt considering electricity purchase and sale tariffs. The performance of integrated energy systems, to a large extent, depends on the EMS strategy, because it's responsible for scheduling, monitoring and controlling power exchanges between different components of the system [11]. The objectives of EMS strategies include charging BES through excess power [13], minimizing energy costs [1, 4, 8], reducing grid pressure [12], and also reducing system losses [14, 15]. Thus, the main focus here is on presenting an EMS strategy for an integrated energy system, which aims to reduce energy costs and facilitate EV participation in V2X programs, as a potential energy source. Also, the maximum energy that can be stored in batteries (BES and EV), based on the battery health curve-the number of dis/charge cycles, is included in the proposed EMS. In this way, the gradual effect of battery degradation on EMS scheduling can be investigated. However, in order to set up and examine the overall performance of such a system, aspects of the topology of the power electronics interface and its control system have also been discussed. In general, the major contributions of this paper are stated below:

- Proposing an EMS based on mathematical relations with easy implementation
- Integrated MIMO converter control
- Facilitating EV connection/disconnection to/from EMS and Reducing energy price of EV charging
- Facilitating use of EV in V2X programs and Considering SOH of batteries in EMS based on dis/charge number

This paper is organized as follows: description of the MIMO converter topology and control schematic is provided in the second part. The relations and flowchart of the proposed method are presented in the third part. The third section includes the system model, how to apply restrictions and process of EMS. The fourth part examines and discusses the simulation results. In the end, a summary and results of the study are presented in conclusion in the fifth section. The novelty of this paper lies in the development of a power electronic-based control algorithm for power flow management in a grid-connected PV/BES/EV energy systems. This algorithm dynamically allocates power based on real-time energy generation, consumption, and EV charging requirements as well as aiming to maximize overall financial benefits of the system and considering battery degradation. By referring to Table 1, it is evident that there is a clear lack of studies addressing all major aspects of EMS implementation. In addition, different mannagment scenarios of such system was investigated [16-21].

2. SYSTEM DESCRIPTION

The general schematic of the integrated energy system is displayed in Figure 1. This system has three parts: Power electronics interface converter, control system and EMS. The power electronics interface is a MIMO converter that has two bidirectional DC/DC ports, one unidirectional DC/DC port and one bidirectional AC/DC port. This converter provides the possibility of power exchange between PV, EV, BES and the power grid under a fewer number of switches. So, compared to topologies with similar capabilities, two less switches are used in its structure. The MIMO converter consists of connecting two DC/DC and DC/AC sections; in the first section, dis/charging of batteries is provided through the DC-link or through each other. In the second part, it’s possible to inject power from the DC-link to the power grid, from the power grid to the DC-link, and from PV to the DC-link. In this way, power exchanges in this topology can be conducted with different objectives, including sharing PV and BES for EV charging and reducing power grid stress. Moreover, this system provides the possibility of dis/charging BES and EV to achieve financial benefits and participate in V2X applications.

In a recent study conducted by Tarassodi et al. [22], performance of the integrated energy system under the above-mentioned topology has been thoroughly

TABLE 1. Comparison table for existing studies

Ref	Sources	Power electronics		Controller system			EMS constrains		
		No. of Bidirectional Ports	**No. of switches**	**Constant power**	**Constant voltage**	**V2X**	**Battery degradation**	**Tariff**	**EV revenue**
[1]	Grid, BES, EV, PV, CHP	2 DC, 1 AC	9	✓	✓	V2G, V2BES	×	MCP	×
[3]	Grid, BES, EV, PV	1 DC, 1 AC	7	✓	✓	×	×	×	×
[4]	Grid, BES, EV, PV	-	-	×	×	V2G	✓	MCP	✓
[5]	Grid, BES, EV, PV	-	-	×	×	V2G, V2BES	✓	MCP	✓
[6]	AC and DC Grids, BES	3 DC, 1 AC	9	✓	✓	×	×	×	×
[7]	Hybrid AC/DC microgrid	1 DC, 1 AC	23	✓	✓	×	×	×	×
[8]	Grid, BES, EV, PV	2 DC, 1 AC	9	✓	✓	V2G, V2BES	×	MCP	×
[9]	EV	1 DC	9	✓	✓	V2H	×	×	×
[10]	BES, EV, PV	1 DC	2	×	×	×	×	×	×
[13]	Grid, BES, PV	1 DC, 1 AC	9	✓	✓	×	×	×	×
[16]	Grid, BES, PV	×	×	×	×	×	✓	MCP and FIT	×
[17]	Grid, BES, PV	1 DC, 1 AC	8	✓	✓	×	×	×	×
[18]	Grid, EV, PV	1 AC	15	✓	✓	×	×	×	×
[19]	Grid, BES	1 DC, 1 AC	14	✓	✓	×	×	×	×
[20]	Grid, BES, PV, Fuel cell	1 DC	14	✓	✓	×	×	×	×
[21]	DC source, Super capacitor	×	6	✓	✓	×	×	×	×
Proposed	Grid, BES, EV, PV	2 DC, 1 AC	9	✓	✓	V2G, V2BES	✓	MCP and FIT	✓

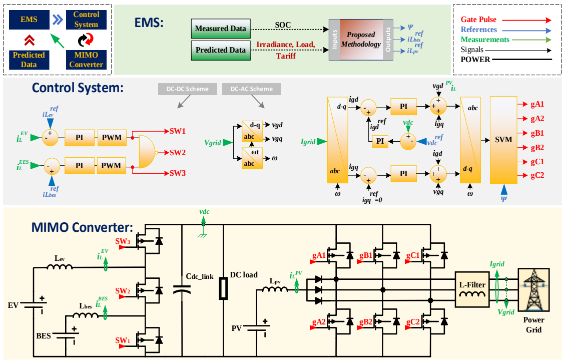


Figure 1. General view of integrated energy system

investigated. With continuous sampling of voltage of the power grid and DC-link as well as current of the inductors, the control system stabilizes values of these quantities on the determined references. The control schematics of DC/DC and DC/AC sections are separated in Figure 1. The control system in DC/DC section includes current loops that apply error of the inductor current to PWM unit after passing through the PI regulator. In this way, switching pulse of SW_1 and SW_3 switches are generated. Also, the switching signal of SW_2 switch is obtained with NAND of two SW_1 and SW_3 signals. In this schematic, the current reference values iL_{EV}^{ref} and iL_{BES}^{ref} are determined by EMS, as discussed in the next section. In the DC/AC control schematic, by using a voltage loop, the voltage error of DC-link is passed through a PI regulator to obtain the value of the direct component of the inverter current reference. This value determines the amount of power that must be exchanged by the inverter to stabilize DC-link voltage at the specified reference value.

In this schematic, the space vector modulation (SVM) method is applied so that in addition to its inherent advantages, PV system integrated with the three-phase, three-leg inverter can be controlled. In this way, one less switch is used in MIMO converter. In order to control PV, the application of ψ parameter to SVM algorithm is used to change the duration of vector [1,1,1] in SVM according to the environmental conditions such as the intensity of radiation and temperature. Because under the other seven vectors, the PV inductor is only charged and as long as vector [1,1,1] is applied, the energy stored in PV inductor is discharged in DC-link capacitor. Although there are some challenges in implementation of such an integrated control system, this paper focuses on EMS, as described in detail in the next section.

3. METHODOLOGY

In this section, the proposed method of this paper is described in how to share resources to gain financial and technical benefits. To begin with, Equation (1) represents the power balance in MIMO converter, in which PV generated power (P_{PV}), BES power in discharge mode (P_{BES}^{DisCh}) and EV power in discharge mode (P_{EV}^{DisCh}) are considered positive. In contrast, BES power in charging mode, EV power in charging mode and load power are considered negative. As a result, the positive value of P_{Grid} is indicative of the injection of excess power of the integrated energy system towards the power grid and vice versa.

$$P_{Grid} = P_{PV} + P_{BES}^{DisCh} + P_{EV}^{DisCh} - P_{BES}^{Ch} - P_{EV}^{Ch} - P_{LOAD} \quad (1)$$

where the amount of load consumption power (P_{LOAD}) is considered available and definite for the day-ahead. Also, P_{PV} is calculated according to the radiation intensity profile, which normally starts from zero and reaches maximum value from early morning to noon. Then, it gradually reaches zero until night. Here, the radiation intensity is considered definite and predetermined; however, to achieve the maximum power point (MPP) for changes in the radiation intensity, the PV system is modeled further. Equation (2) analytically expresses the I-V characteristic of the solar cell [23]. In this model, effect of the resistance of series and parallel branches in the solar cell model is addressed. Despite the existence of more accurate models, this model establishes an interaction between accuracy and simplicity [24, 25].

$$I = N_p I_{pv,cell} - N_p I_{0,cell}\left[exp\left(\frac{q(V+R_s I_{0,cell})}{N_s akT}\right) - 1\right] - \frac{V+R_s I_{0,cell}}{R_p} \quad (2)$$

where, q is the charge of an electron (Coulomb), k is Boltzmann's constant (J/K), T is the temperature of the p-n junction (K), a is the ideality constant of the diode, N_p is the number of parallel cells, N_s is the number of series cells, R_s is the equivalent series resistance and R_p is the equivalent parallel resistance. Also, $I_{0,\ cell}$ is the reverse saturation current of the diode [23]:

$$I_{0,cell} = \frac{I_{sc,n}+K_I\Delta T}{exp\left(\frac{q(V_{oc,n}+K_V\Delta T)}{(N_s akT)}\right)-1} \quad (3)$$

in which, K_V is the voltage coefficient, $I_{\ sc,n}$ is the short circuit current and $V_{\ oc,n}$ is the open circuit voltage of the cell under nominal conditions. $I_{pv,\ cell}$ is the production current of a cell, which has a direct relationship with the intensity of radiation and temperature. Equation (4) describes the mathematical description of the impact of intensity of radiation and temperature on this parameter.

$$I_{pv,cell} = \left(I_{pv,n} + K_I\Delta T\right)\frac{G}{G_n} \quad (4)$$

where, K_I is the current coefficient, G_n is the nominal radiation amount (W/m^2), the difference between the ambient temperature and nominal condition temperature (Kelvin) and $I_{pv,n}$ is the nominal current of the cell under nominal conditions. By specifying the current value, the

PV output voltage value is also obtained from the I-V curve. Multiplying voltage and current of PV has only one maximum point, which may not be on the MPP due to the environmental conditions. To this end, the perturb and observe (P&O) algorithm is used for MPPT in PV model, and Algorithm 1 represents the pseudo code of its modelling [23]. In this algorithm, the applied signal ψ is generated by SVM method instead of calculating duty cycle of the switches.

Algorithm 1. MPPT Algorithm

```
//Input: Instantaneous current, Instantaneous voltage, $\Psi(0)=0$.
Calculate instantaneous power:
if ($p_{pv}$(t)> $p_{pv}$(t-1))
    if (v(t)> v(t-1))
        $\Psi(t) = \Psi(t-1) + \Delta\Psi$; // $\Delta\Psi \approx 1\mu s$
    else
        $\Psi(t) = \Psi(t-1) - \Delta\Psi$;
    end
elseif (($p_{pv}$(t)< $p_{pv}$(t-1))
    if (v(t)> v(t-1))
        $\Psi(t) = \Psi(t-1) - \Delta\Psi$;
    else
        $\Psi(t) = \Psi(t-1) + \Delta\Psi$;
    end
else
        $\Psi(t)= \Psi(t-1)$;
end
//Output: $\Psi(t)$.
```

In this way, by calculating PV generated power and the availability of load consumption power, only by determining BES and EV charging and discharging power values by EMS, it's possible to determine the exchanged power value of the energy system integrated with the power grid. For this purpose, first, the standard discrete time model of the battery is placed in the power balance relation. Then, in each time step, for different values of BES and EV power, the minimum amount of energy cost or maximum financial benefit of the integrated energy system is calculated based on purchase/sale tariff values. Thus, the P_{Grid} value for the next time step is determined. The discrete time model of the stored energy of the battery is shown in Equation (5), in which E[k] is the energy stored in the battery at the instant of k ∈ τ = {0, 1, …, T-1}6. Moreover, η_{Ch} and η_{DisCh} are charge and discharge efficiency of the battery, respectively, and $P_{Ch}[k]$ and $P_{DisCh}[k]$ are charge and discharge power of the battery in kth time step, respectively, and Δt>0 indicate one time step length [26].

$$E[k+1] = E[k] + \eta_{Ch}\Delta t P_{Ch}[k] - \frac{\Delta t}{\eta_{DisCh}} P_{DisCh}[k], \forall k \in \tau \tag{5}$$

The constraints of the problem for ∀k ∈ τ are also given in Equations (6) to (10). The set of Equations (5) to (10) has been used to model BES and EV, with the difference that initial and nominal energy values, maximum dis/charging power, as well as charging and discharging efficiency of BES and EV are considered differently [26-30].

$$E[0] = E_0 \tag{6}$$

$$0 \le P_{Ch}[k] \le P_{Ch}^{max} \tag{7}$$

$$0 \le P_{DisCh}[k] \le P_{DisCh}^{max} \tag{8}$$

$$0 \le E[k+1] \le E_{max} \tag{9}$$

$$P_{Ch}[k]P_{DisCh}[k] = 0 \tag{10}$$

Equation (10) models the condition of complementarity of battery charging and discharging, which proves that a battery cannot be charged and discharged simultaneously. This equality has created a non-linear term in energy management calculations, which challenges the solution of such problems. Therefore, Equation (5) will become Equation (11) by removing restriction (10) and simplifying.

$$E[k+1] = E[k] + \eta \Delta t P_{Bat}[k], \qquad \forall k \in \tau \tag{11}$$

$$-P_{CH}^{max} \le P_{Bat}[k] \le P_{DisCh}^{max} \tag{12}$$

In Equation (11), an input ($P_{Bat}[k]$) is considered to calculate battery energy, defined as follows:

$$\begin{aligned} P_{Bat}[k] &= P'_{DisCh}[k] - P'_{Ch}[k], \qquad \forall k \in \tau \\ P'_{Ch} &= max\{0, -(P_{DisCh} - P_{Ch})\} \\ P'_{DisCh} &= max\{0, P_{DisCh} - P_{Ch}\} \end{aligned} \tag{13}$$

According to Equation (11), for each possible value of E[k+1], a P_{Bat} value is obtained for batteries. Since ΔE changes in each time step is constraint by the inequality (12), for all the values that apply to this inequality, one should look for a P_{Bat} value for BES and EV, so that by placing it in Equation (1), the optimal value of exchanged power between the power grid and the integrated energy system can be obtained. Under the optimal amount of exchanged power, there would be the lowest energy cost, the highest financial profit and the lowest battery degradation. For this reason, after inserting the possible values of E[k+1] and calculating P_{Bat} (for BES and EV separately), taking into account parameters such as electricity purchase/sale tariffs, number of dis/charging cycles, EV mileage and minimum depth of discharge (DOD), the optimal value of exchanged power are obtained. The intended multi-objective function in the proposed EMS is provided in Equation (14). This objective function seeks to increase SOC and simultaneously, increase the financial benefit (reducing the cost paid to the grid or increasing cost received from the grid). For this purpose, in addition to the grid power, the power of the batteries is also multiplied in the purchase and sale tariffs depending on their sign.

$$[P_{Grid}, P_{BES}, P_{BES}, \pi_p, \pi_i] = \begin{cases} if(P_{Grid} \geq 0) \begin{cases} if(P_{BES} \geq 0) \begin{cases} \frac{(P_{BES}+P_{EV})}{P_{Grid}}, if(P_{EV} \geq 0) \\ \frac{(P_{BES}\pi_i+P_{EV}\pi_p)}{P_{Grid}\pi_i}, if(P_{EV} < 0) \end{cases} \\ if(P_{BES} < 0) \begin{cases} \frac{(P_{BES}\pi_p+P_{EV}\pi_i)}{P_{Grid}\pi_i}, if(P_{EV} \geq 0) \\ \frac{(P_{BES}\pi_p+P_{EV}\pi_p)}{P_{Grid}\pi_i}, if(P_{EV} < 0) \end{cases} \end{cases} \\ f(P_{Grid} < 0) \begin{cases} if(P_{BES} \geq 0) \begin{cases} \frac{(P_{BES}\pi_i+P_{EV}\pi_i)}{P_{Grid}\pi_p}, if(P_{EV} \geq 0) \\ \frac{(P_{BES}\pi_i+P_{EV}\pi_p)}{P_{Grid}\pi_p}, if(P_{EV} < 0) \end{cases} \\ if(P_{BES} < 0) \begin{cases} \frac{(P_{BES}\pi_p+P_{EV}\pi_i)}{P_{Grid}\pi_p}, if(P_{EV} \geq 0) \\ \frac{(P_{BES}+P_{EV})}{P_{Grid}}, if(P_{EV} < 0) \end{cases} \end{cases} \end{cases} \quad (14)$$

The electricity purchase/sale tariff information is available for the day-ahead and like PV and load, is considered to be predetermined and definite. The number of dis/charging cycles, as a criterion in determining SOH of the BES battery is applied as a constraint of the problem and overshadows decision-making process in the proposed EMS. For this purpose, the declining curve of the maximum storable energy of BES in terms of changes in number of dis/charging cycles is applied to the problem as an input. The number of dis/charging cycles for a new battery is considered zero. But, for each charge or discharge, its value increases in the proposed algorithm. Moreover, EV mileage is considered as a criterion of its SOH in the proposed EMS. Similarly, the declining curve of SOH with respect to mileage changes is used to limit the problem in terms of the maximum energy that can be stored in EV.

The application of E[k+1] quantification restriction is well shown in Figure 2. As shown in the figure, in k+1 time step, due to an increase in the number of dis/charging cycles or operation compared to before, the final limit of the energy that can be stored in the battery has slightly decreased. Hence, the range of changes of possible value of E[k+1] in k+1 time step has become more limited than before. This would contribute to significant changes in EMS decision-making in the long run. Also, some E[k+1] values are not acceptable. Because in order to achieve these values, the exchanged battery power will exceed the nominal value of condition (12). These values are considered unacceptable and are not considered in the EMS process.

With respect to EV application, in addition to the constraints mentioned in Figure 2, another constraint is applied to E[k+1] value. As it's clear, EV initially behaved as a consumer that needs charging at various home, fast and ultra-fast levels. Depending on the conditions of the integrated energy system, power grid and electricity tariff, one of the above-mentioned charging levels is selected. Anyway, EV working mode is X2V; While based on the objectives of this paper, it may also be used in V2X applications if the EV owner approves. In this case, in X2V mode, E[k+1] quantification takes place only for $\forall k \in \tau \rightarrow E[k] \leq E[k+1]$. In other words, the inequality (12) will change to expression (15) in X2V program. However, in V2X program, it is conducted according to the explanations of Figure 2.

$$-P_{CH}^{max} \leq P_{EV}[k] \leq 0\,, \quad \forall k \in \tau \quad (15)$$

By calculating the average BES dis/charge times and EV performance in 24 hours, it is possible to determine the effect of battery degradation on power sharing in the long term. Moreover, the proposed EMS seeks to reduce the charging time in attaining higher SOCs (lower DOD), which is restricted by the maximum power that can be transferred to the battery, maximum exchanged power with the power grid, financial benefit and, the costs.

The performance description of the proposed EMS is provided in flowchart of Figure 3, in which N_{Ch}^{BES}, N_{Ch}^{EV}, π_p and π_i are the number of BES dis/charge, number of dis/charging cycles, electricity purchase tariff from the power grid and electricity feed-in-tariff (FIT) to the power grid, respectively. The quantification of $E_{BES}[k+1]$ and $E_{EV}[k+1]$ is carried out from the minimum to maximum value of the inequality (9) under

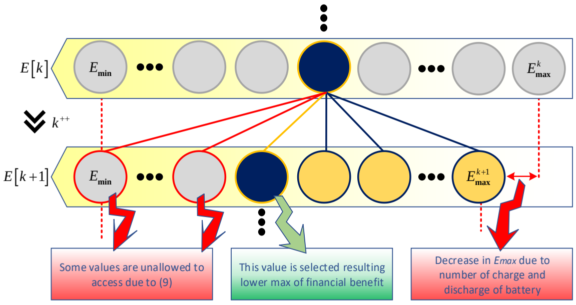

E[k]
E_min
E^k_max
k^++
E[k+1]
E_min
E^k+1_max
Some values are unallowed to access due to (9)
This value is selected resulting lower max of financial benefit
Decrease in Emax due to number of charge and discharge of battery

Figure 2. Quantification and restriction of $E[k+1]$

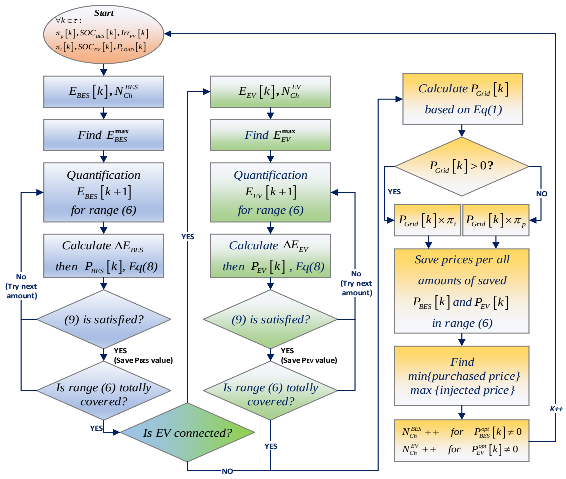
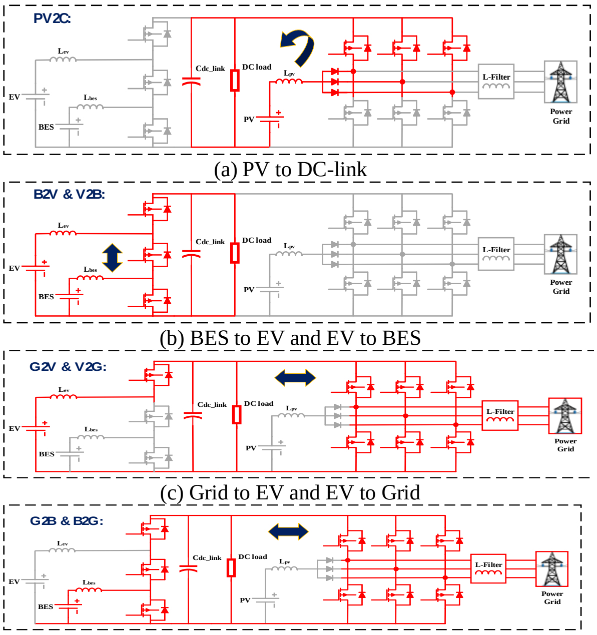

Start
$\forall k \in \tau$:
$\pi_p[k], SOC_{BES}[k], Irr_{PV}[k]$
$\pi_i[k], SOC_{EV}[k], P_{LOAD}[k]$
$E_{BES}[k], N_{Ch}^{BES}$
Find E_{BES}^{max}
Quantification $E_{BES}[k+1]$ for range (6)
Calculate ΔE_{BES} then $P_{BES}[k]$, Eq(8)
(9) is satisfied?
YES (Save PBES value)
Is range (6) totally covered?
No (Try next amount)
YES
Is EV connected?
YES
NO
$E_{EV}[k], N_{Ch}^{EV}$
Find E_{EV}^{max}
Quantification $E_{EV}[k+1]$ for range (6)
Calculate ΔE_{EV} then $P_{EV}[k]$, Eq(8)
(9) is satisfied?
YES (Save PEV value)
Is range (6) totally covered?
No (Try next amount)
YES
Calculate $P_{Grid}[k]$ based on Eq(1)
$P_{Grid}[k] > 0$?
YES
NO
$P_{Grid}[k] \times \pi_i$
$P_{Grid}[k] \times \pi_p$
Save prices per all amounts of saved $P_{BES}[k]$ and $P_{EV}[k]$ in range (6)
Find min{purchased price} max {injected price}
N_{Ch}^{BES} ++ for $P_{BES}^{opt}[k] \neq 0$
N_{Ch}^{EV} ++ for $P_{EV}^{opt}[k] \neq 0$
K++

Figure 3. Flowchart of proposed EMS

certain intervals. The maximum value of this inequality is obtained according to N_{Ch}^{BES} and N_{Ch}^{EV} values, based on the battery degradation diagram.

By specifying $P_{EV}[k+1]$, $P_{BES}[k+1]$ and $P_{PV}[k+1]$ values, the MIMO converter is switched under one of the scenarios shown in Figure 4 or a combination of them. As shown, except for part (a), the rest of the scenarios are bidirectional power exchange. Also, it is possible to simultaneously implement one of the scenarios of one part with the scenarios of other parts. For example, PV2*DC*link, B2V, and G2V scenarios can be simultaneously implemented, which represents EV charging by PV, BES, and the power grid. This means that the proposed EMS performs both decision-making and determination of power references at the same time. In the previous research conducted by Tarassodi et al. [22], performance of the MIMO converter in terms of the control system, efficiency, reliability, as well as the implementation of different performance scenarios, has been fully investigated and discussed. Therefore, in the next section, only results of the implementation of the proposed EMS on the integrated energy system under study have been examined and compared.



(a) PV to DC-link

(b) BES to EV and EV to BES

(c) Grid to EV and EV to Grid

(d) Grid to BES and BES to Grid

Figure 1. Different operational scenarios of proposed EMS

4. RESULTS AND DISCUSSION

In this section, a case study has been simulated in MATLAB, results of which are presented below. Also, in order to check the simulation results, the Simulated Annealing (SA) optimization method has been used as a benchmark. To this end, the SA method has been implemented for a system similar to the proposed method in MATLAB, then the SA analysis has been investigated under the same inputs.

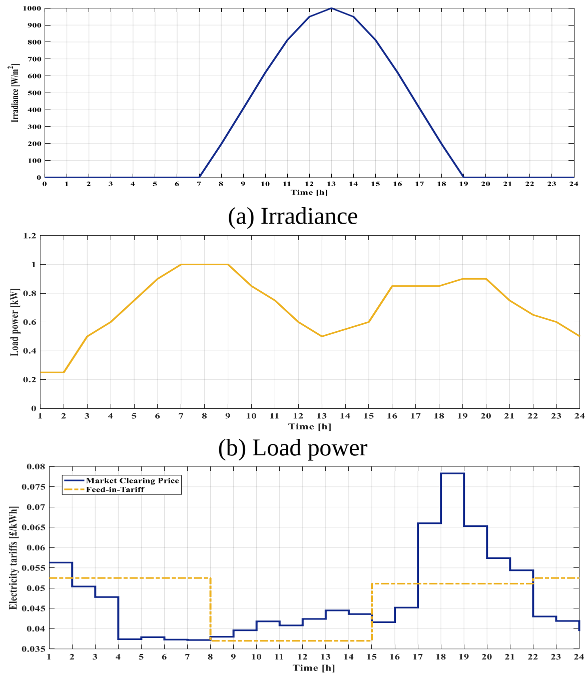
The simulation features of the intended case study are provided in Table 1. The nominal capacity of EV and BES batteries is given in this Table, which by taking into account the EV performance and the number of BES dis/charge cycles**,** determine the maximum energy that can be stored in the batteries. The primary SOC of the batteries is selected as a sample and can change within the permissible range during the planning. The maximum dis/charge power of the batteries are also determined given the capacity of the batteries and MIMO converter. The PV system is also an array of 16 cells connected in series/parallel with the specifications provided in Table 2. Simulation specifications include predicted input data, such as radiation intensity, load consumption power, and electricity tariffs, as shown in Figure 5.

The battery degradation curve is shown in Figure 6. Changes in SOH of BES in relation to an increase in frequency of dis/charging is illustrated in Figure 6(a). As shown in this figure, SOH remains constant until about 100 dis/charge times, then, it gradually reduces. Since the number of EV dis/charge cycles for charging point is not known, the SOH of EV is obtained from Figure 6(b), which is based on mileage changes.

TABLE 2. Simulation parameters value

Description	Value	Unit
Nominal Capacity of BES	1.7	kWh
Nominal Capacity of EV	20	kWh
BES usage	300	Cycle
EV mileage	50000	Miles
BES initial SOC	70	%

EV initial SOC	50	%
BES Dis/Charging power	0.5	kW
EV Dis/Charging power	5	kW
Dis/Charging efficiency	96	%
Lower/upper charge limits	[20-90]	%
Max of irradiance	1000	W/m^2
PV cell current in MPP	7.61	A
PV cell voltage in MPP	26.3	V
Number of PV cells connected in series	8	-
Number of PV cells connected in parallel	2	-

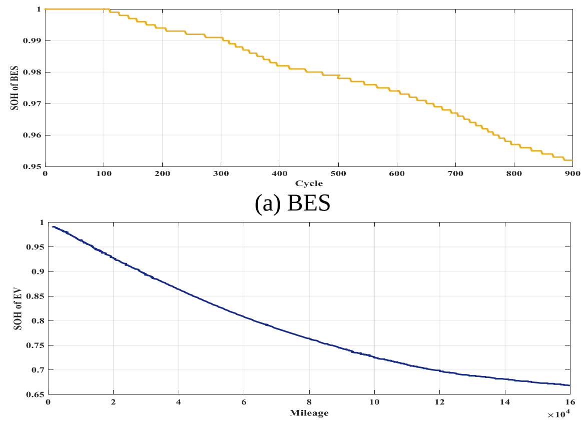


(a) Irradiance

(b) Load power

(c) Electricity tariffs

Figure 5. Day-ahead data



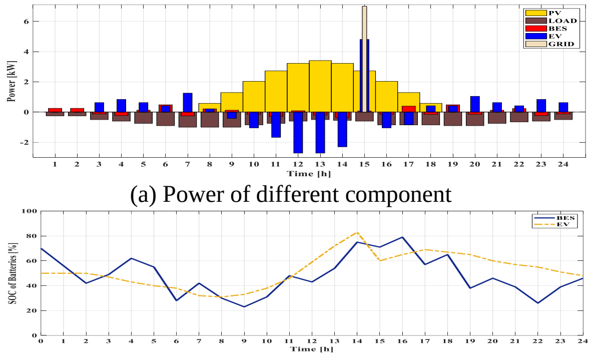
(a) BES

(b) EV

Figure 6. Battery degradation curves

Given the flowchart in Figure 3, with the availability of the predicted inputs as well as measurement of the primary SOC of the batteries, E[k+1] quantification begins. As shown in the explanation of Figure 2, the quantification is conducted under the interval (9) and then, based on conditions such as V2G working mode, the SOH value and the maximum dis/charging power of the batteries would be restricted. Changes in the resource power and SOC of the batteries during 24 hours are shown in Figure 7. These results are achieved under the number of dis/charge cycles of 10 BES and operating range of 1000 miles of EV. In Figure 7(a), the power of the sources that are in the generator state (such as PV and batteries in discharge mode) is positive, and power of the load and batteries in charging mode is negative. As it is shown, with an increase in the PV power, the excess power available in the integrated energy system is used for charging EV. However, in time steps of 14 to 15 (plotted as a column in 15), due to the increase of FIT and increase of EV SOC, not only the EV charging is stopped, but also it is discharged to some extent and the obtained excess power is injected into the power grid. SOC changes of batteries corresponding to paragraph (a) are shown in Figure 7(b). As it's clear in the figure, depending on the conditions of each time step, the batteries are charged or discharged during 24 hours, so that the objective function of the problem is minimized in that time step. Hence, a curve including several charging scenarios and several discharging scenarios has been created.

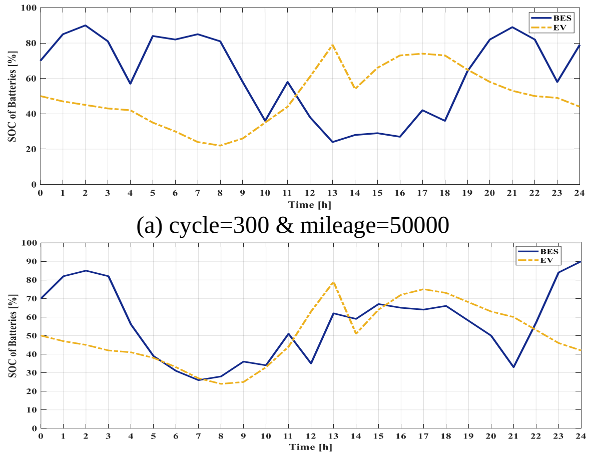
SOC changes of the batteries under the condition of 300 and 700 cycles, and 50 and 100 thousand miles of operation are illustrated in Figure 8. It is clear in this figure that increasing life of batteries directly affects the integrated energy system planning, such that the dis/charge profile of the batteries has changed compared to before. These changes are more evident in the BES profile, which has a much lower capacity than EV. In the following, the numerical results of these analyses are summarized in Table 2.



(a) Power of different component

(b) SOC of BES and EV

Figure 7. Changes in power and SOC per cycle=10 & mileage=1000



(a) cycle=300 & mileage=50000

(b) cycle=700 & mileage=100000

Figure 8. SOC of BES and EV under different conditions

On the other hand, increasing EV lifetime, as a high-capacity source, has a greater impact on the amount of EMS income. The 24-hour income of the integrated energy system under cycle and operation changes is shown in Figure 9. It is well known that the increasing EV battery life, as a higher capacity resource, has the greatest impact on the revenue from BES and EV integration in the energy system under study.

Results shown in Figure 10 were obtained under the condition of no EV presence. As it is evident, without the presence of EV, more excess power is available for injection into the power grid. However, the interesting point is that this does not necessarily cause the integrated energy system to achieve higher financial benefits, because it lost the opportunity to benefit from a high-capacity energy storage; and this has reduced the financial benefit of the system for 24 hours. This is because of the fact that FIT is low in the range where more excess power is available. In paragraph (b), SOC BES was obtained under different cycles that did not experience any specific changes without the presence of EV. The numerical results of these analyses are demonstrated in Table 2.

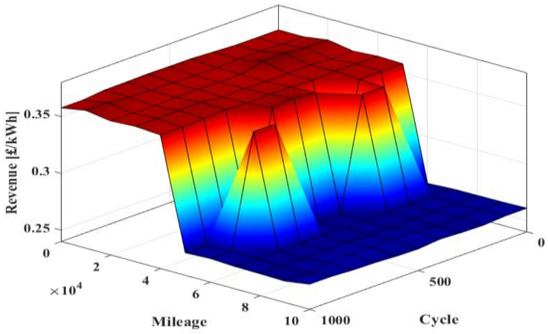
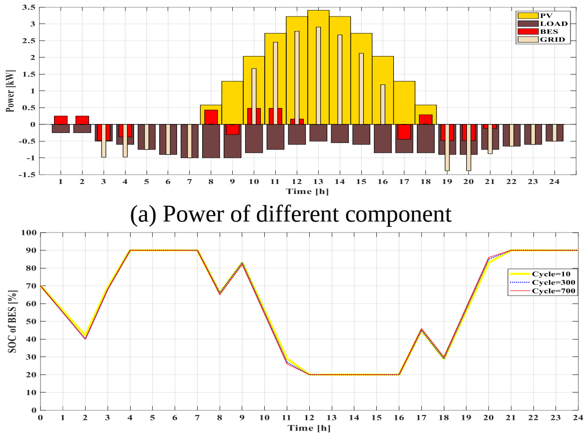


Figure 9. 24-hour energy price per mileage changes



(a) Power of different component

(b) SOC of BES

Figure 10. Changes in power and SOC without EV

The impact of adding EV on the financial benefit of the integrated energy system is shown in Figure 11. This figure proves the effect of the presence of a high-capacity energy storage source (i.e., EV) on the overall financial benefit of the system. Such that in the presence of low-functioning EV, the financial benefit has increased more than 2.3 times compared to the absence of EV. Over time, as the performance of EV increases, this ratio decreases until it reaches 1.6 times per 100,000 miles of operation.

A comparison of the performance of the proposed method with a basic method, namely, SA is illustrated in Figure 12. The time step of this comparison is 1 second and performance of the system under the two proposed and benchmark methods is performed for 86400 seconds.

Results obtained for EV indicate that the final SOC is almost equal, but the dis/charging profile of the batteries under the benchmark method is such that, in total, much less financial benefit is achieved. These results are given in Table 2. So, under the SA method, not only the system has not achieved the financial benefit, but also it has to pay an amount to the power grid. By comparing these results, it’s clear that a 10% reduction in SOC in the proposed method is totally acceptable in return for the financial benefit gained. For BES, the newer the batteries are, the greater the final SOC difference under the two

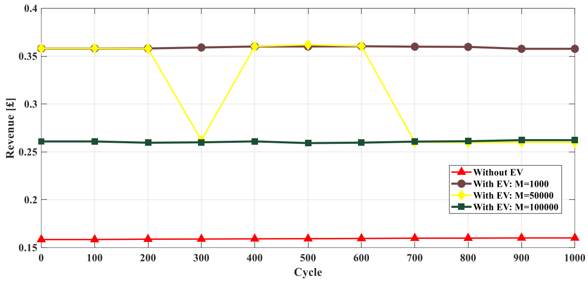
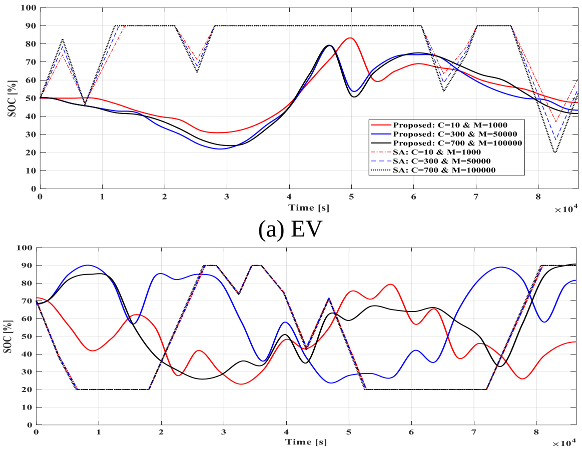


Figure 11. 24-hour energy price per mileage changes, Cycle=0



(a) EV

(b) BES

Figure 12. Batteries dis/charging schedule comparison

proposed and SA methods. However, with an increase in lifespan, the final SOC of the batteries has also been obtained similar to each other. In fact, this is a good indication that under the proposed method, an acceptable interaction has been created between the gained financial benefit and the final SOC value of the batteries.

The numerical comparison between the results obtained from the proposed methods and SA, under different cycle and mileage conditions are provided in Table 3. As mentioned in the explanation of the above analysis, the maximum financial benefit under the proposed method is obtained in the presence of EV and SOH above EV.

TABLE 3. Simulation parameters value

Method	Inputs	Cycle	Mileage	Cycle	Mileage	Cycle	Mileage
		10	1000	300	50000	700	100000
Proposed	With EV	0.3580		0.2621		0.2607	
	Without EV	0.1584		0.1589		0.1598	
SA		-0.2979		-0.2344		-0.2082	

5. CONCLUSION

In this paper, an EMS based on mathematical relations is proposed for an integrated PV/BES/EV energy system. The proposed EMS generates BES and EV power references to minimize a multi-objective function, such that it would lead to the reduction of system energy payment costs to the grid and the increase of system costs received from the grid, taking into consideration the final SOC. In the proposed EMS relations, the maximum energy that can be stored in BES and EV are constraint based on the number of dis/charge cycles and mileage, respectively. In this way, an analysis of the effect of increasing the life of batteries and the presence or absence of EV as a high-capacity storage device on the financial benefit is provided. Based on the results obtained from the numerical analysis, it was found that although in the absence of EV, the system sells more energy to the grid, it does not achieve more financial benefits. Because it loses the possibility of using a main source of energy storage in optimal scheduling. Because a significant part of the energy sold was in the absence of EV when FIT is low. On the other hand, in the presence of EV, it is possible to sell significant energy to the grid at high FITs, and the system would achieve a higher financial benefit as a whole.

6. REFERENCES

1. Gholinejad, H.R., Adabi, J. and Marzband, M., "An energy management system structure for neighborhood networks", ***Journal of Building Engineering***, Vol. 41, (2021), 102376. doi: 10.1016/j.jobe.2021.102376.
2. Sepehrzad, R., Moridi, A.R., Hassanzadeh, M.E. and Seifi, A.R., "Intelligent energy management and multi-objective power distribution control in hybrid micro-grids based on the advanced fuzzy-pso method", ***ISA Transactions***, Vol. 112, (2021), 199-213. doi: 10.1016/j.isatra.2020.12.027.
3. Gohari, H.S., Safaeinasab, A. and Abbaszadeh, K., "Family of multifunctional controllable converters for grid, battery, and pv-powered ev charging station applications", in 2022 30th International Conference on Electrical Engineering (ICEE), IEEE. (2022), 681-687.
4. Li, D., Zouma, A., Liao, J.-T. and Yang, H.-T., "An energy management strategy with renewable energy and energy storage system for a large electric vehicle charging station", ***Etransportation***, Vol. 6, (2020), 100076. doi: 10.1016/j.etran.2020.100076.
5. Ali, Z., Putrus, G., Marzband, M., Gholinejad, H.R., Saleem, K. and Subudhi, B., "Multiobjective optimized smart charge controller for electric vehicle applications", ***IEEE Transactions on Industry Applications***, Vol. 58, No. 5, (2022), 5602-5615. doi: 10.1109/TIA.2022.3164999.
6. Zhang, Z., Jin, C., Tang, Y., Dong, C., Lin, P., Mi, Y. and Wang, P., "A modulized three-port interlinking converter for hybrid ac/dc/ds microgrids featured with a decentralized power management strategy", ***IEEE Transactions on Industrial Electronics***, Vol. 68, No. 12, (2020), 12430-12440. doi: 10.1109/TIE.2020.3040660.
7. Wang, R., Jiang, S., Ma, D., Sun, Q., Zhang, H. and Wang, P., "The energy management of multiport energy router in smart home", ***IEEE Transactions on Consumer Electronics***, Vol. 68, No. 4, (2022), 344-353. doi: 10.1109/TCE.2022.3200931.
8. Gholinejad, H.R., Adabi, J. and Marzband, M., "Hierarchical energy management system for home-energy-hubs considering plug-in electric vehicles", ***IEEE Transactions on Industry Applications***, Vol. 58, No. 5, (2022), 5582-5592. doi: 10.1109/TIA.2022.3158352.
9. Savrun, M.M., İnci, M. and Büyük, M., "Design and analysis of a high energy efficient multi-port dc-dc converter interface for fuel cell/battery electric vehicle-to-home (V2H) system", ***Journal of Energy Storage***, Vol. 45, (2022), 103755. doi: 10.1016/j.est.2021.103755.

10. Yi, W., Ma, H., Peng, S., Liu, D., Ali, Z.M., Dampage, U. and Hajjiah, A., "Analysis and implementation of multi-port bidirectional converter for hybrid energy systems", ***Energy Reports***, Vol. 8, (2022), 1538-1549. doi: 10.1016/j.egyr.2021.12.068.

11. Engelhardt, J., Zepter, J.M., Gabderakhmanova, T. and Marinelli, M., "Energy management of a multi-battery system for renewable-based high power ev charging", ***Etransportation***, Vol. 14, (2022), 100198. doi: 10.1016/j.etran.2022.100198.

12. Mumtaz, S., Ali, S., Ahmad, S., Khan, L., Hassan, S.Z. and Kamal, T., "Energy management and control of plug-in hybrid electric vehicle charging stations in a grid-connected hybrid power system", ***Energies***, Vol. 10, No. 11, (2017), 1923. doi: 10.3390/en10111923.

13. Yi, Z., Dong, W. and Etemadi, A.H., "A unified control and power management scheme for pv-battery-based hybrid microgrids for both grid-connected and islanded modes", ***IEEE Transactions on Smart Grid***, Vol. 9, No. 6, (2017), 5975-5985. doi: 10.1109/TSG.2017.2700332.

14. Gamboa, G., Hamilton, C., Kerley, R., Elmes, S., Arias, A., Shen, J. and Batarseh, I., "Control strategy of a multi-port, grid connected, direct-dc pv charging station for plug-in electric vehicles", in 2010 IEEE Energy Conversion Congress and Exposition, IEEE. (2010), 1173-1177.

15. Acha, S., Green, T.C. and Shah, N., "Effects of optimised plug-in hybrid vehicle charging strategies on electric distribution network losses", in IEEE PES T&D 2010, IEEE. (2010), 1-6.

16. Azuatalam, D., Paridari, K., Ma, Y., Förstl, M., Chapman, A.C. and Verbič, G., "Energy management of small-scale pv-battery systems: A systematic review considering practical implementation, computational requirements, quality of input data and battery degradation", ***Renewable and Sustainable Energy Reviews***, Vol. 112, (2019), 555-570. doi: 10.1016/j.rser.2019.06.007.

17. Tang, C.-Y., Chen, P.-T. and Jheng, J.-H., "Bidirectional power flow control and hybrid charging strategies for three-phase pv power and energy storage systems", ***IEEE Transactions on Power Electronics***, Vol. 36, No. 11, (2021), 12710-12720. doi: 10.1109/TPEL.2021.3083366.

18. Kar, R.R. and Wandhare, R.G., "Energy management system for photovoltaic fed hybrid electric vehicle charging stations", in 2021 IEEE 48th Photovoltaic Specialists Conference (PVSC), IEEE. (2021), 2478-2485.

19. Tang, W., Li, Z., Wang, Y., Zhang, C., Shao, L. and Wang, K., "Fpga-based real-time simulation for multiple energy storage systems", in Journal of Physics: Conference Series, IOP Publishing. Vol. 1659, (2020), 012047.

20. Jafari, M., Malekjamshidi, Z., Platt, G., Zhu, J.G. and Dorrell, D.G., "A multi-port converter based renewable energy system for residential consumers of smart grid", in IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society, IEEE. (2015), 005168-005173.

21. Ahmeti, F. and Arnaudov, D., "Energy flows management of a multi-port dc-dc converter for an energy storage system", in 2022 13th National Conference with International Participation (ELECTRONICA), IEEE. (2022), 1-4.

22. Tarassodi, P., Adabi, J. and Rezanejad, M., "A power management strategy for a grid-connected multi-energy storage resources with a multiport converter", ***International Journal of Circuit Theory and Applications***, https://doi.org/10.1002/cta.3540

23. Gholinejad, H.R., Loni, A., Adabi, J. and Marzband, M., "A hierarchical energy management system for multiple home energy hubs in neighborhood grids", ***Journal of Building Engineering***, Vol. 28, (2020), 101028. doi: 10.1016/j.jobe.2019.101028.

24. Villalva, M.G., Gazoli, J.R. and Ruppert Filho, E., "Comprehensive approach to modeling and simulation of photovoltaic arrays", ***IEEE Transactions on Power Electronics***, Vol. 24, No. 5, (2009), 1198-1208. doi: 10.1109/TPEL.2009.2013862.

25. Carrero, C., Amador, J. and Arnaltes, S., "A single procedure for helping pv designers to select silicon pv modules and evaluate the loss resistances", ***Renewable energy***, Vol. 32, No. 15, (2007), 2579-2589. doi: 10.1016/j.renene.2007.01.001.

26. Nazir, N. and Almassalkhi, M., "Guaranteeing a physically realizable battery dispatch without charge-discharge complementarity constraints", ***IEEE Transactions on Smart Grid***, (2021). doi: 10.1109/TSG.2021.3109805.

27. Sagar, G. and Debela, T., "Implementation of optimal load balancing strategy for hybrid energy management system in dc/ac microgrid with pv and battery storage", ***International Journal of Engineering, Transactions A: Basics*** Vol. 32, No. 10, (2019), 1437-1445. doi: 10.5829/ije.2019.32.10a.13.

28. Basu, A. and Singh, M., "Design and real time digital simulator implementation of a takagi sugeno fuzzy controller for battery management in photovoltaic energy system application", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 35, No. 12, (2022), 2275-2282. doi: 10.5829/ije.2022.35.12c.01.

29. Ahmadigorji, M. and Mehrasa, M., "A robust renewable energy source-oriented strategy for smart charging of plug-in electric vehicles considering diverse uncertainty resources", ***International Journal of Engineering, Transactions A: Basics***, Vol. 36, No. 4, (2023), 709-719. doi: 10.5829/ije.2023.36.04a.10.

30. Ashabi, A., Peiravi, M.M., Nikpendar, P., Salehi Nasab, S. and Jaryani, F., "Optimal sizing of battery energy storage system in commercial buildings utilizing techno-economic analysis", ***International Journal of Engineering, Transactions B: Applications***, Vol. 35, No. 8, (2022), 1662-1673. doi: 10.5829/ije.2022.35.08b.22.

Persian Abstract

چکیده

سیستم‌های انرژی ادغام شده شامل منابع انرژی تجدیدپذیر (RES) و ذخیره‌ساز انرژی باتری (BES)، از پتانسیل بالایی برای مقابله با مشکلات ناشی از نفوذ بالای خودروهای الکتریکی در سیستم‌های قدرت برخودار می‌باشند. تحقق کامل مزایای چنین سیستم‌هایی در گرو پیاده‌سازی یک سیستم مدیریت انرژی (EMS) به منظور نظارت بر اشتراک‌گذاری توان میان اجزای مختلف سیستم می‌باشد. در این مقاله، یک EMS برای یک مبدل چند پورت، به عنوان یک سیستم انرژی ادغام شده PV/BES/EV، پیشنهاد شده است. EMS پیشنهادی، با در نظر گرفتن میزان EV mileage، تعداد شارژ/دشارژ BES و منافع مالی، به برنامه‌ریزی بهینه‌ی شارژ/دشارژ (dis/charging) باتری‌ها پرداخته و EV را در برنامه‌های V2X نیز شرکت می‌دهد. بدین ترتیب، می‌توان از پتانسیل بالقوه‌ی EVها به عنوان یک ذخیره‌ساز پرتابل، در ارائه خدمات جانبی به شبکه قدرت نیز استفاده کرد. مزایای بارز عملکرد EMS پیشنهادی با انجام شبیه‌سازی و مقایسه با روش معیار مشخص شده‌اند. با توجه به نتایج بدست آمده، به ازای یک بازه زمانی مشخص، میان میانگین دستیابی به SOC نهایی و سود مالی سیستم انرژی ادغام شده تحت EMS پیشنهادی تعامل بهتری برقرار شده است. به طوریکه تحت روش پیشنهادی، در ازای ۱۰ درصد کاهش در SOC نهایی نسبت به روش معیار، حداقل منفعی مالی حدود 0.2607 پوند (دریافتی از شبکه) بدست آمده که معادل آن در روش معیار 0.2082- پوند (پرداختی به شبکه) می‌باشد.

International Journal of Engineering

Journal Homepage: www.ije.ir

The Effect of Using Reinforced Granular Blanket and Single Stone Column on Improvement of Sandy Soil: Experimental Study

A. Shahmandi[a], M. Ghazavi*[b], K. Barkhordari[a], M. Hashemi[c]

[a] *Department of Civil Engineering, Yazd University, Yazd, Iran*
[b] *Civil Engineering Department, K.N. Toosi University of Technology, Tehran, Iran*
[c] *Department of Civil Engineering, University of Isfahan, Isfahan, Iran*

PAPER INFO

Paper history:
Received 12 April 2023
Received in revised form 08 June 2023
Accepted 10 June 2023

Keywords:
Geogrid Reinforcement
Granular Blanket
Laboratory Model Test
Sand Bed
Stone Column
Unit Cell

ABSTRACT

A series of large-scale laboratory model tests in a unit cell was performed to explore the behaviour of loose sandy soil due to improvement. An unreinforced and geogrid reinforced granular blanket, a single end-bearing stone column, and their combination were used for this purpose. Since the rupture of the geosynthetic reinforcement in the reinforced granular blanket has never been experimentally investigated. A novel method of installing the geogrid was used. Thus, geogrid was allowed to completely mobilize and fail under loads. In this investigation, load-settlement characteristics have been generated by continuing loading even after geogrid rupture until the desired settlement. Parametric studies were carried out to observe the effect of important factors, such as the blanket thickness and the layout of geosynthetic sheets, including the number and place of geogrid layers within the granular blanket. Reinforcing the blanket with geogrid while changing the usual form of the load-settlement characteristics has had a significant effect on enhancing load-carrying capacity and reducing settlement. It can be said using a stone column, granular blanket, or combination of both techniques to boost load-carrying capacity was more effective than reducing settlement. However, the effect of single-layer and double-layer geogrid reinforcement on settlement reduction depends on their placement within the granular blanket. In addition, the efficiency of improvement methods has been superior under looser bed conditions. The best layout was to arrange one layer of geogrid near the top of the blanket or two layers in the middle and near the top.

***doi**: 10.5829/ije.2023.36.08b.13*

NOMENCLATURE

D_r	Relative density	D	Diameter of the footing
LR	Load ratio	S/D	Settlement ratio
LR_{max}	Maximum load ratio	BCR	Bearing capacity ratio
LR_{final}	Final load ratio	B_f	Footing width
S	Footing settlement	Z_u	Distance of the uppermost reinforcing row from the base of the footing

1. INTRODUCTION

Nowadays, soil improvement techniques [1-5] are widely used. Knowing about soil improvement techniques and their applications is essential to ensure the safety and cost-effectiveness of projects. These methods are used to modify the properties of soil to improve its stability, strength, and bearing capacity, which is vital when constructing structures such as buildings, bridges, and roads.

Stone columns have been used successfully to improve the engineering properties of different types of soils, such as soft clays, silts, and silty sands. However, despite the stone column's advantages in improving ground behaviour, its performance is challenging in soft or loose soils. In these soils, the circumferential

*Corresponding Author Email: ghazavi_ma@kntu.ac.ir (M. Ghazavi)

Please cite this article as: A. Shahmandi, M. Ghazavi, K. Barkhordari, M. Hashemi, The Effect of Using Reinforced Granular Blanket and Single Stone Column on Improvement of Sandy Soil: Experimental Study, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1532-1547

confinement offered by the surrounding soil may not be sufficient to develop the appropriate load-carrying capacity. As a result, the stone column bulges and pushes the surrounding soil radially, reducing efficiency. Hence, studies have tried to employ various geosynthetics to reinforce the stone column and provide confining pressure around it [6-12], and reduce stress concentration at the column's top by placing a reinforced granular blanket [13, 14]. In response to this improvement, the stone column can take vertical loads and reduce bulging. The load-carrying capacity of the stone column and soil is increased due to the reinforced blanket's beam-like behaviour and ability to withstand some bending.

Deb et al. [15] developed a mechanical model to predict the behaviour of a geosynthetic-reinforced granular fill over soft soil that improved with stone columns. They found that adding the geosynthetic layer reduced the total and differential settlement, while the settlement reduction increased with further load intensity and modular ratio. Moreover, Deb et al. [16] extended a mechanical model to investigate the behaviour of multi-layer geosynthetic-reinforced granular fill over stone column-reinforced soft soil. It has been reported that compared to a single layer of reinforcement, using multi-layer geosynthetic reinforcement along with stone columns had less effect in reducing the settlement. However, when soft soil had not included stone columns, the multilayer-reinforced system remarkably efficiently reduced maximum settlement. Laboratory model investigations on the unreinforced and geogrid-reinforced sand bed over an end-bearing stone column-improved soft clay were performed by Deb et al. [17] in a cubic tank. They determined the optimum thickness of unreinforced and geogrid-reinforced sand beds and the optimal size of geogrid reinforcement placed at the bottom of the sand bed. Elsewhere, Debnath and Dey [18] conducted a series of laboratory model tests on an unreinforced sand bed and a geogrid-reinforced sand bed placed over a group of vertically encased stone columns floating in soft clay, as well as their numerical simulations. They reported increased load-carrying capacity, optimal thickness of the unreinforced and reinforced sand bed, and the optimum diameter of the geogrid placed at the bottom of the sand bed while utilizing a reinforced sand bed along with encased stone columns. Finally, Mehrannia et al. [19] studied the effect of floating stone columns, granular blankets, and the combination of both methods in unreinforced and reinforced modes on improving the bearing capacity of scaled physical models in a cubic tank. Their findings showed the enhanced bearing capacity of the clay bed by using improvement methods. It has also been noted that applying geogrid reinforcement in the middle of the granular blanket and geotextile as stone column encasement has considerably improved their efficiency.

Most stone column experimental studies have thus far been conducted on saturated clay beds. Meanwhile, laboratory studies on a stone column overlying with a granular layer of sand or gravel in a unit cell have been insufficiently examined. In none of the prior studies, the granular blanket has been reinforced with the horizontal geosynthetic reinforcement sheets in the unit cell. In addition, a few studies have been reported in the literature indicating the geosynthetic-reinforced granular blanket can noticeably enhance the bearing capacity of the foundation system [17, 18, 20]. However, earlier laboratory studies have modelled a single stone column with a reinforced granular blanket in cubic tanks, which has not considered the concept of the stone column within a group. Moreover, the reinforcement has been applied in optimum dimensions with sizes larger than the stone column diameter and the free end within the granular blanket.

The failure mechanism of planar geosynthetics reinforced foundations has yet to be well investigated and understood. Therefore, to better analyze the failure mechanisms, it is necessary to examine the rupture of the geosynthetic reinforcement layers during loading. Besides, the placement of geogrid reinforcement near the top of the blanket positioned over the stone column-improved bed has yet to be investigated. The present study investigates the effect of end-bearing stone columns, unreinforced and geogrid-reinforced granular blankets, and their combinations in a laboratory unit cell for improving the loose silty sand bed. A novel approach is also adopted to install granular blanket reinforcement in the unit cell, allowing the geogrid reinforcement to mobilize under the applied loads. A principal objective of this study is to conduct large-scale laboratory model testing to examine the effect of some parameters, such as the thickness of the blanket and reinforcing layout, including the number and place of geogrid reinforcement within the blanket. Other objectives of the present study are to reveal the rupture of geogrid reinforcement and to discover the relationship between the failure of the geogrid layers and the characteristics of load-carrying capacity and settlement of the model tests. It should note that compared to the load-carrying capacity, fewer experimental investigations have been conducted on reducing settlement, especially in physical modelling with the unit cell.

2. MATERIALS AND EQUIPMENT USED

2. 1. Sand and Aggregate Materials

Since the reinforcement of fine-grained sandy soil using a stone column is intended, the test bed sample was prepared from an admixture of desert sand and clay, with a particle size distribution curve within the range of the Vibro Replacement method [21, 22]. Hence, the mixed sample can be classified as SM per the Unified classification system. A mix of fine sand and clay, with a relative density (D_r) of 25%, was used to provide the test sand bed in loose mode.

A significant number of blows for the compaction of the stone column and blanket materials in the laboratory process can affect the relative density of the loose sand bed and crush the aggregates of the stone column. Because of this, sand and aggregate materials with self-compacting properties and a particular grain size range were employed to lower the number of required hammer blows. The particle size distribution curves of materials used as loose sand beds, granular blankets and stone columns are presented in Figure 1.

According to guidelines suggested by Nayak [23] and Fattah et al. [24], the size of the aggregate used in the construction of the stone column should be 1/7 to 1/6 of the diameter of the stone column. Based on the works of Fox [25], Stoeber [26], and Mohapatra et al. [27], a value of 1/6 is satisfactory for this ratio. Therefore, crushed stone materials passing through sieve No 3/8 inch and remaining on sieve No 4 with grain sizes ranging from 4.75 mm to 9.5 mm were used to construct stone columns. The ratio of this material's largest to the smallest grain size is equal to 2. However, it is tough to ensure a uniform diameter of a stone column with a relative density greater than 50% [28]. Therefore, the relative density of the stone column was chosen as 50%, and for the granular blanket, it was 70%. Sand with grain sizes ranging from 0.6 mm to 1.7 mm was chosen as a granular blanket, with a grain size ratio of 2.8. Table 1 summarizes the properties of the sand and aggregate materials utilized in the laboratory model tests.

2. 2. Geogrid Reinforcement Finding geosynthetics with the required and reduced stiffness on a laboratory scale is extremely difficult. It is because manufacturers do not generate materials with the appropriate stiffness for the physical models. Therefore, fibreglass mesh with the specifications listed in Table 2 has been used to reinforce the blanket in the tests. Its resistance parameters have been determined based on the ASTM D6637 [29]. Gniel and Bouazza [30] also used fibreglass and aluminium mesh as reinforcement in their laboratory research to model a stone column encased by geogrid. The desirable aperture size of geogrid reinforcement is roughly 3.5 times the average soil particle size, D_{50} [31]. Accordingly, an available geogrid of the aperture size of 5 mm × 5 mm was used to reinforce the granular blankets. In addition, a comparison of the average size of sand blanket grains and the size of fibreglass mesh aperture indicates that Koerner [31] recommendation, as well as the scale effect, was taken into account.

2. 3. Unit Cell Stone columns are often installed in a triangle or square pattern in a group with a certain influence area for each column. In the present experimental study, unit cell idealization of a single stone column within a triangular pattern of a group of columns has been used. The unit cell is the equivalent cylindrical influence area of a single stone column within a group of columns [21].

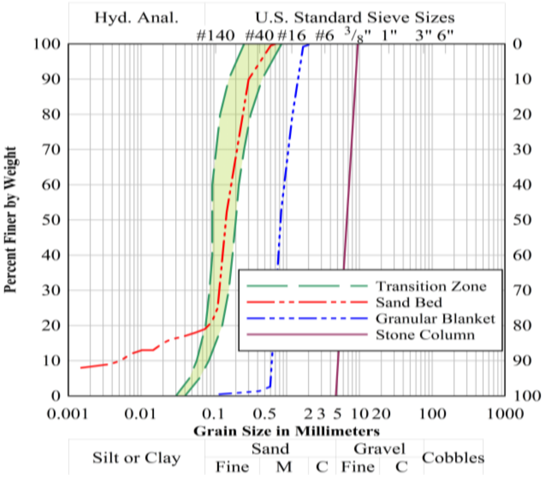


Figure 1. Particle size distribution curves for sands and aggregate materials

TABLE 1. Properties of sands and aggregate materials

Parameter	Sand Bed	Granular Blanket	Stone Column
Specific gravity	2.67	2.66	2.65
Minimum dry unit weight (kN/m^3)	14.67	13.96	14.03
Maximum dry unit weight (kN/m^3)	18.71	17.36	17.21
Bulk unit weight (kN/m^3)	15.51 (D_r =25%)	16.18 (D_r =70%)	15.46 (D_r =50%)
Internal friction angle (degree)	32 (D_r =25%)	35 (D_r =70%)	41 (D_r =50%)
Uniformity coefficient (C_u)	34.8	1.48	1.43
Curvature coefficient (C_c)	15	0.9	0.91
USCS classification	SM	SP	GP

TABLE 2. Properties of geogrid reinforcement

Parameter	Value
Ultimate tensile strength (kN/m)	8
Strain at ultimate strength (%)	3.17
Stiffness at ultimate strain (kN/m)	250
Mesh aperture (mm)	5×5
Mass (g/m^2)	75

In this research, all large-scale laboratory model tests have been performed in a cylindrical steel tank representing the unit cell with 208 mm inside diameter, 6 mm thickness, and 525 mm initial height. The height of the used unit cell can be increased up to 675 mm, by adding modular rings, each with a height of 15 mm made of the unit cell materials. Thus, while carrying out the blankets with variable thickness, in the cases where a fibreglass mesh is reinforcing the blanket, it can be appropriately restrained within the distance between the rings using the pressure resulting from the closure of the retaining nuts and the drop glue (Figure 2(a)).

As displayed in Figure 2(b), a support grid is placed on top of the unit cell, keeping the pipe's head in place. The inner surface of the unit cell was coated with electrostatic paint to reduce friction between the tank's wall and the materials within. In the unit cell theory, radial stiffness is infinite; thus, the outer body was braced by two steel rings to prevent any radial deformation.

2. 4. Test Setup

In this study, the pressure was applied to the surface of the models in the unit cell using a hydraulic jack-frame arrangement with a nominal capacity of 10 tons and load cells connected to it with capacities of 5 tons and 10 tons. A circular steel plate of diameter 200 mm and thickness 20 mm was used as test footing to transfer the uniform stress on the model's surface. The footing has a diameter of 8 mm less than the inner diameter of the unit cell, and the foam has been rolled around it to provide three following functions:

- It prevents inaccuracy in the test caused by the footing's contact with the body of the unit cell.
- It positions the footing in the unit cell's centre, parallel to the unit cell wall.
- It maintains soil grains, especially the granular pad, from migrating around the footing.

The intended load was applied as displacement control with a 1 mm/min strain rate in all tests [32-37]. This strain rate has been set based on the type of materials, their moisture, and the performance of trial tests to control the gradual densification of sand throughout the unit cell's entire height.

To ensure that pressure is applied uniformly on the whole surface of the model tests, the load cell is joined to the footing with a pin connection, according to Figure 3. Another steel plate with a 100 mm diameter and a 15 mm thickness is also welded in the centre of the primary footing. A hole with a diameter and depth equal to 25 mm and 10 mm, respectively, was made in the second plate to adjust the steel ball placement connected to the bolt's end. Finally, a bolt and a steel ball attached to it transfer the load as a perpendicular force from the load cell to the footing. Two LVDTs with a displacement range of 100 mm and an accuracy of 0.01 mm were utilized to record changes in the footing's settlement. Two LVDTs were installed diagonally near the edges of the footing, as depicted in Figure 3(a).

3. PREPARATION OF MODEL TESTS

An identical procedure was utilized in all tests to prepare the sand bed and construct the stone column and the granular blanket. At each step, a given amount of each material was poured into the unit cell based on their determined unit weight and desired volume. Before filling the unit cell with sand, the inner surface was coated with oil and grease to minimize friction between the wall and the materials. A stone column with a displacement construction method was formed in the centre of the unit cell with an open-ended thin-walled pipe, having an approximate thickness of 1.8 mm, and an outside diameter of 75 mm. The surrounding of the stone column was covered by much thin nylon with a meagre tensile strength to prevent sand migration into the coarse-grained material of the column.

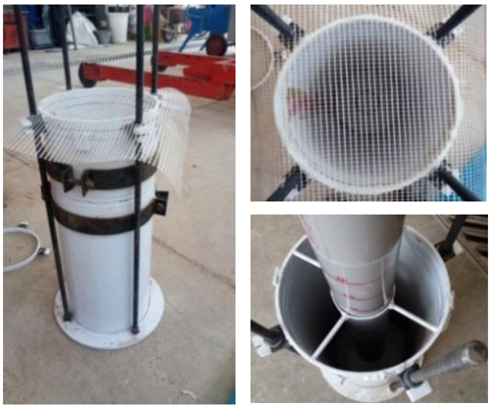

(a) (b)

Figure 2. Unit cell: (a) unit cell head modular rings and ring bracings, (b) PVC pipe supporting grid and method of reinforcement installation

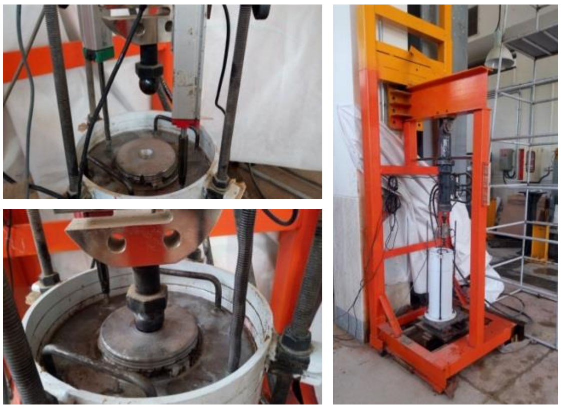

(a) (b)

Figure 3. (a) Load cell hinge connection to footing and LVDT installation, (b) loading frame and jack

The weight of sand materials required to fill the tank in 50 mm thick layers of test bed was determined with the known unit weight of sand. This volume of sand was poured from a shallow 50 mm height at each step until a specific level of the unit cell to prepare a uniform sand bed of the desired relative density. Similarly, the stone column was constructed by dividing its height into equal parts of 50 mm. At each step, a particular weight of self-compacting aggregate material was poured into the PVC pipe. The filled depth of the pipe was measured at each step to monitor the proper relative density. If compaction has been required, mild tapping with a wooden tamper has been performed. No steel rod was used for the compaction of stone column materials due to the crushing of stone grains caused by the impact. The PVC pipe was slowly pulled out every 50 mm, according to the execution of each layer of the sand bed and stone column, so that the bottom of the pipe was always 50 mm in the sand bed and remained buried. In this approach, a 75 mm diameter end-bearing stone column with a length-to-diameter ratio equal to 7 was physically constructed in the centre of the unit cell.

Blankets with thicknesses of 35 mm and 65 mm were also prepared from self-compacting sand grains by pouring materials from a height in layers of 15 mm to 20 mm. Each layer was compacted with a wooden hammer to achieve the desired relative density. On reaching the predetermined depth of the reinforcement layer, the soil surface was levelled, and a reinforcement layer was laid on the sand surface. Finally, drop glue was utilized to attach the reinforcing mesh throughout the perimeter of the unit cell edge. Also, the contact pressure of the upper and lower rings caused by tightening the retaining nuts has helped to restrain the geogrid fully. This geogrid installation method might be regarded as one of the present study's innovations.

Figure 4 displays the models prepared for testing in the unit cell. To prepare all model tests, including the reinforced blanket, a 5 mm layer of granular fill was poured between the sand bed and the geogrid, as reported in the literature geogrid [38]. Vertical spacing between two geogrid layers was 30 mm, with 5 mm of granular material poured on top of the geogrid in models where the geogrid was placed near the top of the blanket.

4. SCALE EFFECTS FOR EXPERIMENTS AND TESTS PROGRAM

The similitude ratio is the ratio of each length size of the prototype model to its physical model equivalent size [18]. Dimensionless variables in the model and prototype must be equal; accordingly, it is feasible to calculate the ratio of the model's values to reality [39]. Based on the Buckingham [40] similitude theory, the ratio of the length scale of the model test to the prototype model is $1/\lambda$, which has been taken as 1/10 in this study. According to the laws of scale [36], the ratio of reinforcement stiffness in the prototype scale (J_p) to the model scale (J_m) could be calculated as $J_p = J_m^2\lambda$ [42]; the same relationship also held in terms of tensile strength.

In earthen constructions, geogrid reinforcements with tensile strengths of more than 400 kN/m and up to 1200 kN/m have often been utilized [43]. Using the average values and according to the laws of similarity, geogrid reinforcement with a tensile strength of 8 kN/m has been used for laboratory model tests. Typically, in the physical model of earthen constructions, there should be infinite soil grains in the contact surfaces of the soil and the structure or the contact surfaces of the soil layers, as well as in the model's boundaries. However, as there are no infinite grains on the contact surfaces of aggregates and the number of grains on these surfaces is finite, the size of the aggregates must be decreased [39]. Thus, the size of the stone column material depends on its diameter. In the current study, the maximum size of the aggregates has been taken as 9.5 mm, while the stone column diameter was 75 mm. In most projects, the diameter of the stone column ranges from 60 cm to 120 cm. Since the stone column diameter in the laboratory model tests has been considered equal to 75 mm, the similitude ratio becomes 8 to 16.

The prototype stone column has a length-to-diameter ratio ranging from 5 to 20 [44]; this parameter is equivalent to 7 in this investigation, with a stone column having a length of 525 mm and a diameter of 75 mm. Compared to other laboratory studies with a similar background [17-19, 45], some distinctions could note. In the present study, the bed soil is fine-grained sand. The test tank shape has been changed from a large cube tank to a laboratory-scale unit cell. The stone column's diameter has been extended to 75 mm, and the number of geogrid layers within the blanket has increased to two layers. Extensive studies have been carried out on geosynthetic-reinforced soil systems [46-48]. However, there is no unified understanding of the failure mode of reinforcement, and few experimental investigations have been conducted on this topic [49].

In the present study, the geogrid installation mode has been changed from applying with an optimal length and

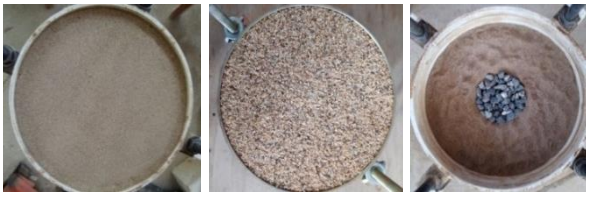

(a) (b) (c)

Figure 4. (a) sand bed, (b) sand bed and granular blanket, (c) sand bed with stone column a free end to an utterly restrained connection. As such, the geogrid's tensile strength is fully mobilized until it fails, significantly improving the load-carrying capacity and decreasing settlement. According to one of the study's principal purposes, the loading was continued once the geogrid failed. It continued until the desired settlement of 20 mm was attained, and the load-settlement characteristics of models involving the reinforced blanket were recorded. The performed tests are presented in Table 3, and abbreviations are according to the general plan developed for the investigation. According to the types of tests mentioned in Table 3, the flowchart of Figure 5 shows the research methodology. It is observed that considerable studies have been conducted to study the effectiveness of geosynthetic reinforcement on load-carrying capacity. As compared, limited experimental investigations have been conducted on reducing settlement [49]. The points noted are also seen in the reinforced blanket used to improve the performance of stone columns.

When the laboratory study of the improved soil with the geosynthetics-reinforced blanket with free ends is carried out in cubic tanks, further experiments should develop to determine the optimum reinforcement size. However, there has been no need for studies to identify the appropriate diameter of the geogrid in the current study because of the new method of connecting the geogrid sheets to the edges of the unit cell. One of the most significant challenges in confirming the accuracy of results in laboratory investigations is reproducibility. Hence, some tests were repeated to validate the findings. Inaccuracies can cause potential mistakes in material weighing and non-uniformity in the constructed test bed, stone column, or blanket.

5. EXPERIMENTAL OBSERVATIONS

5. 1. Effect of Unreinforced Blanket and Stone Column

The results of loading on several model tests until they settled 20 mm, as reported in the literature [17, 18], have been reported in this section.

In addition, because of the laboratory limitations and increasing the internal pressure in the tank while applying load on the model tests, the settlement value of 20 mm was chosen.

The load-settlement characteristics of the unimproved loose sand bed, loose sand improved by only stone column, loose sand along with 35 mm and 65 mm thick unreinforced blanket, and a combination of these scaled physical models are depicted in Figure 6. The settlement given is the average of two LVDTs placed at diametrically opposite ends on the footing. Because of the application of load on the entire surface of the model and the impossibility of lateral soil deformation, the sandy soil was gradually densified upon increasing the overburden pressure. As a result, its stiffness slowly increased, and the models behaved similarly to hydrostatic loading conditions. The model loaded in the unit cell with a rigid plate can be compared to the one-dimensional consolidation test.

TABLE 3. Summary of the experimental tests

Tests series	Type of tests	Test name
1	Sand bed (without stone column and blanket)	SB
2	Sand bed with a 35 mm and 65 mm thick unreinforced blanket	SB+UB35 SB+UB65
3	Sand bed with a layer reinforcement at the bottom or near the top of the 35 mm thick blanket	SB+1bRB35 SB+1tRB35
4	Sand bed with two-layer reinforcement at the bottom and middle of the 65 mm thick blanket	SB+2b&mRB65
5	Sand bed with two-layer reinforcement at the middle and near the top of the 65 mm thick blanket	SB+2m&tRB65
6	Sand bed with stone column	SB+SC
7	Sand bed with stone column and a 35 mm and 65 mm thick unreinforced blanket	SB+SC+UB35 SB+SC+UB65
8	Sand bed with stone column and a layer reinforcement at the bottom or near the top of the 35 mm thick blanket	SB+SC+1bRB35 SB+SC+1tRB35
9	Sand bed with stone column and two-layer reinforcement at the bottom and middle of the 65 mm thick blanket	SB+SC+2b&mRB65
10	Sand bed with stone column and two-layer reinforcement at the middle and near the top of the 65 mm thick blanket	SB+SC+2m&tRB65

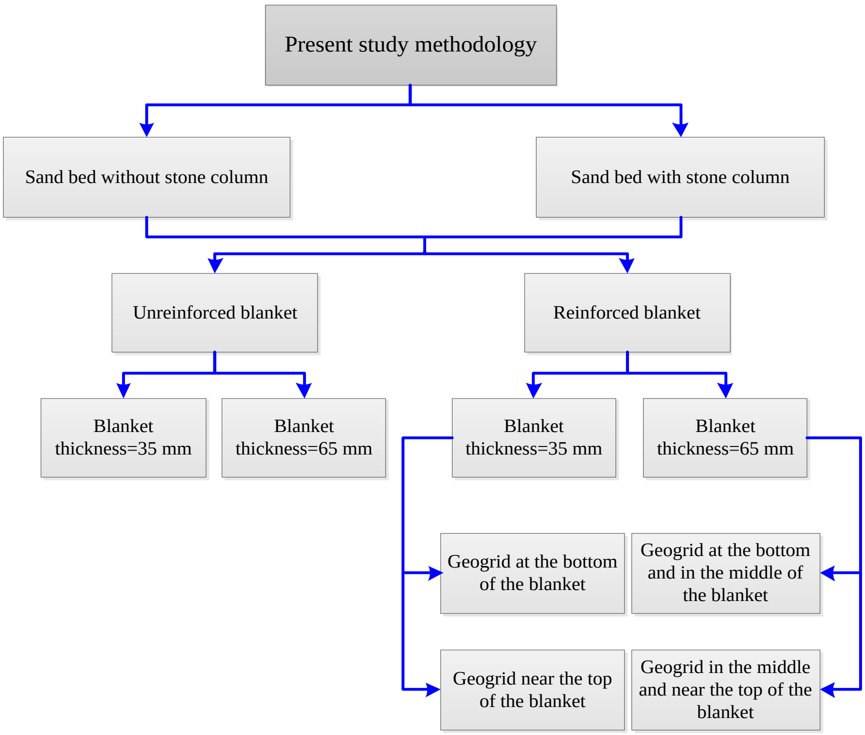

Present study methodology
Sand bed without stone column
Sand bed with stone column
Unreinforced blanket
Reinforced blanket
Blanket thickness=35 mm
Blanket thickness=65 mm
Blanket thickness=35 mm
Blanket thickness=65 mm
Geogrid at the bottom of the blanket
Geogrid at the bottom and in the middle of the blanket
Geogrid near the top of the blanket
Geogrid in the middle and near the top of the blanket

Figure 5. Research methodology of laboratory model tests

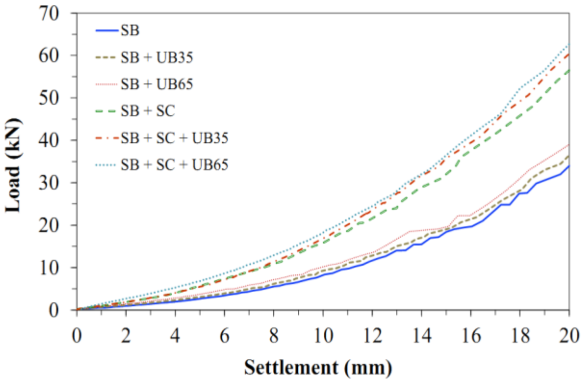


Figure 6. Load-settlement characteristics of loose sand, unreinforced granular blanket, and stone column model tests

In this kind of experiment, since the loading is in the stress path line K_0, failure does not occur in terms of bearing capacity [16]. Soils subjected to hydrostatic loading exhibit nonlinear behaviour [50]. Hardening behaviour in load-settlement characteristics has rarely been reported in laboratory studies conducted by researchers in the unit cell under rigid loading on the entire model surface.

Similar behaviour has been observed only in the laboratory modelling undertaken by Gniel and Bouazza [30], which considers unit cell idealization on a saturated clay bed improved with a geogrid-encased stone column. However, an approximate similar behaviour in laboratory models loaded in the unit cell on saturated clay improved with the stone column by Ambily and Gandhi [51] can be seen. Figure 6 shows the load rises with the settlement and the chart deviation of models without stone columns from those with columns increases. The slope of the load-settlement charts becomes steeper while the presence of a stone column. Based on a comparison of the load-carrying capacity of the models at a constant settlement value, it can be said the effectiveness of the improvement methods is more considerable under looser bed soil conditions. For example, in the case of a 10 mm settlement, the stone column enhances the load-carrying capacity by 92%. While placing 35 mm and 65 mm thick unreinforced blankets on top of the stone column and circumferential soil boosts the load-carrying capacity up to 105% and 122%, respectively. In the 20 mm settlement, the load-carrying capacity of the sand bed with a stone column rises by 66%. In contrast, combining the stone column and unreinforced blanket with the given thicknesses improves the load-carrying capacity by 78% and 84%, respectively.

According to Deb et al. [17], the load-carrying capacity of a soft clay bed with an end-bearing stone column was improved by 69%. Again, it is noted that the load-carrying capacity of a sand bed with an optimum thickness of 50 mm (0.5 times the diameter of the footing) over a stone column-improved soft soil was grown by 141%. The findings of their study are related to a 20 mm settlement and the presence of a single stone column in a large cubic tank. Debnath and Dey [18] observed that the floating stone column group augmented the load-carrying capacity of the improved clay bed by 172%. They also reported a 363% rise in load-carrying capacity in settlement of 20 mm, where a sand bed with the optimal thickness of 40 mm (0.2 times the diameter of the footing) was positioned over the geotextile-encased floating stone columns group.

It can be said, compared to the stone column, the usage of the unreinforced blanket has a far lower effect on the improvement rate, especially when the thickness of the blanket is lower and roughly half the diameter of the stone column. While compared to an unreinforced granular blanket placed on the surface of loose soil, an end-bearing stone column will be more able to carry the load and reduce settlement. The effect of the stone column and the unreinforced blanket in enhancing the load-carrying capacity diminishes as the sand bed becomes gradually dense in the unit cell. The stone column causes a considerable role in decreasing settlement, whereas the unreinforced blanket has a minor effect. Exampling, at a loading intensity of 20 kN, the extent of settlement reduction of the model improved with a stone column reaching 30%. In contrast, with 35 mm or 65 mm thick unreinforced blankets positioned over the stone column, the settlement drops 32% or 35%, respectively. The percentage of settlement reduction under 34 kN loading intensity for models improved with a stone column alone, a stone column along with a 35 mm thick unreinforced blanket, and a stone column along with a 65 mm thick unreinforced blanket is estimated to be around 23%, 26%, and 28%, respectively. Deb et al. [17] reported that for a loading intensity of 0.5 kN, compared to unimproved soil, the settlement has been reduced by 67% and 91%; when the soil is improved by only stone column and by stone column along with unreinforced, respectively.

It suggests that the effect of stone columns and unreinforced blankets in reducing settlement is declined due to the sand bed's gradual densification while loading and its hardening behaviour. Based on the points noted, the role of the unreinforced granular blanket and the stone column in enhancing the load-carrying capacity is more significant than reducing settlement.

5. 2. Effect of Geogrid-reinforced Blanket

Several studies about the effect of geosynthetic reinforcement on soil foundation improvement have applied the reinforcement with the free end and the optimum length. Based on a literature review undertaken by Guo et al. [49], it is 4 to 5 times the width of the foundation. The optimum reinforcement length is affected by the number of reinforcing layers, density and type of soil [52]. As the geogrid has been installed and restrained in the current study, experiments have not been required to identify the appropriate size. Thus, blankets with thicknesses of 35 mm and 65 mm were reinforced with one and two layers of geogrid. The 35 mm thick blankets were reinforced with a single layer of geogrid at the bottom or near the top of the blanket, while blankets with a thickness of 65 mm were reinforced with two layers of geogrid at the bottom and middle or middle and near the top. A geogrid layer was applied at the bottom of the granular fill over stone column-improved soft soil in the studies by Deb et al. [15-17] and Debnath and Dey [18]. Mehrannia et al. [19] used single and double-layers geogrid reinforcement within the middle of the blanket but did not explain how the two layers were arranged relative to each other. Hamidi and Lajevardi [45] reinforced the granular mattress with a geogrid layer at the bottom or middle of it. Figure 7 displays the ruptured geogrid after emptying the unit cell from the granular material of the blanket at the end of the experiment.

It can be seen the geogrid has been ruptured throughout the inner perimeter of the tank. A similar rupturing mechanism has been observed in all blankets, including one or two rows of geogrid reinforcement.

The following section examines the influence of some factors on load-carrying capacity and settlement variation by comparing the load-settlement characteristics under various improving conditions. Figures 8 and 9 illustrate the load-settlement features of the unimproved sand bed and the sand bed, along with the 35 mm and 65 mm thick blankets reinforced with geogrid. Figure 9 indicates the cases where the loose sand bed has also included a stone

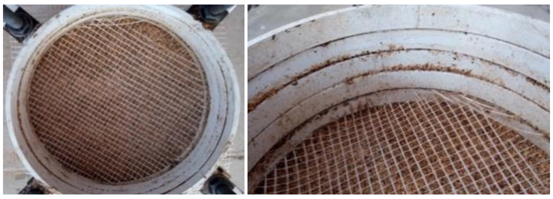


Figure 7. A ruptured sample of the blanket's geogrid reinforcement at the end of the test

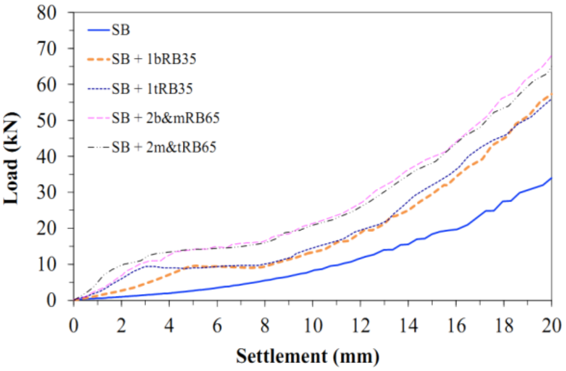


Figure 8. Load-settlement characteristics of sand bed without stone column having a geogrid-reinforced granular blanket

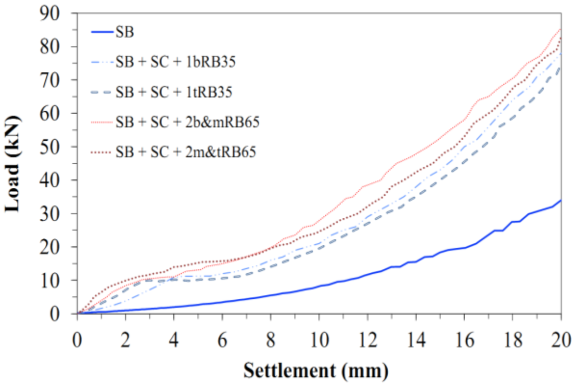


Figure 9. Load-settlement characteristics of sand bed including stone column and geogrid-reinforced granular blanket

column. The charts reveal that reinforcing the blanket with geogrid significantly boosted the load-carrying capacity and reduced the settlement of the model tests. First, the slope of the load-settlement graphs increases until reaching a certain value; then, it becomes nearly constant within a range of the chart, after which the gradient rises again. As compared to unreinforced models, the inclusion of geogrid in the blanket alters the charts' shape and slope.

In addition, a noticeable prominence in load-settlement features and a change of direction of chart concavity at the threshold of geogrid rupture in the settlement ranging from 1-5 mm is observed. The shift in concavity direction and varying the slope from ascending to constant trend are related to the yielding of geogrid.

With the continuance of loading, geogrid rupture ultimately, and with the gradual process of sand densification, the chart returns to its ascending mode.

Increasing the number of reinforcing layers helps to increase load-carrying capacity further and reduce settlement more. The charts in Figures 8 and 9 show that the load-settlement characteristics would be somewhat different with the inclusion of two layers of geogrid reinforcement. During the load enhancement process, two stages of slope variation and concavity direction change are observed when two geogrid layers are placed in the blanket. The first prominence is related to the failure of the first layer of geogrid reinforcement, followed by the failure of the second layer, which forms another prominence. There have been no reports of changes in the slope and direction of the concavity of the load-settlement characteristics in investigations of reinforced blankets with sheet geosynthetic reinforcement. These changes are caused by the way reinforcement operations and their failures. In the studies of Chen et al. [37], the rupture of sheet reinforcement layers under the foundation was reported, and the change in the form of a load-settlement curve was observed.

The comparison of Figure 6 with Figures 8 and 9 reveal that despite the geogrid rupture, the final load value at the settlement of 20 mm has grown compared to the case where the geogrid was not used. In addition, it can be said because of the sandy soil's hardening characteristic; its densification has been possible under the conditions causing tension in the geogrid. Under conditions with a stone column, the geogrid rupture at a higher intensity of load and less settlement due to the stiffer bed caused by the presence of a stone column. Models with reinforced blankets have similar load-settlement characteristics in the geogrid rupture range, regardless of whether stone columns are included. All models with a layer of geogrid near the top of the blanket have load-settlement characteristics with steeper slopes and less settlement at the same load extent compared to the model with geogrid at the bottom. Similar findings have been observed while using two geogrid layers in the middle and near the top of the blanket, compared to placing the geogrid in the bottom and middle of the blanket. While the overburden pressure over the model developed, loose sand hardened, and its density and strength grew as the settlement increased. Therefore, the effect of all improving methods for reducing the settlement diminishes as the load-settlement curves grow gradually. The reduction in settlement following the failure of the geogrid reinforcement has a considerable drawdown in the models, including reinforced blankets. For example, the settlement of the model with a 35 mm thick reinforced blanket, including a layer of geogrid at its bottom resting on the stone column-improved sand bed for a loadings intensity of 5 kN, 15 kN, and 25 kN is reduced by 69%, 44%, and 38%, respectively. However, when the geogrid reinforcement is placed near the top of the blanket, with the given loads, the settlement decreases by 80%, 38%, and 35%, respectively. The comparison suggests that the drawdown in settlement reduction following the geogrid rupture is more severe in the model tests with a single layer of geogrid near the top of the blanket. It is also observed for reinforced blankets, including two geogrid layers in the middle and near the top.

Based on the investigation of Deb et al. [17], for a loading intensity of 1.0 kN, as compared to an unreinforced sand bed, a 44% reduction in the settlement has been observed when the geogrid-reinforced sand bed is used, whereas, for a loading intensity of 1.3 kN, the settlement reduction is 55%. They resulted that the geogrid reinforcement is more effective for higher loading intensity than for lower loading intensity.

As remarked in the introduction, Deb et al. [16] investigated mechanical models with multi-layer reinforced granular fill. They concluded that, compared to single-layer reinforcement, a granular fill reinforced with multi-layer geosynthetic had less effect on reducing settlement since a significant reduction in the settlement was related to the stone column. In addition, they discovered that when stone columns have not been used, the multi-layer reinforcement curtails the settlement. The results of the present study show that at the geogrid rupture threshold, the settlement reduction is relative and depends on the place of reinforcing layers. For example, as compared the model having a layer of geogrid near the top of the blanket (35 mm thick) over a stone column-improved sand bed, with employing two layers of geogrid at the bottom and middle of the blanket (65 mm thick) over the stone column-improved sand bed, the settlement drops by 40% more. When two geogrid layers are placed in the middle, and near the top of the blanket, settlement reduction grows by 60%. Compared to a 35 mm thick blanket reinforced with a geogrid layer at the bottom, the extent of settlement reduction with the placement of two layers at the bottom and middle of the 65 mm thick blanket grows by up to 63%. By placing two layers in the middle and near the top of the blanket, settlement is lower by up to 75%. Thus, the number of reinforcing layers and places will affect the settlement reduction.

Furthermore, the final load-carrying capacity of the model tests at 20 mm settlement has been compared. In the case of using one row of geogrid at the bottom of the blanket or two rows in the middle and bottom of the blanket, up to 5% higher load-carrying capacity has been observed compared to placing a single layer of geogrid near the top of the blanket or two layers in the middle and near the top of the blanket. Due to the distance of the geogrid from the bottom of the footing, the reinforcement is ruptured in a more amount of settlement. As a result, the bed soil has reached a higher density, and the load-carrying capacity has increased. Also, while the presence of a stone column, the final load-carrying capacity grows up to 38% in models having reinforced blankets with one row of geogrid. It is up to 28% in models with reinforced blankets, including two geogrid layers, compared to similar models without stone columns.

5. 3. Improved Load Ratio

The load ratio parameter [42] is derived by dividing the improved sand bed load-carrying capacity (with a blanket, stone column, or a combination of both methods) by the sand bed load-carrying capacity without improvement. This parameter, known as "LR", is related to the improved and unimproved models' load-carrying capacity in an equal settlement. In addition, the settlement ratio parameter (S/D), which is by dividing the footing settlement by the diameter of the footing, can be defined. Therefore, the preceding charts can be generated in different spaces when the axes are dimensionless, as illustrated in Figures 10 and 11. The LR curve related to models without reinforced blankets peaks and then drops with a mild downward trend that the inclusion of blankets or the presence of stone columns causes the maximum load ratio (LR_{max}). Fine-grained sand bed compressibility is higher than coarser-grained materials used in stone columns and blankets.

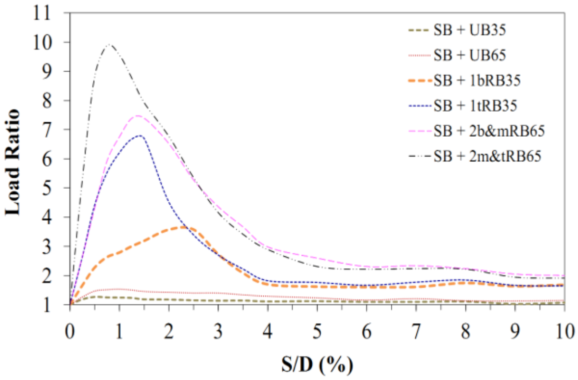


Figure 10. Load ratio-settlement ratio characteristics of improved model tests with an unreinforced and geogrid-reinforced granular blanket

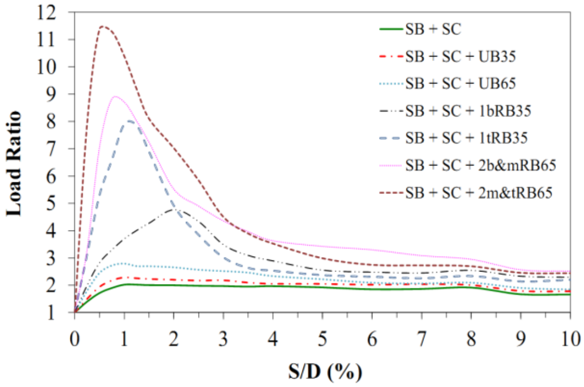


Figure 11. Load ratio-settlement ratio characteristics of improved model tests with stone column and granular blanket

The range of variations in the dry unit weight of these materials confirms this. However, adding a blanket or the presence of a stone column changed the stress distribution and affected the sand's hardening behaviour to some extent. Therefore, reducing the amount of stress in the depth of the improved sand bed models can be attributed to the fact that the stone column carries a significant share of the vertical stress. However, there is also the potential for relative displacement of stone column aggregates under pressure. In addition, the granular blanket's performance on the carriage of some overburden pressure has also affected the sand bed's hardening behaviour.

The load ratio-settlement ratio characteristics for the models with reinforced blankets reveal a prominent peak. These noticeable peaks are caused by the geogrid's tensile strength mobilization, followed by a sudden drop yielded by the geogrid's rupture. After the failure of the reinforcement layers, the resistance was only generated by sand and aggregate materials, which explains the sudden drop in LR variations. The mobilization of the tensile strength of the geogrid reinforcement lay within 0.5-2.5% of the settlement ratio. In the model tests with a reinforced blanket including two layers of geogrid, the LR increases with the settlement ratio, then drops suddenly after the prominent peak point. It indicates that all reinforcement layers ruptured within a relatively short period. As Figure 11 points, the LR_{max} is enhanced to 4.77 in model tests with a layer of geogrid at the bottom of the reinforced blanket with a thickness of 35 mm resting on the stone column-improved sand bed. In this model, when the geogrid reinforcement is placed near the top of the blanket, the LR_{max} grow to 7.9. It is while the settlement has been reduced from 4 mm to 2.5 mm. In other words, altering the geogrid's position from the bottom to near the top of the blanket results in 66% further growth of LR_{max} and 38% less settlement. Therefore, placing the geogrid near the top of the blanket is significantly boosted load-carrying capacity and is reduced settlement; thus, it could be regarded as the optimum place for a layer of geogrid reinforcement.

Upon adding the stone column to the model with layer(s) of geogrid reinforcement, the growth of the load ratio increased further. Similar to using a single layer of geogrid, when two layers of the geogrid move away from the base of the footing while getting closer to the top of the stone column, the effect of the column in enhancing the load-bearing and reducing the settlement is intensified. Although, placing two geogrid reinforcement layers in the middle and near the top of the blanket is the optimal arrangement. With the presence of a stone column and the blanket reinforced with two layers of geogrid in the middle and near the top of the blanket, the maximum value of LR_{max} has been obtained equal to 11.38.

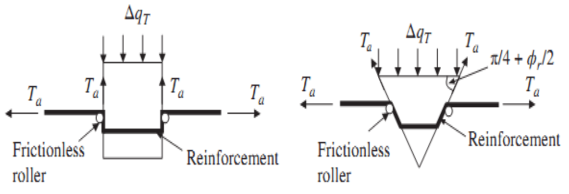
Mehrannia et al. [19] reported that at a settlement of 50 mm, the bearing capacity rose by 85% and 92%, respectively, for the model including a layer of geogrid in the middle of the blanket with a thickness of 75 mm over the clay bed as well as for a clay bed model having floating stone column along with a similar blanket. Moreover, in the research of Deb et al. [17], the maximum enhancement in the load-carrying capacity of geogrid-reinforced sand bed over stone column-improved soft clay was reported as 233%; such a condition that the sand bed had an optimum thickness of 30 mm (0.3 times the diameter of the footing), which included a geogrid layer at the bottom. Debnath and Dey [18] obtained 8.45 times the load-carrying capacity with the geogrid-reinforced sand bed over the geotextile-encased stone column group floating in a soft clay bed; the geogrid reinforcement has been placed at the sand bed's bottom with a 30 mm optimum thickness (0.15 times the diameter of the footing).

According to Figures 11 and 12, the LR_{max} could be derived within the geogrid rupture range and the LR_{final} at the end of loading of model tests. The main differences between LR_{max} and LR_{final} can be summarized as follows:

- LR_{final} values are lower than LR_{max} values in all model tests.
- The model tests that improved with the stone column show further LR_{final} compared to experiments without the stone column but with the reinforced blanket.
- The difference between LR_{max} and LR_{final} in models without the reinforced blanket ranges from 19%-52%, whereas the variation in models with the reinforced blanket is between 108%-417%.
- The difference between LR_{max} and LR_{final} for models with the reinforced blanket and the stone column is less than that of similar ones without the stone column.
- The models with one row of geogrid near the top of the blanket have a higher LR_{max} than the one with a stone column and one row of geogrid at the bottom.
- The LR_{max} of models with a single layer of geogrid near the top of the blanket is 1.5 to 2 times that of models with one geogrid layer at the bottom, but it is not valid for LR_{final}.
- The LR_{final} for models improved by the stone column alone and the models improved with one geogrid layer reinforced blanket without stone column are almost the same, which differs from the LR_{max}.
- The models with stone column and unreinforced blanket have a higher LR_{final} than those without stone column but with one geogrid layer reinforced blanket, which is the inverse of LR_{max}.
- The models with a stone column and reinforced blanket with a single layer of geogrid have a higher LR_{final} than models with the reinforced blanket including two geogrid layers, which is the inverse of LR_{max}.
- In the final load, improving the sand bed with the stone column and the unreinforced blanket is a better alternative than improving the bed only with a reinforced blanket. In addition, it can be said utilizing the stone column along with the reinforced blanket is a more suitable alternative than employing each of these techniques alone.

6. DISCUSSIONS

The blanket material and geogrid reinforcement in the reinforced zone move downward when the footing settles under the applied load. However, since the geogrid reinforcement under the footing is curved, an upward force is mobilized to resist the applied load, increasing the load-carrying capacity [53, 54]. This force is one of the main reinforcing mechanisms with horizontal geosynthetic layers, known as the membrane effect. As illustrated in Figure 12, Das [55], Wayne et al. [56], and Chen [57] presented complete reinforcement rotation to model the membrane tensioned effect. Chen [57]



(a) (b)

Figure 12. Complete rotation of geosynthetic: (a) vertical punching, (b) active triangular wedges (modified from Das [55], Wayne et al. [56], Chen [57])

attributed the contribution of geosynthetic reinforcement to providing lateral confinement to the punching wedge.

The effect of lateral confinement could be noted among other geosynthetic reinforcing mechanisms. It is related to the relative movement of soil grains along the surface of the geogrid reinforcement under the foundation load, which mobilizes the frictional force at the reinforcement-soil interface. The interaction between the geogrid reinforcement and the soil effectively limits the soil grains' horizontal movement, increasing the soil's lateral confining stress and compressive strength beneath the foundation [54, 58]. Based on the method of installing the geogrid reinforcement and conditions of restraining its edges in the present study, there seems to be only the possibility of relative movement of soil grains and geogrid support under the conditions of developing strain in the geogrid during loading. Therefore, it can be said the membrane tension effect has dominated the development of lateral confinement in these types of tests. According to Giroud and Han [59], the influence of the membrane effect becomes increasingly significant with large deformations. When geogrid gets closer to the base of the footing, further reinforcement deformation occurs; hence the development of the membrane effect increases.

Numerical studies of Debnath and Dey [18] conducted in the 3D software ABAQUS 6.12 confirm this. They reported that most geogrid deformations and stresses occurred mainly in the area immediately below the footing, with small deformations away from the loaded area.

Given the restraint of the geogrid at the unit cell's edges and the conditions for its rupture, it is feasible to infer the full participation of the membrane tension effect and reinforcing tensile strength in enhancing the load-carrying capacity and reducing the settlement. When the geogrid is placed near the top of the blanket, more curvature occurs on the surface of the geogrid reinforcement under the footing, mobilizing the membrane effect and increasing the contribution of its tensile strength. Under these conditions, the vertical component of the geogrid's tensile strength somewhat balances the upper loads on the reinforcement. In response to the combined effect of tensile mobilization strength and the geogrid reinforcement membrane effect due to its curvature, vertical stress diminishes in the region under the geogrid [17, 18, 60-62].

Placing the geogrid at the bottom of the blanket causes a reduction of curvature of geogrid reinforcement under the applied load as well as the reduction of both membrane effect contribution and tensile strength mobilization [63]. Therefore, it has reduced the effectiveness of reinforcement, resulting in a further transferred part of the load being to the stone column. In this condition, the stone column is more involved in carrying the load and reducing the settlement. Also, when geogrid reinforcement is further away from the load, it ruptures at a higher footing settlement. But the model's further settlement will correspond to more densification of the sand bed.

Also, since the load ratio parameter is calculated by dividing the improved model's load-carrying capacity by the model without improvement in the same settlement, if the geogrid fails at more amount of settlements blanket, the LR_{max} would be lower. This issue reveals the benefit of placement of a single layer or two layers of geogrid reinforcement closer to the top of the blanket in models without and with stone columns.

The shallow failure is more likely to happen when the spacing above the uppermost reinforcement is greater than 2/3 times the width of the footing, according to Binquet and Lee [53]. Mandel and Salencon [64] developed a solution for a footing on sand bounded by a rigid base. Figure 13 presents footings with finite width and illustrates that the bearing capacity ratio (BCR) grows as the sand friction angle increases. However, when the distance of the uppermost reinforcing row (z_u) from the base of the footing with width B_f increases, the bearing capacity ratio approaches one. The results of three laboratory investigations are consistent with these curves [65]. According to these curves, the bearing

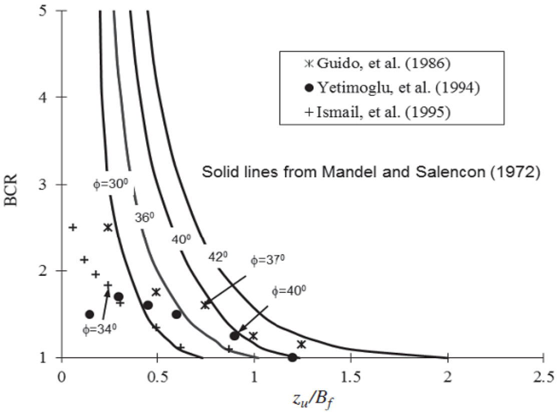


Figure 13. Bearing capacity ratio due to shallow failure above the uppermost reinforcement (after Wayne et al. [56], with permission from ASCE)

capacity ratio grows as the spacing between the reinforcement and the base of the footing diminishes, mainly when the friction angle and density of the soil are low. When z_u is minimal, the overburden pressure related to the shallow footing on the uppermost reinforcing layer is low, and the reinforcement's pullout capacity is limited. Under such a condition, the slip surface extends below the uppermost reinforcement [65]. The current laboratory study is related to the unit cell and the simulation of the centre part of the soil from a wide loading area, which is different from the condition of applying a load via a finite-width foundation. As such, there is no reason for concern about the low overburden pressure on the uppermost reinforcing layer.

Thus, in response to the overall outcome of these curves, if the load is applied over a large area or through multiple adjacent footings, the load-bearing capacity growth will be more remarkable where the geogrid reinforcement is closer to the base of the footing. In addition, an increase in normal stress would increase the shear strength at the contact surface between soil and geogrid [66, 67]. The conditions of the side stone columns differ from those of the others in an infinite group, and the unit cell assumption is unrealistic for them. As a result, the findings of this study cannot be generalized to side stone columns or stone columns of a small group.

7. SUMMARY AND CONCLUSIONS

In this study, the load-settlement characteristic of sand bed models improved with unreinforced and geogrid-reinforced granular blankets, the end-bearing stone column, and with the combination of these methods investigated through large-scale laboratory model tests. The unit cell was used in this study to simulate the behaviour of a single stone column in an infinite group of stone columns. The thickness of the blankets has been taken as 35 mm and 65 mm, and the stone column was 75 mm in diameter with a length-to-diameter ratio equal to 7. A new approach was utilized to install the geogrid, which allows complete mobilizing of the tensile strength and rupture of the geogrid reinforcement.

The role of this mechanism on the load-carrying capacity and settlement characteristic of the physical models was identified. It should be noted that the geogrid reinforcement rupture mechanism has not been investigated earlier in reinforced blanket studies; thus, the findings of this research can be applied in practice. The following are the most prominent conclusions from the current laboratory study:

- As compared to the stone column, the unreinforced granular blanket had a far lower effect on enhancing the load-carrying capacity and reducing settlement. It can be said using a stone column, granular blanket, or combination of both techniques to boost load-carrying capacity was more effective than reducing settlement. However, when the sand bed gradually densified under loading, the effect of the stone column and granular blanket on increasing the load-carrying capacity and reducing settlement was diminished. In addition, the efficiency of improvement methods has been superior under looser bed conditions.
- The results indicate that including geogrid reinforcement in the blanket significantly improves the load-carrying capacity and reduces the settlement of all model tests. However, the effect of single-layer and double-layer geogrid reinforcement on settlement reduction depends on their placement within the granular blanket.
- The comparison of reinforcement layouts of the reinforced blanket with the geogrid indicates that when the geogrid is closer to the base of the footing, it will play a more effective role in enhancing the load-carrying capacity and decreasing the settlement. In models with reinforced blankets, the extent of reduction in the settlement after the rupture of the geogrid reinforcement has a significant drop.
- In models with stone columns causing stiffer beds, the geogrid reinforcement ruptured under more loading intensity and at less extent of settlement. Regardless of the number of reinforcing layers, the stone column significantly improves the LR_{max} and reduces settlement in models with a geogrid layer at the bottom of the blanket.
- In the final load, improving the sand bed with stone columns and unreinforced blankets is preferred over improving the sand bed with only reinforced blankets. Overall, the combination approach of the stone column and reinforced blanket is a preferable alternative rather than using either of these techniques individually.

8. REFERENCES

1. Abbas, H. O., “Laboratory study on reinforced expansive soil with granular pile anchors”, ***International Journal of Engineering, Transactions A: Basics***, Vol. 33, No. 7, (2020), 1167-1172. 10.5829/ije.2020.33.07a.01

2. Umravia, N. B. and Solanki, C. H., “Numerical analysis to study lateral behavior of cement fly ash gravel piles under the soft soil”, ***International Journal of Engineering, Transactions B: Applications***, Vol. 35, No. 11, (2022), 2111-2119. 10.5829/ije.2022.35.11b.06

3. Chandiwala, A. and Vasanwala, S., “Experimental study of lateral loading on piled raft foundations on sandy soil”, ***International Journal of Engineering, Transactions A: Basics***, Vol. 36, No. 1, (2023), 28-34. 10.5829/ije.2023.36.01a.04

4. Russo, G., Marone, G. and Girolamo, L. D., “Hybrid energy piles as a smart and sustainable foundation”, ***Journal of Human, Earth, and Future***, Vol. 2, No. 3, (2021), 306-322. http://dx.doi.org/10. 28991/HEF-2021-02-03-010

5. Vali, R., "Water table effects on the behaviors of the reinforced marine soil-footing system", ***Journal of Human, Earth, and Future***, Vol. 2, No. 3, (2021), 296-305. http://dx.doi.org/10.28991/HEF-2021-02-03-09

6. Farah, R. E. and Nalbantoglu, Z., "Behavior of geotextile-encased single stone column in soft soils", ***Arabian Journal of Science and Engineering***, Vol. 45, (2020), 3877-3890. https://doi.org/10.1007/s13369-019-04299-3

7. Hataf, N., Nabipour, N. and Sadr, A., "Experimental and numerical study on the bearing capacity of encased stone columns", ***International Journal of Geo-Engineering***, Vol. 11, No. 4, (2020), 1-19. https://doi.org/10.1186/s40703-020-00111-6

8. Alkhorshid, N. R., Araujo, G. L. S. and Palmeira, E. M., "Consolidation of soft clay foundation improved by geosynthetic-reinforced granular columns: Numerical evaluation", ***Journal of Rock Mechanics and Geotechnical Engineering***, Vol. 13, No. 5, (2021), 1173-1181. https://doi.org/10.1016/j.jrmge.2021.09.017

9. Bahrami, M. and Marandi, S. M., "Large-scale experimental study on collapsible soil improvement using encased stone columns", ***International Journal of Engineering, Transactions B: Applications***, Vol. 34, No. 5, (2021), 1145-1155. 10.5829/IJE.2021.34.05B.08

10. Akosah, S., Chen, J. and Bao, N., "Reinforcement of problematic soils using geotextile encased stone/sand columns", ***Arabian Journal of Geosciences***, Vol. 15, (2022), 1-21. https://doi.org/10.1007/s12517-022-10561-0

11. Gu, M., Mo, H., Qiu, J., Yuan, J. and Xia, Q., "Behavior of floating stone columns reinforced with geogrid encasement in model tests", ***Frontiers in Materials***, Vol. 9, (2022), 1-10. https://doi.org/10.3389/fmats.2022.980851

12. Kang, B., Wang, J., Zhou, Y. and Huang, Sh., "Study on bearing capacity and failure mode of multi-layer-encased geosynthetic-encased stone column under dynamic and static Loading", ***Sustainability***, Vol. 15, No. 6, (2023), 1-18. https://doi.org/10.3390/su15065205

13. Nazari Afshar, J., Mehrannia, N., Kalantari, F. and Ganjian, N., "Bearing capacity of group of stone columns with granular blankets", ***International Journal of Civil Engineering***, Vol. 17, (2017), 253-263. https://doi.org/10.1007/s40999-017-0271-y

14. Ramadan, E. H., Abdel-Naiem, M. A., Senoon, A. A. and Megally, A. A., "Stone columns and reinforced sand bed for performance improvement of foundations on soft clay", ***International Journal of Advances in Structural and Geotechnical Engineering***, Vol. 6, No. 3, (2022), 57-64. 10.21608/ASGE.2022.274736

15. Deb, K., Basudhar, P. K. and Chandra, S., "Generalized model for geosynthetic-reinforced granular fill-soft soil with stone columns", ***International Journal of Geomechanics***, Vol. 7, No. 4, (2007), 266-276. https://doi.org/10.1061/(ASCE)1532-3641(2007)7:4(266)

16. Deb, K., Chandra, S. and Basudhar, P. K., "Response of multi-layer geosynthetic-reinforced bed resting on soft soil with stone columns", ***Computer and Geotechnics***, Vol. 35, No. 3, (2008), 323-330. https://doi.org/10.1016/j.compgeo.2007.08.004

17. Deb, K., Samadhiya, N. K. and Namdeo, J. B., "Laboratory model studies on unreinforced and geogrid-reinforced sand bed over stone column-improved soft clay", ***Geotextiles and Geomembranes***, Vol. 29, No. 2, (2011), 190-196. https://doi.org/10.1016/j.geotexmem.2010.06.004

18. Debnath, P. and Dey, A. K., "Bearing capacity of geogrid-reinforced sand over encased stone columns in soft clay", ***Geotextiles and Geomembranes***, Vol. 45, No. 6, (2017), 653-664. https://doi.org/10.1016/j.geotexmem.2017.08.006

19. Mehrannia, N., Kalantary, F. and Ganjian, N., "Experimental study on soil improvement with stone columns and granular blankets", ***Journal of Central South University***, Vol. 25, No. 4, (2018), 866-878. https://doi.org/10.1007/s11771-018-3790-z

20. Abdullah, C. H. and Edil, T. B., "Behaviour of geogrid-reinforced load transfer platforms for embankment on rammed aggregate piers", ***Geosynthetics International***, Vol. 14, No. 3, (2007), 141-153. https://doi.org/10.1680/gein.2007.14.3.141

21. Barksdale, R. D. and Bachus, R. C., "Design and construction of stone columns, Federal Highway administration Office of Engineering and Highway Operations Research and Development, FHWA/RD-83/029", School of Civil Engineering, Georgia, Georgia, UAS, (1983).

22. Jamshidi Chenari, R., Karimpour Fard, M., Jamshidi Chenari, M. and Shamsi Sosahab, J., "Physical and numerical modeling of stone column behavior in loose sand", ***International Journal of Civil Engineering***, Vol. 17, (2019), 231-244. https://doi.org/10.1007/s40999-017-0223-6

23. Nayak, N. V., "Recent advances in ground improvements by stone column", Proceedings of Indian Geotechnical Conference, IGC-83, Madras, India, (1983).

24. Fattah, M. Y., Shlash, K. T. and Al-Waily, M. J., "Stress concentration ratio of model stone columns in soft clays", ***Geotechnical Testing Journal***, Vol. 34, No. 1, (2011), 1-11. 10.1520/GTJ103060

25. Fox, Z. P., "Critical state, dilatancy and particle breakage of mine waste rock", PhD Dissertation, Colorado State University, Colorado, USA, (2011).

26. Stoeber, J. N., "Effects of maximum particle size and sample scaling on the mechanical behavior of mine waste rock: A critical state approach", PhD Dissertation, Colorado State University, Colorado, USA, (2012).

27. Mohapatra, S. R., Rajagopal, K. and Sharma, J., "Direct shear tests on geosynthetic-encased granular columns", ***Geotextiles and Geomembranes***, Vol. 44, No. 3, (2016), 396-405. https://doi.org/10.1016/j.geotexmem.2016.01.002

28. Ali, K., Shahu, J. T. and Sharma, K. G., "Model tests on geosynthetic-reinforcement stone columns: a comparative study", ***Geosynthetics International***, Vol. 19, No. 4, (2012), 292-305. https://doi.org/10.1680/gein.12.00016

29. ASTM D6637/D6637M-15. "Standard test method for determining tensile properties of geogrids by the single or multi-rib tensile method", ASTM International, West Conshohocken, PA, USA, (2015).

30. Gniel, J. and Bouazza, A., "Improvement of soft soils using geogrid encased stone columns", ***Geotextiles and Geomembranes***, Vol. 27, No. 3, (2009), 167-175. https://doi.org/10.1016/j.geotexmem.2008.11.001

31. Koerner, R. M., "Designing with geosynthetics", 6th Ed. New Jersey, Prentice Hall, USA, (2005).

32. Murugesan, S. and Rajagopal, K., "Studies on the behavior of single and group of geosynthetic encased stone columns", ***Journal of Geotechnical and Geoenvironmental Engineering***, Vol. 136, No. 1, (2010), 129-139. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000187

33. Ali, K., Shahu, J. T. and Sharma, K. G., "Model tests on single and groups of stone columns with different geosynthetic reinforcement arrangement", ***Geosynthetics International***, Vol. 21, No. 2, (2014), 103-118. https://doi.org/10.1680/gein.14.00002

34. Hasan, M. and Samadhiya, N. K., "Experimental and numerical analysis of geosynthetic-reinforced floating granular piles in soft clays", ***International Journal of Geosynthetics and Ground Engineering***, Vol. 2, No. 3, (2016), 1-13. https://doi.org/10.1007/s40891-016-0062-6

35. Hong, Y. S., Wu, C. S. and Yu, Y. S., "Model tests on geotextile-encased granular columns under 1-g and undrained conditions", ***Geotextiles and Geomembranes***, Vol. 44, No. 1, (2016), 13-27. https://doi.org/10.1016/j.geotexmem.2015.06.006

36. Ghazavi, M., Ehsaniyamchi, A. and Nazari Afshar, J., "Bearing capacity of horizontally layered geosynthetic reinforced stone columns", ***Geotextiles and Geomembranes***, Vol. 46, No. 3, (2018), 312-318. https://doi.org/10.1016/j.geotexmem.2018.01.002

37. Chen, J. F., Guo, X. P., Xue, J. F. and Guo, P. H., "Load behavior of model strip footings on reinforced transparent soils", ***Geosynthetics International***, Vol. 26, No. 3, (2019), 251–260. https://doi.org/10.1680/jgein.19.00003

38. Han, J. and Gabr, M. A., "Numerical analysis of geosynthetic-reinforced and pile-supported earth platform over soft soil", ***Journal of Geotechnical and Geoenvironmental Engineering***, Vol. 128, No 1, (2002), 44-53. https://doi.org/10.1061/(ASCE)1090-0241(2002)128:1(44)

39. Fakher, A., "Research methods in geotechnics", University of Tehran Press, Tehran, Iran, (2014).

40. Buckingham, E., "On physically similar systems; illustrations of the use of dimensional equations", ***Physical Review***, Vol. 4, No. 4, American Physical Society, (1914) 345. https://doi.org/10.1103/PhysRev.4.345

41. Iai, S., "Similitude for shaking table tests on soil-structure fluid models in 1g gravitational field", ***Soils and Foundations***, Vol. 29, No. 1, (1989), 105–118. https://doi.org/10.3208/sandf1972.29.105

42. Ghazavi, M. and Nazari Afshar, J., "Bearing capacity of geosynthetic encased stone columns", ***Geotextiles and Geomembranes***, Vol. 38, (2013), 26-36. https://doi.org/10.1016/j.geotexmem.2013.04.003

43. NAUE GMBH & CO KG., "Naue Products Manual. Espelkamp, Germany", www.naue.com/products, (Status 26 October 2021).

44. Shahu, J. T. and Reddy, Y. R., "Clayey soil reinforced with stone column group: model tests and analyses", Journal of ***Geotechnical and Geoenvironmental Engineering***, Vol. 137, No. 12, (2011), 1265-1274. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000552

45. Hamidi, M. and Lajevardi, S. H., "Experimental study on the load-carrying capacity of single stone columns", ***International Journal of Geosynthetics and Ground Engineering***, Vol. 4, (2018), 1-10. https://doi.org/10.1007/s40891-018-0142-x

46. Arjomand, M. A., Abedi, M., Gharib, M. and Damghani, M., "An experimental study on geogrid with geotextile effects aimed to improve clayey soil", ***International Journal of Engineering, Transactions B: Applications***, Vol. 32, No. 5, (2019), 685-692. 10.5829/ije.2019.32.05b.10

47. Hoseini, M. H., Noorzad, A. and Zamanian, M., "Physical modelling of a strip footing on a geosynthetic reinforced soil wall containing tire shred subjected to monotonic and cyclic loading", ***International Journal of Engineering, Transactions A: Basics***, Vol. 34, No. 10, (2021), 2266-2279. 10.5829/IJE.2021.34.10A.08

48. Sarfarazi, V., Tabaroei, A. and Asgari, K., "Discrete element modeling of strip footing on geogrid-reinforced soil", ***Geomechanics and Engineering***, Vol. 29, No. 4, (2022), 435-449. https://doi.org/10.12989/gae.2022.29.4.435

49. Guo, X., Zhang, H. and Liu, L., "Planar geosynthetic-reinforced soil foundations: a review", ***SN Applied Sciences***, Vol. 2, (2020), 1-18. https://doi.org/10.1007/s42452-020-03930-5

50. Chen, W. F. and Saleeb, A. F., "Constitutive equations for engineering materials", 2nd Revised Ed, Elsevier Science B.V, New York, USA, (1994).

51. Ambily, A. P. and Gandhi, S. R., "Behavior of stone columns based on experimental and FEM analysis", ***Journal of Geotechnical and Geoenvironmental Engineering***, Vol. 133, No. 4, (2007), 405-415. https://doi.org/10.1061/(ASCE)1090-0241(2007)133:4(405)

52. Mosallanezhad, M., Hataf, N. and Ghahramani, A., "Three-dimensional bearing capacity analysis of granular soil, reinforced with innovative grid-anchor system", ***Iranian Journal of Science and Technology, Transactions B: Engineering***, Vol. 34, (2010), 419–431. 10.22099/IJSTC.2012.693

53. Binquet, J. and Lee, K. L., "Bearing capacity tests on reinforced earth slabs", ***Journal of Geotechnical and Geoenvironmental Engineering***, Vol. 101, (1975), 1241–1255. https://doi.org/10.1061/AJGEB6.0000219

54. Chen, Q. M. and Abu-Farsakh, M., "Ultimate bearing capacity analysis of strip footings on reinforced soil foundation", ***Soils and Foundations***. Vol. 55, No. 1, (2015), 74–85. https://doi.org/10.1016/j.sandf.2014.12.006

55. Das, B. M., "Principles of foundation engineering", 4th Ed. PWS Publishing, Boston, USA, (1998).

56. Wayne, M. H., Han, J. and Akins, K., "The design of geosynthetic reinforced foundations", ASCE Geo-Institute Geotechnical Special Publication, ASCE Press, USA, (1998).

57. Chen, Q., "An experimental study on characteristics and behavior of reinforced soil foundation", PhD Dissertation, Louisiana State University, Louisiana, USA, (2007).

58. Fazeli Dehkordi, P., Ghazavi, M. and Karim, U. F. A., "Bearing capacity-relative density behavior of circular footings resting on geocell-reinforced sand", ***European Journal of Environmental and Civil Engineering***, Vol. 26, No. 11, (2021), 5088-5112. https://doi.org/10.1080/19648189.2021.1884901

59. Giroud, J. P. and Han, J., "Design method for geogrid-reinforced unpaved roads, Part I: theoretical development", ***Journal of Geotechnical and Geoenvironmental Engineering***, Vol. 130, No. 8, (2004), 776-786. https://doi.org/10.1061/(ASCE)1090-0241(2004)130:8(775)

60. Basudhar, P. K., Dixit, P. M., Gharpure, A. and Deb, K., "Finite element analysis of geotextile-reinforced sand-bed subjected to strip loading", ***Geotextiles and Geomembranes***, Vol. 26, No. 1, (2008), 91-99. https://doi.org/10.1016/j.geotexmem.2007.04.002

61. Lee, K. M., Manjunath, V. R. and Dewaikar, D. M., "Numerical and model studies of strip footing supported by a reinforced granular fill-soft soil system", ***Canadian Geotechnical Journal***, Vol. 36, No. 5, (1999), 793-806. https://doi.org/10.1139/t99-053

62. Burd, H. J., "Analysis of membrane action in reinforced unpaved roads", ***Canadian Geotechnical Journal***, Vol. 32, No. 6, (1995), 946-956. https://doi.org/10.1139/t95-094

63. Shahu, J. T., Madhav, M. R. and Hayashi, S., "Analysis of soft ground-granular pile-granular mat system", ***Computers and Geotechnics***, Vol. 27, No. 1, (2000), 45-62. https://doi.org/10.1016/S0266-352X(00)00004-5

64. Mandel, J. and Salencon, J., "Force portante d'un sol sur une assise rigide", ***Geotechnique***, Vol. 22, (1972), 79–93. https://doi.org/10.1680/geot.1972.22.1.79

65. Han, J., "Principles and practice of ground improvement", John Wiley & Sons, New Jersey, USA, (2015).

66. Vieira, C. S. and Lopes, M. D. L., "Sand-nonwoven geotextile interfaces shear strength by direct shear and simple shear tests", ***Geomechanics and Engineering***, Vol. 9, No. 5, (2015), 601-618. https://doi.org/10.12989/gae.2015.9.5.601

67. Safa, M., Maleka, A., Arjomand, M. A., Khorami, M. and Shariati, M., "Strain rate effects on soil-geosynthetic interaction in fine-grained soil", ***Geomechanics and Engineering***, Vol . 19, No. 6, (2019), 533-542. https://doi.org/10.12989/gae.2019.19.6.533

Persian Abstract

چکیده

در این پژوهش مطالعه نمونه‌های بزرگ مقیاس آزمایشگاهی در سلول واحد به منظور بررسی رفتار خاک ماسه‌ای سست بهسازی شده با بالشتک دانه‌ای غیر مسلح و مسلح، ستون سنگی اتکایی و ترکیبی از آن‌ها انجام شده است. با توجه به اینکه تا کنون در مطالعات تجربی به گسیختگی مسلح کننده در بالشتک دانه‌ای مسلح پرداخته نشده، روشی نوین جهت نصب ژئوگرید در سلول واحد به کار رفته تا در نتیجه آن امکان بسیج کامل مقاومت کششی ژئوگرید و گسیختگی آن تحت تنش‌های وارده محقق گردد. در این تحقیق رفتار بار- نشست نمونه‌ها حتی پس از گسیختگی ژئوگرید و تا رسیدن به نشست موردنظر ادامه یافته است. تمرکز مطالعات در راستای بررسی تاثیر متغیرهایی چون ضخامت بالشتک و آرایش تسلیح شامل تعداد و محل قرارگیری صفحات ژئوگرید در بالشتک دانه‌ای، طی ساخت مدل‌های فیزیکی بستر ماسه‌ای بدون ستون سنگی و دارای ستون است. مسلح سازی بالشتک با ژئوگرید ضمن متمایز نمودن شکل نمودارهای بار- نشست نسبت به سایر مطالعات با زمینه مشابه، اثر قابل توجهی بر افزایش باربری و کاهش نشست نمونه‌ها داشته است. میزان تاثیر استفاده از ستون سنگی، بالشتک ماسه‌ای یا هر دو روش بر افزایش توان باربری بستر بیشتر از کاهش نشست بوده است. چگونگی تاثیر مسلح‌کننده بالشتک بر کاهش نشست بستر، به محل قرارگیری آن در بالشتک وابسته است. این روش‌های بهسازی در حالتی که خاک بستر در شرایط سست‌تری قرار داشته موثرتر بوده‌اند. می‌توان قرارگیری یک لایه ژئوگرید در بالا یا دو لایه در میانه و بالای ضخامت بالشتک را موقعیت‌های بهینه قرارگیری مسلح‌کننده در ضخامت بالشتک تلقی نمود.

International Journal of Engineering

Journal Homepage: www.ije.ir

Feature Extraction from Several Angular Faces Using a Deep Learning Based Fusion Technique for Face Recognition

E. Charoqdouz, H. Hassanpour*

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

PAPER INFO

Paper history:
Received 19 February 2023
Received in revised form 20 June 2023
Accepted 23 June 2023

Keywords:
Feature Extraction
Angled Face Recognition
Fusion of Feature Vectors
Generative Adversarial Neural Networks

ABSTRACT

Due to its non-interfering nature, face recognition has been the most suitable technology for designing biometric systems in recent years. This technology is used in various industries, such as health care, education, security, and surveillance. Facial recognition technology works best when a person is looking straight into the camera. On the contrary, the performance of facial recognition degrades when encountered with an angled facial image, because they are generally trained using images of a full face. The purpose of this paper is to estimate the feature vector of a full face image when there are several angular facial images of the same person, one example being angular faces in a video. This method extracts the basic features of a facial image using the non-negative matrix factorization (NMF) method. Then, the feature vectors are fused using a generative adversarial network (GAN) to estimate the feature vector associated with the frontal image. The experimental results on the angular images of the FERET dataset show that the proposed method can significantly improve the accuracy of facial recognition technology methods.

***doi**: 10.5829/ije.2023.36.08b.14*

1. INTRODUCTION

Among the biometric techniques to identify and authenticate people, the facial recognition method is widely used because of its many benefits, including simplicity and easy access [1]. Self-driving cars, criminal identification, video surveillance, and building access control are some of the applications of the facial recognition system. Despite many improvements, this system still faces problems such as angle changes, occlusion, lighting, and other factors [2]. Several face recognition methods exist in literature, including Eigenfaces [3], Fisherfaces [4], independent component analysis [5], method based on the analysis of local features [6], hashing in Uncontrolled environment [7], and sparse processing in the recognition of thermal face images [8] which are able to achieve a good result. Many facial recognition methods require the facial image be frontal (full face) to accurately identify the person. In other words, as the angle of the face to the camera increases, the accuracy of face recognition methods decreases. Recently, feature fusion technique [9, 10] has improved the performance of facial recognition systems to some extent. In a facial recognition system, the fusion of information can be done at the decision level or at the feature level [11]. Feature level techniques combine input characteristic sets into fused sets, then use them in a typical classifier, while decision-level techniques combine different classifiers [12, 13]. AL-Shatnawi et al. [14] proposed a face recognition method based on the Laplace Pyramid (LP) fusion technique at the level of fused features. Based on this, key facial features are identified, general features are extracted using Principal Component Analysis (PCA), and local features are extracted using Local Binary Pattern (LBP) method. Finally, using the LP fusion technique, the extracted features are combined, then classified by the artificial neural network classifier. Often the fusion at the decision level is based on the combination of the output scores from the classifiers. The fusion was performed by Štruc et al. [15] based on LBP, Gabor, and pixel scores. Hu et al. [16] used feature-dense SIFT, multi-scale SIFT, and

*Corresponding Author Email: h.hassanpour@shahroodut.ac.ir (H. Hassanpour)

LBP to train a deep neural network. Finally, the features were combined at the score level.

The process of taking multiple angular faces and combining the identification information is followed in this work. This study presents a scheme based on which recognition can be performed, using feature fusion, by receiving several angular faces of the same person. This paper uses the non-negative matrix factorization (NMF) for face feature extraction. The feature vectors of various faces from the same person are fused using a weight vector which is the same size as the feature vector. The weight vectors, which depend to face angle, are obtained using a genetic algorithm, and indicate the significance of each feature in fusion. Finally, the fusion result is fed to a generative adversarial ***network (GAN)*** to appropriately estimate the feature vector of the front face.

Other sections of the article are the following. In Section 2, we review the literature related to facial recognition technology. Section 3 describes the proposed method in detail. Section 4 presents the experimental results of the proposed method. And finally, section 5 contains the conclusion.

2. LITERATURE REVIEW

Shanthi and Nickolas [11] proposed a method for integrating different descriptors in face recognition, which is generally done at two levels: feature level and decision level. By fusing different descriptors, strong descriptors can be obtained. In the feature-level method, the extracted features are fused into a feature vector and sent to a classifier. The advantage of this method is the simplicity of training and exploiting the correlation of multiple features in the early stages. Alternatively, the integration approach can be used at the decision level, where separate classifiers are used to obtain the score of the extracted features, and local decisions are combined to obtain the final decision. The advantage of this method compared to the feature level is the easy possibility of fusing decisions compared to the fusion of features. Jabid et al. [17] proposed a method based on local orientation pattern (LDP) for face recognition system. LDP obtains different values of edge response in all eight directions from each pixel. LDP histograms are generated from multiple blocks that are regularly concatenated into a unified feature vector. In the proposed method, the weighted chi-square criterion is used, which determines different weights in the facial block areas due to the better recognition capacity of facial features such as mouth, eyes, and nose. The performance of the proposed model has been compared with PCA and LBP, and the experimental results show that the proposed method can improve the accuracy of face recognition in aging and light conditions compared to PCA and LBP methods. In the same vein, Al-Dabagh et al. [18] proposed a method based on feature fusion for face recognition. In this method, features are extracted from face images by local binary pattern (LBP) and Gabor and then fused. In the next step, for recognition, distinct features are extracted from the fusion feature vector by Conventional Correlation Analysis (CCA). After that, classification and identification are done using Support Vector Machine (SVM). The experimental results indicate that this method can achieve 97.14% recognition accuracy. Similarly, Liu et al. [19] proposed a face recognition method based on feature fusion which fuses hybrid color space, Gabor, Discrete Cosine Transform (DCT), and local binary patterns (LBP). In this method, the combined color space is obtained by merging the R component from RGB color space, Cr from YCbCr color space, and Q from YIQ color space. The experimental results demonstrate that the proposed method can achieve a recognition accuracy of 92.43%. In a unique study, a new method for hyperspectral face recognition was introduced by Uzair et al. [20] the proposed method uses a band fusion strategy based on spectral-spatial covariance. The fusion algorithm incorporates local spatial information. After obtaining the composite image, Partial Least Square regression (PLS) is used for classification. The experimental results on three standard databases of PolyU, CMU, and UWA demonstrate that the proposed method has been able to improve the accuracy of hyperspectral face recognition in the range of 95.2% to 99.1%. Bi et al. [21] introduced a thermal face recognition method that is based on multi-feature fusion. The proposed approach extracts features from the input image by using Gabor descriptor, LBP, Weber descriptor, and downsampling, which are then fused. The experimental results indicate that the proposed method can achieve a recognition accuracy of 91.5%. Additionally, Zhu et al. [22] proposed a novel face recognition method based on big data. This method extracts global features of the face by using the two-Dimensional Principal Component Analysis (2DPCA) and local features by using the Local Binary Pattern (LBP) algorithm, which are then fused. In the subsequent step, the fusion features are employed as input to the convolutional neural network. Finally, the trained feature vector is utilized for face recognition. The results demonstrate that the proposed method achieves a recognition accuracy of 95%. Wang et al. [23] proposed a method for integrating facial and finger vein biometric features by using a Convolutional Neural Network (CNN). The method utilizes AlexNet and VGG-19 networks for feature extraction. After feature extraction and fusion, a fusion feature vector is obtained. In this method, the fusion feature vector and vein and face features are recombined to prevent information loss and optimize the effective information. Experimental results indicate that the proposed method enhances identification accuracy in both networks by over 98.4%. Medjahed et al. [24] aimed to enhance the performance of unimodal biometric-based security systems by matching face, right and left palm scores. They utilize CNN to extract features from biometric data such as face, right palm and left palm. Following feature extraction, a fusion operation is conducted at the score level. Finally, a K-Nearest Neighbor (KNN) classifier is used for identification. In this method, testing is carried out on both healthy data without noise, data with salt and pepper noise, Gaussian noise, and data rotation with various degrees. Experimental results reveal that the proposed method is more robust to disturbances than the one-way biometric system. Zhang et al. [10] proposed the idea of combining features from each layer of the CNN network for face recognition. Since the operation after the convolution layer in the CNN network is usually nonlinear, some useful features for identification might be lost. Thus, this method extracts shallow, middle, and deep features of the image, which are then fused together through the CNN network. The experimental results indicate that this method has improved the accuracy of face recognition against occlusion. Xu et al. [25] conducted a study based on fusion biometric features using a CNN network. In this method, feature extraction is performed from the face, iris, and palm by using a CNN, and then the extracted feature vectors are integrated. Finally, classification is obtained based on the fusion feature vector. Experimental results on three databases, including CMU PIE, CASIA, and Poly-U, show that the proposed method improved recognition accuracy in the range of 96-97%. Likewise, Almabdy and Elrefaei [26] presented a face recognition method based on a combination of features. In the proposed method, feature extraction is performed by AlexNet and ResNet-50 convolutional neural networks, and the extracted features are combined. In the next step, support vector machine (SVM) is used to classify the fusion feature vector. Different datasets, including FEI, ORL, and LFW have been used for testing. The experimental results show that this method has been able to improve the accuracy of face recognition in the range of 96.21% to 100% by using the combination of features.

Previous studies have utilized feature extraction and fusion techniques to perform face recognition. However, a new approach is proposed in this study for recognizing faces in angled images. The method involves estimating the feature vector of the frontal state of the face by fusing the features extracted from the angled image of the individual. By doing so, this method aims to accurately recognize faces despite the presence of angles in the input image.

3. PROPOSED METHOD

This study utilizes the NMF method to extract features from the angular images of an individual. The feature vectors associated with the angular faces are fused and then fed into a GAN neural network to estimate the feature vector of the frontal face. The general structure of the proposed method is depicted in Figure 1.

3. 1. NMF-Based Feature Extraction NMF is a feature extraction technique that utilizes a non-negative constraint, which distinguishes it from other methods [27]. In the NMF method, the constraint of non-negative elements in two matrices, W and H, are consistent with the intuitive concept, and therefore, the method learns component-based features [28]. In this technique, the image dataset is considered as a V matrix, which is an n × m matrix. Each column of the matrix represents n nonnegative values from one of the m face images. The matrix V is divided into two matrices W and H. According to Equation (1), each column of the matrix V is obtained as a linear combination of r columns of the matrix W:

$$V_{mn} \cong (WH)_n = \sum_{a=1}^{r} W_{ma}H_{an} \quad (1)$$

After obtaining the feature vector (H) of the dataset images based on Equation (1), the image (Y_i) is first converted into a vector, then by using the matrix W (obtained from the matrix analysis of the dataset images,W_{Train}) and the vector (Y_i), the feature vector of an angular image is obtained by using Equation (2).

$$H_i^{\sim} = W_{Train\_i}^{-1} \times Y_i \quad (2)$$

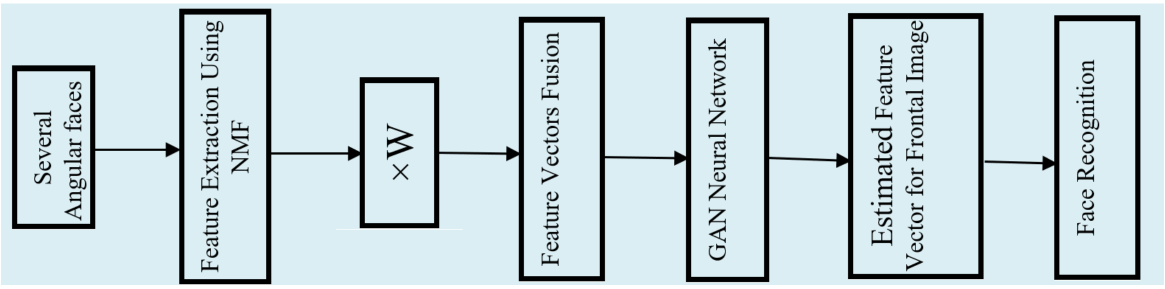


Figure 1. The overall structure of the proposed method. The weight values during the training phase of the proposed method are determined by using a genetic algorithm.

In this method, the feature vectors, extracted based on the component, have spatial dependence on each other, which can be used in the GAN network. In the next step, the optimal weight vector for the extracted feature vectors is obtained by the genetic algorithm. The steps for calculating the optimal weight vector are as follows:

3. 2 Optimal Weight Vector The feature vectors associated with angular images of the same person are fused by using a weight vector that is the same size as the feature vector and indicates the significance of each vector in the fusion process. During the training phase of the proposed method, the optimal weight vector is calculated based on the following steps:

1. After calculating the frontal image feature vector (H_i) according to Equation (1) and the feature vector of the angled images of the person($H_i^{\sim}$) based on Equation (2), the Euclidean distance between them is calculated according to Equation (3).

$$E_i = (H_i, H_i^{\sim}) = \sqrt{\sum_{i=1}^{M}(H_i - H_i^{\sim})^2} \qquad (3)$$

2. The Euclidean distance based on Equation (4) between the feature vector of the person's front image (H_i) and the feature vector of the person's angular image ($H_i^{\sim}$) is obtained by multiplying the optimal weight vector (W) which is the same size as the feature vector ($H_i^{\sim}$). The initial values of W are chosen randomly and then are optimized through several steps of the genetic algorithm. The algorithm selects chromosomes by using a roulette wheel at each iteration and forms new chromosomes in the next population by combining genes of two chromosomes based on the one-point crossover operator. The mutation operator randomly assigns new values within [0, 1] at the gene level on each of the chromosomes. Such an important feature helps the genetic algorithm to break out the local trap.

$$F_i = \left(H_i, (H_i^{\sim} \times W_i)\right) = \sqrt{\sum_{i=1}^{N}(H_i - (H_i^{\sim} \times W_i))^2} \qquad (4)$$

3. If F_i is smaller than E_i, it can be concluded that the optimal weight vector can reduce the Euclidean distance between the angular feature vector and the frontal feature vector of the person.

4. After calculating the optimal weight vector, the fusion vector is calculated for each person's angular feature vectors through the weighted averaging operation according to Equation (5).

$$H_{Fusion\_Person} = \frac{\sum_{i=1}^{k} H_i^{\sim} \times W_i}{k} \qquad (5)$$

In Equation (5), the value of k indicates the number of angular images of the person.

After generating the fusion feature vector for each individual, the GAN neural network is used to estimate the fusion feature vector associated with the frontal face feature vector.

3. 3. Network Structure In this study, we utilize a GAN network to estimate the feature vector which is linked with the frontal face. The GAN deep learning network comprises two sub-models, namely generative and discriminative [29]. A method based on feature enhancement GAN (FI-GAN) for face recognition was proposed by Rong et al. [30]. In this method, the difference between the front face and the profile is estimated, and FI-GAN maps the features of profile face images to the front space. The experimental results demonstrate that the proposed method can improve the accuracy of face recognition in large situations. Shahbakhsh and Hassanpour [31] utilized a GAN network for detecting low-resolution facial images. The proposed method employed feature-level image resolution enhancement to preserve the structure of low-resolution faces. This method primarily focuses on edges and reconstructing high-frequency details in the images. The proposed method successfully increased the accuracy of face recognition for low-resolution images within the range of 71.84 to 79.51. Han et al. [32] introduced a GAN-based method for face recognition from various angles. In this method, the front view is initially trained by a CNN network. Subsequently, the original face and the synthesized face are merged together. The results indicate that the recognition accuracy of the proposed method has improved compared to some existing methods.

Figure 2 illustrates the general structure of the proposed GAN network for estimating the frontal face feature vector.

Figure 3 shows the structure of the generator. The convolution layer in the proposed **generator structure** has a filter (1×1) with stride=1 and output =32. The proposed **discriminator network** utilizes a single convolution layer with Leaky Relu activation. The convolution layers are characterized by kernel sizes of (1×1) and filters of 32.

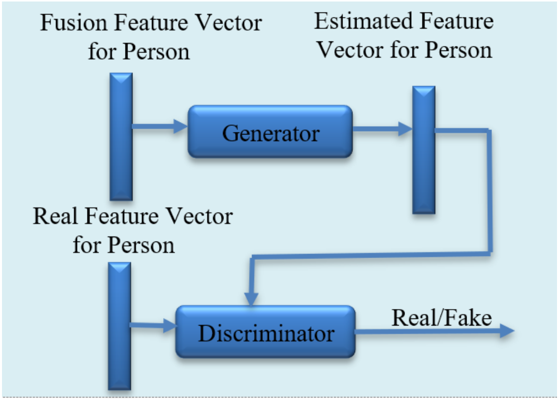

Fusion Feature Vector for Person
Estimated Feature Vector for Person
Generator
Real Feature Vector for Person
Discriminator
Real/Fake

Figure 2. The overall structure of the proposed GAN network for estimated feature vector associated with the frontal image

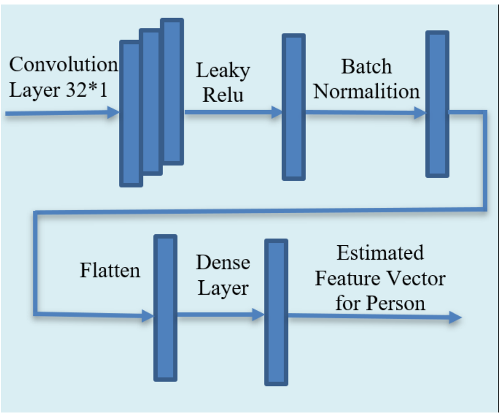


Figure 3. The general structure of the GAN network generator in Figure 2

This study employs two loss functions to train the generator structure. Each loss function compares the estimated feature vector with the frontal feature vector from different perspectives Finally, the loss function is obtained from the sum of these functions:

The Mean Absolute Error (MAE) function is used to minimize the distance between the estimated feature vector and the frontal feature vector. The MAE loss function is as follows:

$$\mathbf{Loss_{MAE}} = \frac{1}{n}\sum_{i=0}^{n} |\mathbf{l_i} - \hat{\mathbf{l}}_\mathbf{i}| \tag{6}$$

where l_i is the frontal feature vector, and $\hat{l}_i$ is the estimated frontal feature vector.

The Binary Cross Entropy (BCE) compares the predicted probabilities with the actual class output (0 or 1). It then calculates a score to penalize the probabilities based on their distance from the expected value. The BCE loss function is as follows:

$$\mathbf{Loss_{BCE}} = \mathbf{abs}(\hat{\mathbf{l}}_\mathbf{i} - \mathbf{l_i}) \tag{7}$$

where l_i is the frontal feature vector, and $\hat{l}_i$ is the estimated frontal feature vector.

Finally, the total loss results from the sum of all loss functions:

$$\boldsymbol{Total\ Loss} = \boldsymbol{Loss_{MAE}} + \boldsymbol{Loss_{BCE}} \tag{8}$$

4. EXPERIMENTS AND DISCUSSION

In this section, first, the dataset which was used for the proposed network is introduced. After that, we will explain the implementation details of the proposed method and evaluate its face recognition accuracy for angular faces.

4. 1. Dataset

The FERET dataset [33] contains 1684 face images, of which 1500 are used for training and 184 for testing. These images vary in light, face angle, pose position, etc. In the present study, for each person 6 images were used at angles of 5, 10, 15, 20, 30 and 40. Figure 4 shows images of the FERET dataset.

4. 2. Implementation Details

In the proposed method, images up to an angle of 40° have been selected for each person. As mentioned in section 3, first, feature vectors are obtained for the images using NMF. After that, the fusion operation is performed for the angular feature vectors. Experimentally, it was found that averaging was the most suitable fusion operation for appropriately estimating the frontal feature vector. In this research, the size of each image is 200 × 200, and by using the NMF method, the length of the feature vector obtained for each image is 538.

Subsequently, the estimated feature vector for each individual was obtained through the fusion feature vector by using the GAN neural network. The Adam optimizer (learning_rate=0.0001) was used to train the GAN neural network. All the code was written in Python 3.7 by using the Keras platform. In addition, network training and evaluation were performed by using the GeForce GTX 3060 GPU.

4. 3. Face Recognition

This article employs the correlation coefficient similarity criterion to compare the estimated feature vectors with the feature vectors of the images in the dataset. The correlation coefficient is calculated using the following formula:

$$\mathbf{C} = \frac{\sum_m \sum_n (\mathbf{X_{mn}} - \bar{\mathbf{X}})(\mathbf{Y_{mn}} - \bar{\mathbf{Y}})}{\sqrt{(\sum_m \sum_n (\mathbf{X_{mn}} - \bar{\mathbf{X}})^2)(\sum_m \sum_n (\mathbf{Y_{mn}} - \bar{\mathbf{Y}})^2)}} \tag{9}$$

where X is the estimated feature vector and Y is the feature vector of the image in the dataset. Also, $\overline{X}$ is the mean of X, and $\overline{Y}$ is the mean of Y.

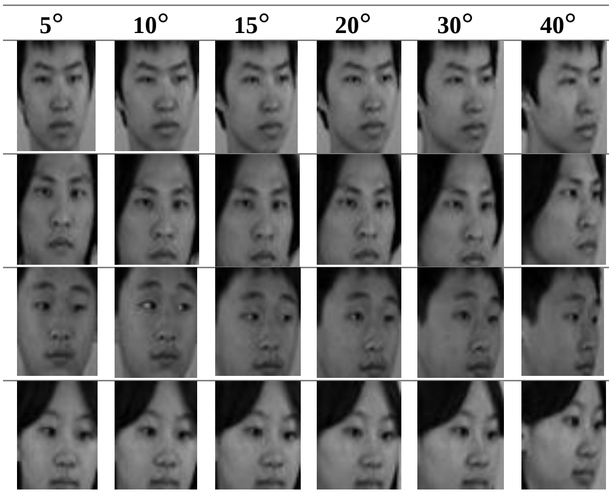


Figure 4. FERET Dataset images in different angles

The correlation coefficients of two feature vectors are considered in the decision-making process. Ultimately, the coefficients of the estimated feature vector are compared with the coefficients of all the feature vectors in the dataset. If the correlation coefficient between all the feature vectors in the dataset reaches the highest value, the two feature vectors are considered similar.

4. 3. 1. Face Recognition Results This section presents a comparison of the accuracy of the proposed method with PCA and Nikan [34] in Table 1. Based on the results, the existing identification methods are very sensitive to the angles of the image and can obtain good results when facing the camera. However, they lose their effectiveness when there is a slight change in the image angle. In contrast, the proposed method has been able to significantly improve the accuracy of angular face recognition by estimating the frontal feature vector. Figure 5 shows the results of the proposed method,

TABLE 1. Comparing face recognition accuracy between the proposed method and Nikan [34] , PCA

		RECOGNITION RATE (%)	
Face Recognition	Nikan [34]	up to 15 degrees	23
		up to 40 degrees	9
	PCA	up to 15 degrees	11
		up to 40 degrees	5
	PROPOSED METHOD	fusion feature vector up to 15 degrees	80
		fusion feature vector up to 40 degrees	63

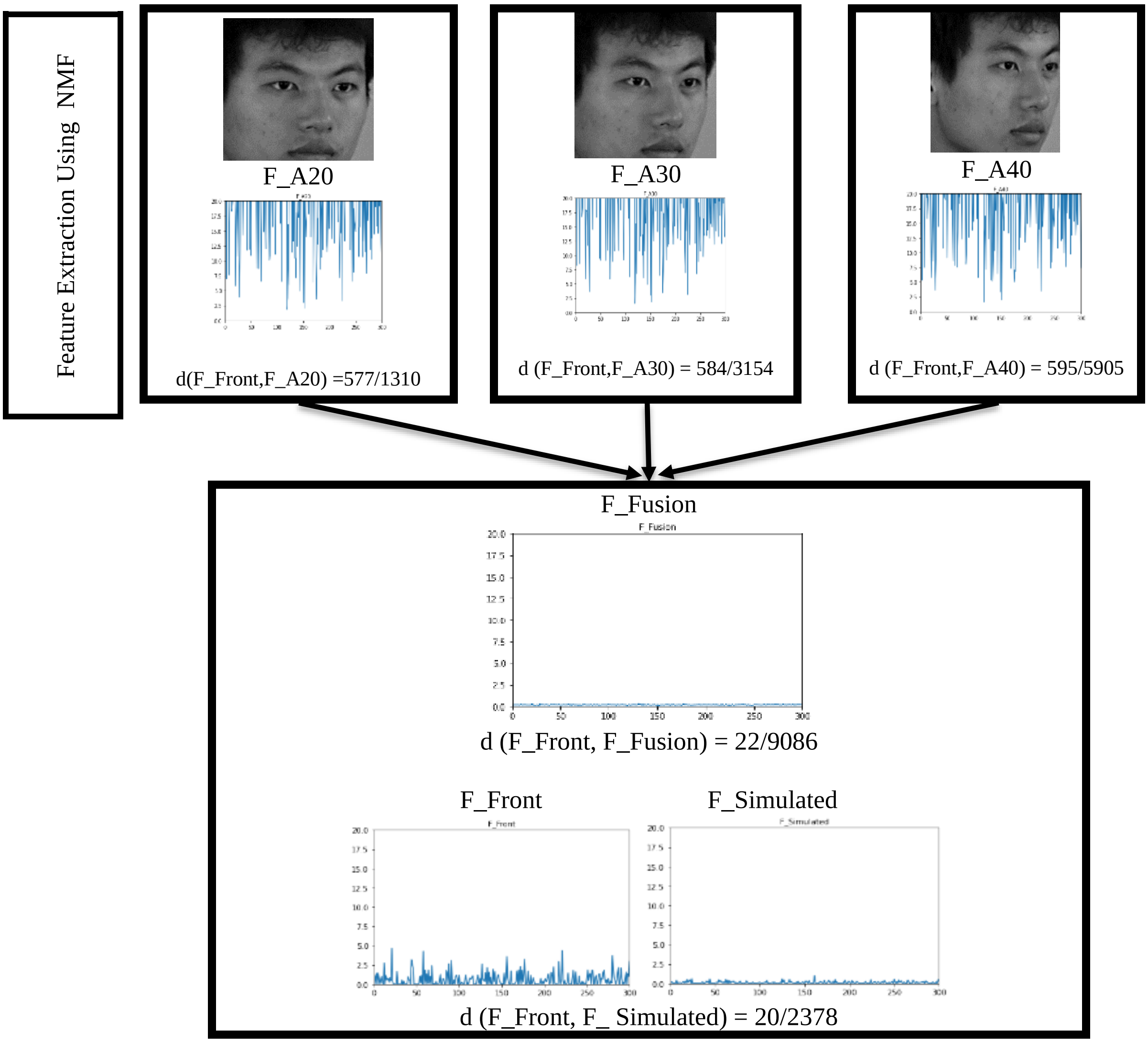


Figure 5. Feature vector estimated using the GAN network

including the representation of each feature vector and the similarity between the angular feature vector of the image and the frontal feature vector, based on the Euclidean distance. A smaller Euclidean distance between these two vectors indicates a higher degree of similarity.

5. CONCLUSION

This article proposed a technique to improve the accuracy of face recognition in the presence of angular faces. Feature vectors extracted from different angles of a person are fused and the obtained vector is fed to a GAN neural network to estimate the feature vector associated with the frontal face. Experimental results on the FERET dataset containing pose images with the angle of up to 40 degree indicate capability of the proposed method in detecting angular faces in video face recognition system.

6. REFERENCES

1. Kortli, Y., Jridi, M., Al Falou, A. and Atri, M., "Face recognition systems: A survey", ***Sensors***, Vol. 20, No. 2, (2020), 342. doi: 10.3390/s20020342.
2. Annalakshmi, M., Roomi, S.M.M. and Naveedh, A.S., "A hybrid technique for gender classification with slbp and hog features", ***Cluster Computing***, Vol. 22, (2019), 11-20. doi: 10.1007/s10586-017-1585-x.
3. Dinesh, P. S. and Manikandan, Dr.M., "Face Reconstruction using Eigenface and Neural Network", ***Tierärztliche Praxis***, Vol. 40, (2020), 1940-1968.
4. Aliyu, I., Ali Bom, M. and Maishanu, M., "A Comparative Study of Eigenface and Fisherface Algorithms Based on OpenCV and Sci-kit Libraries Implementations", ***International Journal of Information Engineering and Electronic Busin***, Vol. 3, (2022), 30-40. doi: 10.5815/ijieeb.2022.03.0.
5. Maghari, A.Y.A., "Recognition of partially occluded faces using regularized ICA", ***Inverse Problems in Science and Engineering***, Vol. 29, No. 8, (2021), 1158-1177. doi: 10.1080/17415977.2020.1845329.
6. Rakshit, R.D. and Kisku, D.R., "Face Identification via Strategic Combination of Local Features", ***Computational Intelligence in Pattern Recognition***, Vol. 999, (2020), 207-217.
7. Hassanpour, H. and Ghasemi, M., "A three-stage filtering approach for face recognition", ***International Journal of Engineering, Transactions B: Applications***, Vol. 34, No. 8, (2021), 1856-1864. doi: 10.5829/ije.2021.34.08b.06.
8. Shavandi, M. and Afrakoti, I., "Face recognition in thermal images based on sparse classifier", ***International Journal of Engineering, Transactions A: Basics***, Vol. 32, No. 1, (2019), 78-84. doi: 10.5829/ije.2019.32.01a.10.
9. Abed, R., Bahroun, S., Zagrouba, E., "KeyFrame extraction based on face quality measurement and convolutional neural network for efficient face recognition in videos ", ***Multimedia Tools and Applications***, Vol. 80, (2021), 23157-23179. doi: 10.1007/s11042-020-09385-5.
10. Zhang, J., Yan, X., Cheng, Z. and Shen, X., "A face recognition algorithm based on feature fusion", ***Concurrency and Computation: Practice and Experience***, Vol. 34, No. 14, (2022), e5748. doi: 10.1002/cpe.5748.
11. Shanthi, P., Nickolas, S., "An efficient automatic facial expression recognition using local neighborhood feature fusion ", ***Multimedia Tools and Applications,*** Vol. 80, (2021), 10187–10212. doi: 10.1007/s11042-020-10105-2.
12. Ksieniewicz, P., Zyblewski, P., Burduk, R., " Fusion of linear base classifiers in geometric space ", ***Knowledge-Based Systems***, Vol. 227, No. 3, (2021). doi: 10.1016/j.knosys.2021.107231.
13. Singh, M., Singh, R. and Ross, A., "A comprehensive overview of biometric fusion", ***Information Fusion***, Vol. 52, No., (2019), 187-205. doi: 10.1016/j.inffus.2018.12.003
14. AL-Shatnawi, A., Al-Saqqar, F., El-Bashir, M. and Nusir, M., "Face recognition model based on the laplacian pyramid fusion technique", ***International Journal of Advances in Soft Computing & Its Applications***, Vol. 13, No. 1, (2021).
15. Štruc, V., Gros, J.Z., Dobrišek, S. and Pavešic, N., "Exploiting representation plurality for robust and efficient face recognition", in Proceedings of the 22nd Intenational Electrotechnical and Computer Science Conference (ERK'13), Citeseer, (2013), 121-124.
16. Hu, J., Lu, J. and Tan, Y.-P., "Discriminative deep metric learning for face verification in the wild", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2014), 1875-1882.
17. Jabid, T., Kabir, M.H. and Chae, O., "Local directional pattern (ldp) for face recognition", in 2010 digest of technical papers international conference on consumer electronics (ICCE), IEEE, (2010), 329-330. doi: 10.1109/ICCE.2010.5418801.
18. Al-Dabagh, M.Z.N., Ahmad, M.I., Isa, M.N.M. and Anwar, S.A., "Face recognition system based on fusion features of local methods using cca", in 2020 8th International Electrical Engineering Congress (iEECON), IEEE, (2020), 1-4. doi: 10.1109/iEECON48109.2020.229489.
19. Liu, Z. and Liu, C., "Fusion of color, local spatial and global frequency information for face recognition", ***Pattern Recognition***, Vol. 43, No. 8, (2010), 2882-2890. doi: 10.1016/j.patcog.2010.03.003.
20. Uzair, M., Mahmood, A. and Mian, A., "Hyperspectral face recognition with spatiospectral information fusion and pls regression", ***IEEE Transactions on Image Processing***, Vol. 24, No. 3, (2015), 1127-1137. doi: 10.1109/TIP.2015.2393057.
21. Bi, Y., Lv, M., Wei, Y., Guan, N. and Yi, W., "Multi-feature fusion for thermal face recognition", ***Infrared Physics & Technology***, Vol. 77, (2016), 366-374. doi: 10.1016/j.infrared.2016.05.011.
22. Zhu, Y. and Jiang, Y., "Optimization of face recognition algorithm based on deep learning multi feature fusion driven by big data", ***Image and Vision Computing***, Vol. 104, (2020), 104023. doi: 10.1016/j.imavis.2020.104023.
23. Wang, Y., Shi, D. and Zhou, W., "Convolutional neural network approach based on multimodal biometric system with fusion of face and finger vein features", ***Sensors***, Vol. 22, No. 16, (2022), 6039. doi: 10.3390/s22166039.
24. Medjahed, C., Rahmoun, A., Charrier, C. and Mezzoudj, F., "A deep learning-based multimodal biometric system using score fusion", ***IAES International Journal of Artificial Intelligence***, Vol. 11, No. 1, (2022), 65. doi: 10.11591/ijai.v11.i1.
25. Xu, H., Qi, M. and Lu, Y., "Multimodal biometrics based on convolutional neural network by two-layer fusion", in 2019 12th International Congress on Image and Signal Processing,

BioMedical Engineering and Informatics (CISP-BMEI), IEEE, (2019), 1-6. doi: 10.1109/CISP-BMEI48845.2019.8966036.

26. Almabdy, S. and Elrefaei, L., "Feature extraction and fusion for face recognition systems using pre-trained convolutional neural networks", ***International Journal of Computing and Digital Systems***, Vol. 9, No., (2021), 1-7. doi: 10.12785/ijcds/100144.

27. Khosravi, M.H., Hassanpour, H. and Ahmadifard, A., "A content recognizability measure for image quality assessment considering the high frequency attenuating distortions", ***Multimedia Tools and Applications***, Vol. 77, (2018), 7357-7382. doi: 10.1007/s11042-017-4636-7.

28. Aonishi, T., Maruyama, R., Ito, T., Miyakawa, H., Murayam, M., Ota, K.," Imaging data analysis using non-negative matrix factorization ", ***Neuroscience Research***, Vol. 179, (2022), 51-56. doi: 10.1016/j.neures.2021.12.001.

29. Liu, H., Zheng, X., Han, J., Chu, Y. and Tao, T., "Survey on gan-based face hallucination with its model development", ***IET Image Processing***, Vol. 13, No. 14, (2019), 2662-2672. doi: 10.1049/iet-ipr.2018.6545.

30. Rong, C., Zhang, X. and Lin, Y., "Feature-improving generative adversarial network for face frontalization", ***IEEE Access***, Vol. 8, (2020), 68842-68851. doi: 10.1109/ACCESS.2020.2986079.

31. Shahbakhsh, M.B. and Hassanpour, H., "Empowering face recognition methods using a gan-based single image super-resolution network", ***International Journal of Engineering***, Vol. 35, No. 10, (2022), 1858-1866. doi: 10.5829/IJE.2022.35.10A.05.

32. Han, Z. and Huang, H., "Gan based three-stage-training algorithm for multi-view facial expression recognition", ***Neural Processing Letters***, Vol. 53, (2021), 4189-4205. doi: 10.1007/s11063-021-10591-x.

33. Phillips, P.J., Wechsler, H., Huang, J. and Rauss, P.J., "The feret database and evaluation procedure for face-recognition algorithms", ***Image and Vision Computing***, Vol. 16, No. 5, (1998), 295-306. doi: 10.1007/s11063-021-10591-x.

34. Nikan, F. and Hassanpour, H., "Face recognition using non-negative matrix factorization with a single sample per person in a large database", ***Multimedia Tools and Applications***, Vol. 79, No. 37-38, (2020), 28265-28276. doi: 10.1007/s11042-020-09394-4.

Persian Abstract

چکیده

در سال‌های اخیر، تشخیص چهره به دلیل ماهیت غیرتداخلی، به مناسب‌ترین فناوری برای طراحی سیستم‌های بیومتریک تبدیل شده است. این فناوری در صنایع مختلفی از جمله مراقبت‌های بهداشتی، آموزشی، امنیتی و نظارتی مورد استفاده قرار می‌گیرد. فناوری تشخیص چهره زمانی بهترین عملکرد را دارد که فرد مستقیماً به دوربین نگاه کند. برعکس، عملکرد تشخیص چهره زمانی که با یک تصویر چهره زاویه‌دار مواجه می‌شود کاهش می‌یابد، زیرا معمولاً با استفاده از تصاویر یک چهره کامل آموزش داده می‌شود. هدف از این مقاله تخمین بردار ویژگی یک تصویر تمام صورت است، زمانی که چندین تصویر زاویه‌دار از یک فرد وجود دارد، مانند تصاویری که در یک ویدیو یافت می‌شود. این روش ویژگی‌های اساسی یک تصویر چهره را با استفاده از روش فاکتورسازی ماتریس غیر منفی (NMF) استخراج می‌کند. سپس، بردارهای ویژگی با استفاده از یک شبکه متخاصم مولد (GAN) برای تخمین بردار ویژگی مرتبط با تصویر جلویی ترکیب می‌شوند. نتایج تجربی به دست‌آمده بر روی تصاویر زاویه‌ای مجموعه داده FERET نشان می‌دهد که روش پیشنهادی می‌تواند به طور قابل توجهی دقت فناوری تشخیص چهره را بهبود بخشد.

International Journal of Engineering

Journal Homepage: www.ije.ir

Segmenting the Lesion Area of Brain Tumor using Convolutional Neural Networks and Fuzzy K-Means Clustering

S. Fooladi, H. Farsi*, S. Mohamadzadeh

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran

P A P E R I N F O

Paper history:
Received 10 May 2023
Received in revised form 13 June 202
Accepted 14 June 2023

Keywords:
Brain Tumor
Convolutional Neural Networks
Fuzzy K-Means
Segmentation

A B S T R A C T

Brain tumor Segmentation is one of the most crucial methods of medical image processing. Non-automatic segmentations are broadly used in clinical diagnosis and medication. However, this kind of segmentation does not have accuracy in medical images, especially in terms of brain tumors, and it provides a low level of reliability. The primary objective of this paper is to develop a methodology for brain tumor segmentation. In this paper, a combination of Convolutional Neural Network and Fuzzy K-means algorithm has been presented to segment the lesion area of brain tumor. It contains three phases, Image preprocessing to reduce computational complexity, Attribute extraction and selection and Segmentation. At first, the database images are pre-processed using adaptive filters and wavelet transform in order to recover the image from the noise state and reduce the computational complexity. Then feature extraction is performed by the proposed deep neural network. Finally, it is processed through the Fuzzy K-Means algorithm to segment the tumor region separately. The innovation of this article is related to the implementation of deep neural network with optimal parameters, identification of related features and removal of unrelated and repetitive features with the aim of observing a subset of features that describe the problem well and with minimal reduction in efficiency. This results in reduced feature sets, storage of data collection resources during operation, and overall data reduction to limit storage requirements. This proposed segmentation approach has been verified on BRATS dataset and produces the accuracy of 98.64%, sensitivity of 100% specificity of 99%.

***doi**: 10.5829/ije.2023.36.08b.15*

NOMENCLATURE

ψ (t)	Mother wavelet function	Ĉ	showed the purpose
x (t)	signal wavelet transform	$\bar{r}_{cf}$	correlation mean value
$\psi_{s,\tau}$	wavelet function	$\bar{r}_{ff}$	one-to-one mean value of correlation
O $_s$	Feature vector	r_{cfi} and r_{fifj}	variables
X $_r$	r^{th}input channel	N	the number of data points
W $_r$	the kernel for input	M	the fuzzy parameter
X	the input to the ReLU function	η	positive and small number

1. INTRODUCTION

Image segmentation is defined as dividing a digital image into several sections (a collection of pixels also, known as superpixel). The purpose of segmentation is to simplify or/and create a change in displaying pixels to those that are more meaningful and simpler for analysis. Segmentation is usually used for finding the location of objects and boundaries (lines, curves, etc.) in the image. In other words, image segmentation means the process, in which a label is allocated to each pixel so that the pixels with the same labels have similar features.

Cancer can be defined as abnormal and uncontrolled growth and division of body cells. This occurrence means the abnormal growth and divisions of the cells in the brain tissues as a mass, and it is called a brain tumor. Brain tumors are not very common; however, they are from very deadly types of cancers [1]. Brain tumors

*Corresponding Author Email: hfarsi@birjand.ac.ir (H. Farsi)

Please cite this article as: S. Fooladi, H. Farsi, S. Mohamadzadeh, Segmenting the Lesion Area of Brain Tumor using Convolutional Neural Networks and Fuzzy K-Means Clustering, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1556-1568

might have various shapes and sizes, and they grow enough until the diagnosis time. The most common brain tumor among adults is glioma which is made of glial cells and has the highest rate of mortality prevalence [2]. This type of brain tumor is divided into High-Grade Glioma and Low-Grade Glioma based on the severity of the glioma and its origin [3]. Low-Grade Glioma has a less aggressive effect and penetration compared to High-Grade Glioma [4]. At the moment, brain tumor segmentation usually depends on the personal and scientific experience of the doctor. This situation presents various segmentation results due to the different professional knowledge of the doctors, in addition to wasting much time and human mistakes. Further, brain tumor segmentation is challenging because of the different shapes and similarities of the Gray Level between the tumor tissue and its adjacent organs. Thus, the method of accurate and efficient segmentation of brain tumors has been a critical research method using deep neural networks [5]. Recognizing the tissues of brain lesions and determining the situation of these tissues in the medical images are accounted as significant issues in medical images. Analyzing the pathology presents an important role in diagnosing, predicting, and medical planning for brain tumors. Today, great achievements have been provided by studies about brain tissues and their molecular understanding. At the present, some tools can be created for analyzing these complicated images automatically, using digital pathology, such as digital scanning and saving the tumor tissue sections in patients. Therefore, analyzing image data has attracted a lot of attention in recent years. Kernel segmentation from the tissue images is necessary, especially for the approaches relevant to the biological characteristics. The type of tissue, the difference in color, and the type of cell present various visual features, and they lead to many difficulties for segmentation algorithms of the traditional images that work appropriately for all of these cases.

The non-automatic analysis of the many sampled slides of the tissue by the doctors is an intensive and expensive process. Thus, the computerized diagnosis systems that are being converted into influential tools are used by doctors to discover and diagnose tumors [6, 7]. The most common method for brain tumor medication is surgery. However, some methods, such as radiotherapy and chemotherapy are used to reduce the speed of tumor growth. Brain tumor segmentation in images might have a significant effect in diagnosing the tumor properly, predicting its growth speed, and also, planning for the medication. Some tumors, such as meningioma can be easily segmented. Nevertheless, defining the location of tumors such as glioma is much more difficult. These tumors are always more scattered with swelling around them, and they have poor contrast with the healthy tissues around themselves. In addition, they spread with tentacle-like structures that make their segmentation difficult. The other basic problem in brain tumor segmentations is their different shapes and sizes anywhere in the brain [8]. Brain tumor segmentation performed by seasoned radiologists is considered a standard reference. However, the semi-automatic and full-automatic computer segmentation methods result in improving the speed of segmentation and reproducibility of the results. Moreover, the full-automatic segmentation removes inconsistency between the observer and within the observer as the result of some factors, such as differences in expertise, attention, and errors due to visual fatigue [9-11]. Besides, significant progress has been achieved in increasing the similarity of segmentation in the manual and automatic methods with segmentation algorithms using deep neural networks [12].

Healthy brains are usually made of three types of tissue: white matter, gray matter, and cerebrospinal fluid. The purpose of brain tumor segmentation is to diagnose the area and prevent the development of the tumor area, meaning the tissue area of active tumor of necrotic and edema. This action is performed by identifying the abnormal areas compared to the natural tissue [13, 14].

In this study, the researchers segment the MRI (BRATS) images, in which the data are directly controlled as a section of the learning process of the neural network via the proposed deep neural network architecture. Afterward, we observed more accuracy with increasing speed by comparing this model to several common algorithms used in this field.

In this research, by using the fuzzy K-means clustering method, structural similarity is considered as an index and this index is used as an important parameter to find the similarity between segmented results and ground truth images.

Next, in order to select the appropriate feature, we extract the feature that contains information around the target and define a set of features that have a high correlation with the target feature as a suitable set. This definition of the Deep Neural Network with optimal parameters and high learning power at a suitable speed is one of the innovative aspects of this research.

One of the advantages of using deep neural networks is the automatic adjustment of parameters and weights at every moment of training. The mechanism of sharing the weights in each feature also makes it possible that the number of parameters in each layer of the neural network is reduced and the computational load on the processor is avoided.

The structure of this study is as follows:

Section two reviews some studies that have been conducted in this field, section three introduces the proposed method, and section four presents the results of this study. Finally, this study ends with a general conclusion in section five.

2. RELATED WORKS

Many methods have been introduced by researchers for automatic and semi-automatic segmentation of brain tumors in recent years. Making difference among the body tissues in medical images manually is boring and results in human mistakes. The crucial purpose of each method is to identify and classify the tumor area properly. Many studies have been carried out about the automatic segmentation of brain tumors using deep learning considering the success of deep neural networks in terms of medical image processing.

Toğaçar et al. [15] have performed the process of feature extraction for effective segmentation using the architectures of ALEXnet and VGG16. Thus, in the first stage, they enhanced the outstanding features via Hyper column technique, and in the second stage, they combined the extracted features from both architectures. In this method, recurrent feature elimination was utilized to select the most appropriate features. Ultimately, a support vector machine was applied for the segmentation. This method eventually reported 96% accuracy for this study.

Amin et al. [16] used a deep learning algorithm by focusing on preprocessing and MRI image segmentation before presenting it as an input. Their idea was to sharpen the images; thus, they used the median filter which is one of the non-linear filters in digital filtering to remove noise. After that, the tumor area was segmented with accurate adjustment using the growing area to give it as the input to a model of stacked sparse autoencoders (SSAE). This model was trained and examined on the collection of BRATS data. The results indicated the accuracy and sensitivity improvement of the proposed techniques compared to other methods.

Islam et al. [17] focused on multi-level segmentations, and first, they preprocessed the database images to extract the efficient features from the MRI scans of the brain tumors. Afterward, they segmented the areas of brain tumors using methods of thresholding, watershed algorithm, and morphological operations. In this method, the features were extracted from the convolution, and the database images were classified as two cancer and non-cancer classes via the K-SVM method. The proposed algorithm reported an accuracy of 87.4%.

Zhang et al. [18] investigated brain tumor segmentation from MRI images via multiple encoders. This model reduced the difficulty of feature extraction by defining several encoders and improving segmentation accuracy. Besides, this model presented Categorical Dice Loss which provided various weights for different areas of segmentation to solve the problem of unbalanced data. The proposed method illustrated the accuracy of 88.2% for the segmentation.

Hasan et al. [19] proposed an improved model of U-net which was introduced by substituting an inverse convolution stage with an algorithm that was the nearest neighbor for the increased sample. Besides, an elastic transformation was used to enhance the collection of training data to empower the model for database image segmentation.

Rajan and Sundar [20] implemented a system based on a combination of K-Means with FCM methods, and they used the active contour as a post-processing for brain tumor segmentation. Standardizing the image severity was the main purpose of using active contour. The function of the proposed method was evaluated based on the black-and-white pixels and the tumor locations. This study provided a comparable function to other approaches.

The other effective study was conducted by Begum and Lakshmi [21] with the title of combining statistical wavelets and recurrent neural networks for brain tumor segmentation. This study classified and segmented the brain tumor via statistical features. To do so, it preprocessed the images for noise removal, and then, it extracted the statistical features, using longitudinal navigation of tissue and GLCM matrix. Afterward, the features were reduced via gravitational search algorithm (OGSA) and were given to the recurrent neural network to classify the images as tumor and non-tumor classes. After that, the images were entered into the next implementation step for the area segmentation. In this case, the algorithm of modified region growing was used for the segmentation stage.

Thaha et al. [22] introduced the enhanced convolution neural network (E-CNN) by Loss function optimization and using the BAT algorithm to segment the abnormalities from the MRI images of the brain. Therefore, the results of accuracy improvement of the segmentation were shown via intelligent optimization.

Gao and Qian [23] focused on one of the methods of artificial neural networks called as DeepLab. This method made difference between lesion and background using the semantic-based and patch-based segmentation approaches. In the following, it accurately adjusted the borders of the lesion area by combining some other methods, such as conditional random fields (CRF).

Emadi et al. [24] have proposed a new method for improving brain tumor segmentation accuracy based on super-pixel and fast primal dual (PD) algorithms. The proposed method detects brain tumor tissue in Flair-MRI imaging in BRATS2012 dataset. This method detects the primary borders of tumors using a super-pixel algorithm, and improves brain tumor borders using fast PD in Markov random field optimization. Then, post-processing processes are used to delete white brain areas. Finally, an active contour algorithm was employed to display tumor area. Different experiments were carried on the proposed method and qualitative and quantitative criteria such as sensitivity, accuracy and F-measure were used for evaluation. The obtained results showed the efficiency of the proposed method in the accuracy and sensitivity are 86.59% and 88.57% and F1-Measure 86.37%.

Azimi et al. [25] presented a fully-automated method based on graph shortest path layer segmentation and fully convolutional networks (FCNs) for fluid segmentation.

This research presented a fully-automated method for fluid segmentation based on fully convolutional networks (FCNs) applied to OCT scans and their corresponding regions of interest computed by graph shortest path in neutrosophic (NS) domain. From the results of this research, it can be concluded that in the future will be train FCN with augmented training data by random translation, reflection, rotation, flipping and cropping to achieve more accurate results.

Khan et al. [26] have presented the segmentation process for brain tumor images by using the K-Means clustering method and deep learning by increasing the combined data, focusing on the non-invasive feature of MRI images and better display of internal tumor information.

Rai et al. [27] merged CNN with the full fuzzy specialist (NS-CNN) neutrosophic, confident entropy to diagnose brain tumors. These images were then added to the CNN for the extraction of characteristics and finally, extracted features are fed in the SVM classification to be classified as benign or malignant with an averaged 95.62% accuracy.

We carefully find out in related works that the reported methods often used traditional approaches and pre-trained networks and researchers try to classify the created classes and finally the desired segmentation. The proposed method tries to provide an efficient technique with high accuracy and applicable at a suitable speed on ordinary processors. Therefore, we define the proposed research in 3 sections: pre-processing, feature extraction and selection, and segmentation. In the pre-processing stage, a method is presented to remove noise and reduce computational complexity, and further, by using the concepts of deep learning and the definition of convolutional neural network, the high-level features of medical images are extracted, which can be of great help in accurate segmentation. Finally, using the Fuzzy K-Means algorithm, we will try to minimize the distortion and the best clustering for the final segmentation of the images.

3. THE PROPOSED METHOD

Algorithm 1 illustrates the general segmentation process of the lesion area of a brain tumor. First, in the proposed method, the proposed CNN extracted the critical features of the images from the preprocessed images by the adaptive filters. Afterward, the features with high significance were selected via correlation-based feature selection. Finally, the tumor area was extracted from the primary images via the fuzzy K-Means algorithm.

In algorithm 1, database images that are manually segmented are considered as input.

Algorithm 1: Algorithm of the proposed method.

```
Input:
  1) trainImgSet: The medical images Set, with segmented
     brain tumor areas manually in theirs;
  2) targets = The segmented brain tumor areas manually in
     trainImgSet.

get N = The number of images in trainImgSet
get wavelet = The wavelet transform according to Equation (1,2)
get th_w = The threshold limit of wavelet transform
get CNN = Our Convolutional Neural Netwprk
get FKM = The fuzzy-kmeans algorithm
get K = The number of clusters needed to feature clustering in FKM
get th_k = The threshold limit of FKM

for i = 1 to N do:
    WT[i]= wavelet(trainImgSet[i])
    WT_b[i]= remove coefficients less than th_w in WT[i]
    Im_wt[i]= inverse wavelet(WT_b[i])
    RI=Divide Im_wt[i] into 9 equal areas
    CNN_Features=[]
    for region in the RI do :
        RCI=CNN(region)
        add RCI to CNN_Features
    Selcted_features= Applying feature selection algorithm
      on CNN_Features according to Equation 8,9
    Segmented_features=FKM(Selcted_features,K,th_k)

Output:
  Segmented_features = an image, that tumor pixels are
  distinguished.
```

The designed convolution network was regarded based on Table 1, and the N variable was the number of training images. The k variable was the number of required clusters for clustering the image pixels (with and without tumors).

th_w is the threshold limit of the defined wavelet transform, and th_k equaled the threshold limit of K-Means of X variable for each parameter or other network that has been already determined in this study. The wavelet transform on the image has been applied according to formulas 1 and 2. Afterward, the coefficients less than the y parameter were removed, and the inverse wavelet transform (denoised image) was applied. Correspondingly, the image was divided into nine equal areas. The selected features were chosen based on the formula of eight and nine references by applying CNN transform to each area and extracting the features. The fuzzy K-Means were applied to the selected features based on k and z (dividing them into two classes). Ultimately, the output of an image similar to the original

TABLE 1. The Proposed Architecture of CNN for Extracting Feature

Layer number	Layer type	Filter size	Stride	filters	Fc unit	Input
Layer 1	convolution	3*3	1*1	64*64	-	4*33*33
Layer 2	convolution	3*3	1*1	64*64	-	64*66*33
Layer 3	convolution	3*3	1*1	64*64	-	64*33*33
Layer 4	Max-Pooling	3*3	2*2	-	-	64*33*33
Layer 5	convolution	3*3	1*1	128*128	-	64*16*16
Layer 6	convolution	3*3	1*1	128*128	-	128*16*16
Layer 7	convolution	3*3	1*1	128*128	-	128*16*16
Layer 8	Max-Pooling	3*3	2*2	-	-	128*16*16
Layer 9	FC				256	6272
Layer 10	FC				5	256

image, in which the tumor pixels were determined was displayed as the final purpose of this study.

3. 1. Preprocessing

Pre-processing steps, including normalization in order to prevent to lose features and wavelet transformation, a process that reduces unnecessary information for the convolutional neural network and leads to optimal use of the proposed convolutional neural network. And finally, the structure of the adaptive filter, which is used with fixed and predetermined specifications for pre-processing operations in order to reduce the computational complexity.

First, we cut the images relevant to the database in this study to reduce the computational complexity and create a model for better evaluation. After that, we adjusted the image brightness to understand the proposed network from the database images better. Adaptive filters are filters that can change their parameters in some ways despite the traditional filters with fixed and predetermined specifications. Therefore, they can respond to the changes in their surrounding environment, considering specific purposes. Wavelet transform is a method for displaying the image in two dimensions of time and frequency. All the wavelet functions have been made of a wavelet called a mother wavelet. Wavelet transform is a function of scale that is related to the inverse frequency and transform which has been shown in Equation (1).

The modified and extended versions of the mother wavelet can be demonstrated as the signal wavelet transform of x (t) with the mother wavelet function of ψ(t) [28].

$$\psi_{s,r} = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \tag{1}$$

Signal wavelet transform of x (t) and wavelet function of $\psi_{s,\tau}$ are displayed as follows [28]:

$$T\,(s,\ \tau)= \int_{-\infty}^{+\infty} x\ (t)\ \psi^{*}\left(\frac{t-\tau}{s}\right)dt \tag{2}$$

Wavelet transform is a combination of two low-pass and high-pass filters that are applied to the input image during various stages. Two small and large scales were used to introduce the high and low frequencies in these transforms. The purpose of small scales was to achieve the short-term behaviors of the image, and the purpose of large scales was to access the long-term behavior of the image. To do so, this transform used an image and selected one mother wavelet that this study had used a Daubechies wavelet. Correspondingly, wavelet transform presented images with high frequency, representing image details, and images with low frequency, approximately representing input image in each stage. Therefore, the input image was achieved from the output of the approximation low-pass filter, and details of the input image were achieved from the output of the high-pass filter in this study. These filters reported the wavelet coefficient and scaling function. Sub-band coding, including the sequence of the filtering process and reduction of sampling rate, was used in this study. In the first stage, the input image was filtered by two high-pass and low-pass filters. After that, the output of both filters was reduced in sampling by factor 2. In the second stage, the output of the low-pass filter in the first stage was filtered by those low-pass and high-pass filters and its rate was reduced by factor 2 so that the output sequence was produced with a length of N/4. This process of filtering the output of the low-pass filter and rate reduction continued. The database images were decomposed into the wavelet coefficient, using wavelet transform. To do so, the Daubechies wavelet of db2 was used to extract the features. Wavelet thresholding was performed to recover the image from the noisy mode in the wavelet transform method. Therefore, the small coefficient of the wavelet was adjusted to zero, and the coefficients compatible with the image remained. This process decreases the unnecessary information for the CNN, and it leads to optimized use of the proposed CNN. Figure 1 shows the block diagram of image preprocessing and deep neural network training in order to reduce computational complexity and to create a better evaluation model.

The block diagram of Figure 1 shows the combination of pre-processing and deep network training. The purpose of normalization in this section is that all features are involved in our decision-making and features with large values do not remove other features. Wavelet transform parameters are changed in such a way that they are able to respond to the changes in their surroundings according to specific goals and the short-term and long-term behaviors of the images are known. In this block diagram, suitable features are extracted using deep neural network and after denormalization, the output is provided to the K-means Fuzzy algorithm for segmentation.

3. 2. The Proposed Deep Neural Network In this section, the proposed method based on CNN has been introduced to extract the appropriate features of brain tumors. Training network is minimizing the error function based on the real outputs of the network compared to the appropriate outputs of the network. This process was done by modifying free network parameters, meaning weights and biases. The method of training used in the current proposed structure was the training method with an observer. Thus, a supervisor observed the behavior of the learners and reminded them to do the proper action. In other words, the learner system is a set of data pairs, consisting of network input and appropriate output. After applying the network input, its output was compared to the appropriate output. Besides, the learning error of computation was used to modify the network parameters in a way that if it was given to those input networks once again, the network output was closer to the appropriate output. The Loss function should have reached its lowest limit despite being non-linear to train the CNN. In the following, a sliding window (filter) was considered in all the image sections to make difference between the normal areas and tumor areas or the cancer cell nucleus. Therefore, each area of the image determined the local tissue from the image pixels via these windows and introduced them to the CNN. All the information and features received from the local tissues determined by the windows helped identify the tumor area and cancer cell nucleus more accurately. Further, a more accurate decision could be made to identify the healthy and damaged tissues from that image by putting the features of these sections together. In this case, the brain tumor tissues were completely observable. In this stage, the features were extracted from the determined areas of each image using the proposed CNN. The filter size in the convolution layers was considered 3*3; thus, the image resolution might have decreased during this path. Each feature vector of O_s was related to one or several kernels, concerning the convolution filters. The feature vector was achieved from the following equation [29]:

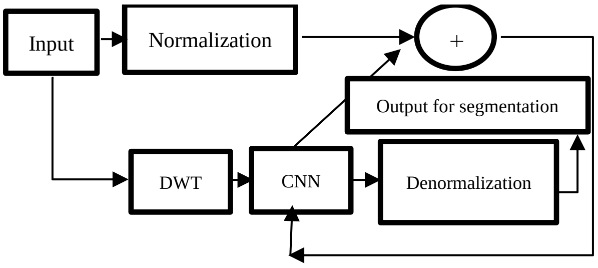


Figure 1. Block diagram of preprocessing the image and training the deep neural network

$$O_s = b_s + \sum_r W_{sr} * X_r \qquad (3)$$

X_r was the r^{th}input channel, W_r was the kernel for input, * showed the convolution operator, and b_s was considered the bias number. In other words, the convolution operation was performed for each feature vector. The collection of convolution filters was added to each pixel by the sum of one number as bias that provided the location of each pixel. However, feature extraction in the traditional methods depended on one fixed instruction. The excellence of the proposed CNN to these methods was due to their ability in learning weights and extracting specific features in particular jobs. Correspondingly, a non-linear element was applied to the convolution results to achieve the transformed non-linear features of the input. In this study, an activation function of ReLU was used because the network could train more quickly without making difference in accuracy due to computational efficiency. This function was very effective in terms of computation and let the network be converged quickly. The reason was that its relationship was linear; thus, it was faster than the Sigmoid and Tanh functions.

The mathematical relation of this function was as follows:

Parameter X was the input to the ReLU function, and here, the values of computed pixels were in the convolution layer [30].

$$f(\mathrm{x}) = \max(0, \mathrm{x}) \qquad (4)$$

The purpose of using non-linear activation functions in the proposed CNN was to create a complex mapping between the inputs and outputs. In other words, these functions provided our model the possibility to adapt itself to complex and non-linear data. The pooling layer was periodically put among the convolution layers in the proposed CNN at certain intervals after each convolution layer. The purpose of putting the pooling layers in this study was to decrease the mapping size of features and parameters of the deep neural network. The function of this layer was to reduce the spatial size of the image to decrease the parameter numbers and computations inside the network, and finally, to control the overfitting. The most common form of using this layer was using the layers with filters by size and Max_Pooling.

The Max_Pooling layer was used after the activation function. This action selected the maximum value in each window of the feature vectors. Thus, it kept the number of feature pages; however, the size of the feature page decreased. The computation relation of this action has been stated in relations 5 and 6 [29]:

$$Z_{s,i,j} = \max\{O_{s,i,j}, O_{s+1,i,j}, \dots, O_{s+K-1,i,j}\} \quad (5)$$

$$H_{s,i,j} = \max Z_{s,i+p,j+p} \quad (6)$$

In relation 5, the p symbol indicated the size of the Max_Pooling window. Max_Pooling actions reduced the size of feature vectors. This action was performed under the built-in windows by controlling the regarded pooling size and the steps in the vertical and horizontal modes. Figure 2 shows one stage of the blocks of the convolution layer, activation functions, and pooling layer.

The convolution networks could extract a hierarchy of increasingly complex features that made them more attractive. This process was performed by processing the feature vectors achieved from the output of a convolution layer that was used as the input of the lateral convolution sublayers. As it is obvious from the fully-connected layer, all the neurons of this layer were connected to the previous layer. The main duty of the fully-connected layer was combining the local feature in the bottom layer, especially the local feature in the top layers. Dropout was used to prevent overfitting in the fully-connected layer. The way of working this layer was that in each stage of training, some nodes of the network were removed with the probability of p-1, and other nodes remained with the probability of p. Therefore, a decreased network remained that prevented overfitting.

Loss function was used in this study, and it was tried to reach in minimum in training and testing. To do so, Categorical Cross-entropy has been used. The C symbol indicated the probable predictions, and Ĉ showed the purpose [31].

$$H = -\Sigma_{j \in voxels} \Sigma_{k \in classes} C_{j,k} \log\left(\hat{C}_j, k\right) \quad (7)$$

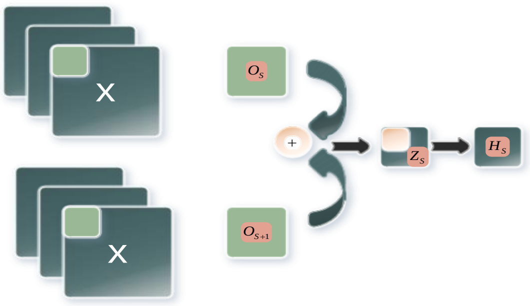


Figure 2. The blocks of the convolution layer, activation functions, and pooling layer

Table 1 illustrates the used layers in the architecture of the deep neural network. The size of digital filters of all layers was considered 3*3 in this study. First, the Max_Pooling layer was used after three convolution layers and an activation function. The size of the steps was 1*1 in the convolution layer, and 2*2in the Max_Pooling layer. If we define these layers as a box, another box is made of the three convolution layers and the activation function same as the first box in the following. Ultimately, two fully-connected layers and no network overfitting were used.

3. 3. Feature Selection

Feature selection can be defined as the procedure of identifying relevant features and removing irrelevant and repetitive features. Correspondingly, the purpose is to observe a subset of features that defines the issue clearly with a minimal reduction in efficiency degree. This method has various advantages that have been explained in this study as folows:

- Improving the efficiency of machine learning algorithms
- Understanding the data, achieving knowledge about the procedure, and helping its visualization
- Decreasing the general data, limiting requirements, and saving and probably helping costs decrease
- Decreasing the features collection, saving the resources in the following period, and collecting data during the use
- Having the simplicity and capability of using simpler models and gaining speed

To recognize a feature relevant to the issue, this definition was used so that one feature is relevant if it has information about the purpose.

The method of correlation-based selection feature was used in this study. In this method of feature selection, the subsets of features were considered good subsets, in which the features had a high correlation with the target feature on one hand, and they were uncorrelated on the other hand. The merit or being good of a subset of features was computed via the following relation in this study [32].

$$Merit_{S_k} = \frac{k\bar{r}_{cf}}{\sqrt{k+k\,(k-1)\,\bar{r}_{ff}}} \quad (8)$$

In this relationship, $\bar{r}_{cf}$ was the correlation mean value that was computed between the target feature and all the features in the data set. Further, $\bar{r}_{ff}$ was the one-to-one mean value of correlation computed among the features. Finally, the correlation-based method was formulated as follows [32]:

$$CFS = \max_{s_k}\left[\frac{r_{cf1}+r_{cf2+\cdots+}r_{cfk}}{\sqrt{k+2\,(r_{f2f1}+r_{fifj}}+\cdots+r_{fkf1}}\right] \quad (9)$$

In this relation, the variables of r_{cfi} and r_{fifj} were regarded as the correlation variables. Further, the correlation-based method was used to select the best features.

3. 4. Segmentation by Fuzzy K-Means

This method was considered an exclusive and flat method in this study. Different forms have been defined for this algorithm; however, all of them had a repetitive procedure that tried to estimate the following items for a fixed number of clusters.

Gaining some points as the cluster centers. These points were actually those point means that belonged to each cluster.

Attributing each given data to a cluster where the data had the shortest distance to the center.

This method was used to reduce distortion [33].

$$j = \Sigma_{j=1}^{k}\Sigma_{j=1}^{N}u_{i,j}^{m}d_{ij} \quad (10)$$

In this relation, the N indicated the number of data points, and the m showed the fuzzy parameter which equaled 2. The cluster numbers were displayed as the K symbol that represented the square of Euclidean distance between selected pixels in the image with a clustering center. u_{ij} should have regarded this limitation for the above relation according to this relation [33].

$$\Sigma_{j=1}^{N}u_{ij} = 1 \qquad i = 1\ to\ N \quad (11)$$

Reducing the Euclidean distance was the first priority of this study to segment the database images, considering that the purpose of Euclidean distance was the target function. Therefore, the distortion was reduced in the target function. The FKM algorithm started clustering by a collection of the primary centers in a way that these centers have been selected completely randomly, and none of the two or several clusters had the same cluster center. Afterward, the function components were updated to compute the new centers using the Euclidean distance. A group was made between those image pixels and the nearest center of the cluster after computing the new centers. Thus, a repetitive procedure was done. The new cluster centers changed their locations for each repetition until the cluster center was stable. Correspondingly, the fuzzy K-Means algorithm reduced the Euclidean distance between image pixels and cluster centers which were our target function. By minimizing the Euclidean distance, the distortion reached its lowest degree. Therefore, the distortions were reduced in the target function. In this technique, the function of the new membership was determined by gaining the value mean of the previous membership function.

The K-Means algorithm worked in a way that it first, selected a set of primary clusters randomly and adjusted P=1. Afterward, the square of the Euclidean distance of d_{ij} was computed, and the membership function of u_{ij} was updated using mathematical relations [33].

$$u_{ij} = \left(\left(d_{ij}\right)^{\frac{1}{m}-1}\Sigma_{l=1}^{k}\left(\frac{1}{d_{il}}\right)^{\frac{1}{m}-1}\right)^{-1} \quad (12)$$

In this mode, l ≠j, if the d_{ij} <η, and $u_{ij} = 1$ is adjusted, where the η is a positive and small number.

In the next step, the new set of cluster centers was computed using the following equation [33].

$$C_j = \frac{\Sigma_{i=1}^{N}u_{ij}^{m}X_1}{\Sigma_{i=1}^{nN}u_{ij}^{m}} \quad (13)$$

Finally, $\|C_j - C_{j-1}\|$< ε is stopped for repeating j=1 to N, otherwise, p+1→p was adjusted, and the second stage was repeated.

The computation complexity for the third step was higher than the fourth step. In this algorithm, the ideal condition was provided, and the repetition stopped after 10 stages of repetition

4. RESULTS

4. 1. Database

The purposive evaluation received from numerous and new methods of brain tumor image segmentation was more difficult. However, a widely accepted criterion was used for the automatic segmentation of brain tumors to develop the BRATS criteria. At the moment, purposive comparison of numerous methods of brain tumor segmentation was possible, using this common database. The BRATS database contains 274 MRI scans of glioma patients, which are divided into HGG and LGG levels, this version was segmented by an expert manually so that the proposed system function was evaluated by these scans. The image dimensions were decreased to 254*254 pixels to increase the processing speed. Some examples of the database are shown in Figure 3[2].

4. 2. Evaluation Criteria

In this section, the output data of deep learning was compared to the

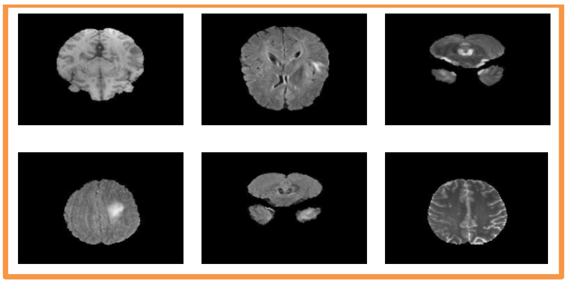


Figure 3. Some Examples of the Database Images [2]

[2] www.kaggle.com/datasets/dschettler8845/brats-2021-task1

diagnostic data in society by specialist doctors, and finally, the efficiency of the proposed methods was validated. Ultimately, the function of the proposed brain tumor segmentation was evaluated by using various criteria, such as sensitivity and accuracy.

TN: Represented the number of records, of which the real cluster was negative, and classification algorithms recognized their cluster as negative properly.

TP: Represented the number of records, of which the real cluster was positive, and classification algorithms recognized their cluster as positive properly.

FP: Represented the number of records, of which the real cluster was negative, and classification algorithms recognized their cluster as positive by mistake.

FN: Represented the number of records, of which the real cluster was positive, and classification algorithms recognized their cluster as positive by mistake.

The ability to assess the sick and healthy cases from other cases was called accuracy. The following relation has illustrated this concept [33].

$$Accuracy=\frac{TN+TP}{TN+FN+TP+FP} \qquad (14)$$

The accuracy criteria did not make difference between FN and FP. Thus, the precision criterion was defined to solve this problem.

The ability of one method to find sick cases, lesion areas, and cancer nuclei is called sensitivity. To compute the sensitivity of a test, the proportion of the true positive rate to the sum of the true positive rate and negative false should be computed which is been shown in the following relation [34].

$$sensitivity=\frac{TP}{TP+FP} \qquad (15)$$

4. 3. The Output Results

In this section, the efficiency of the proposed method was compared to other methods that indicated effective parameter improvement in database image segmentation. The purpose of using the K-Means algorithm in the proposed method was to implement this algorithm easily and quickly. Considering the sensitivity of this algorithm to the primary cluster centers, it could produce a locally optimal response. This algorithm was one of the valid methods of clustering that performed clustering means based on the shortest distance of each data from a cluster center. Table 2 shows the implementation of segmenting lesion areas of brain tumors by various methods that depended on selecting features from the images. It was observed that the proposed method had more segmentation accuracy and precision compared to other methods. This was related to the method of high-level feature extraction, using CNN and the correlation-based feature selection among the feature.

TABLE 2. Comparing the Results of Lesion Area Segmentation of Brain Tumors with Other Methods

Methods	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
ALEX-Net + VGG16 [15]	96	99	98.1
Stacked Sparse Autoencoders [16]	89.2	88.3	92.2
CNN+Multilevel segmentation [17]	87.4	90	90
Multi Encoder – Net (ME-Net) [18]	88.2	90.1	89
Modified U-Net [19]	91	93	90
K-means + FCM [20]	89	-	-
Optimal Wavelet Staistical + RNN+ Modified Region Growing [21]	96	100	92
Enhanced-CNN [22]	92	92	87
K-means + deep learning [26]	94.06	89.9	90.01
NS-CNN feature fed to SVM classifier [27]	95.62	-	-
Deep Lab+CRF [23]	85.7	87	86.7
Proposed method	**98.64**	**100**	**99**

Figures 4 and 5 report the increasing procedure of accuracy and Loss function minimization in two stages of training and testing. The purpose of using fuzzy logic in this study was to develop the classical set theories in mathematics. The elements' membership followed a zero pattern and a binary pattern. However, the theory of fuzzy

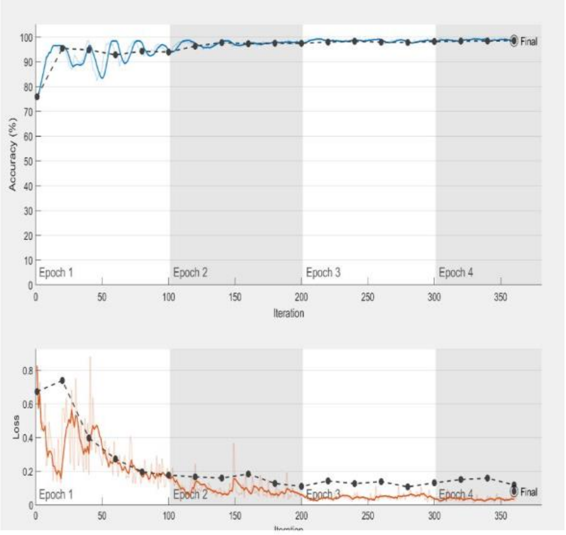


Figure 4. The Progress Procedure in the Training Stage

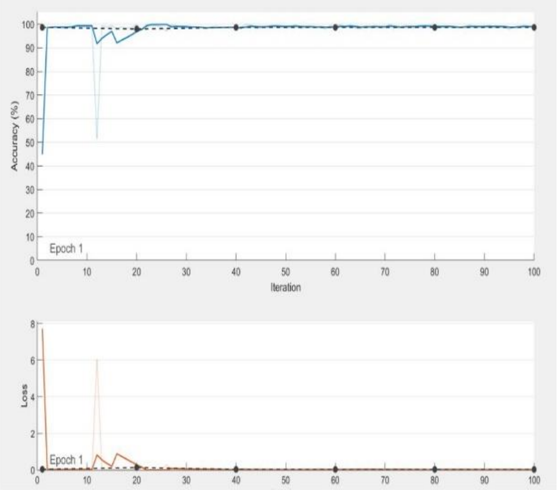


Figure 5. The progress procedure in the testing stage

sets developed this concept and introduced graded membership. Therefore, one element could be a member of a set to some degree and not completely. This concept helped increase image segmentation accuracy and Loss function minimization.

Figures 4 and 5 show the process of maximizing the accuracy of the proposed method and minimizing the loss pan. In this process, there is a period or epoch when the entire data set is transferred back and forth through the neural network only once. Since an era is too large to be entered into the system at once, we divide it into several smaller categories called epochs.

In this study, the fuzzy neural networks clusterer was used as the best separator, and training the network was based on using images segmented and indexed by the K-Means algorithm as the input in this network. Afterward, computing the white points of various brain areas, diagnosing the different brain areas, estimating the tumor area, and diagnosing the exact location of the brain tumor were done by fuzzy clustering. Therefore, the exact location of the tumor could be computed by extracting the best features from the image. In this study, it was tried to help the doctors' diagnosis via computer techniques and tools due to the high significance of diagnosing brain tumors in the later stages of treating the patient. Figures 6 and 7 illustrate the output of this research, showing the lesion area and the area of different tumor types in advanced mode.

In Figure 6, only the tumor and non-tumor area are specified. This figure shows the original image, the image diagnosed by the medical doctor, and the final output image segmented by the proposed method, which shows a more comprehensive and accurate diagnosis than the one specified by the medical doctor. Meanwhile in Figure 7, in addition to identifying the tumor area, different types of tumors are also identified.

In this research, convolutional neural network is trained in order to extract suitable features in database images. In the following, by defining the feature selection method based on correlation, a subset of features are created that have correlation with the target feature and are not correlated with each other. We consider this type of feature selection as the process of identifying related features and removing unrelated and repetitive features. The use of fuzzy logic enables a process-oriented view of the result along with the use of various conditions. Since fuzzy K-Means segmentation is important in this research from the point of view that in the images, we consider points as cluster centers, and the average points belong to the cluster. This increases the accuracy of segmentation. On the other hand, in this case, we assign each data sample that has the smallest distance to the cluster centers to that cluster, which increases the speed of segmentation. The advantage of this regularization method under the Gaussian criterion is to obtain suitable cluster centers, which reduces non-homogeneous interference and better segmentation.

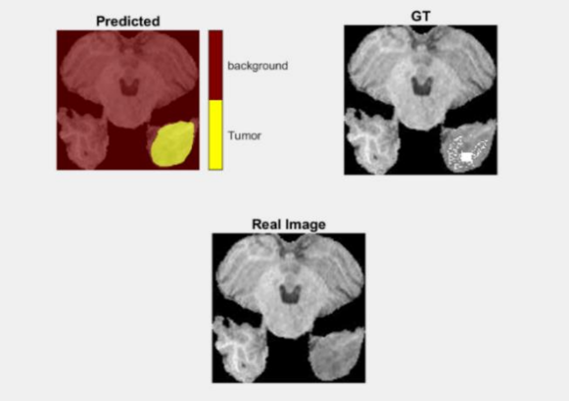


Figure 6. The output image and the segmentation of brain lesion area

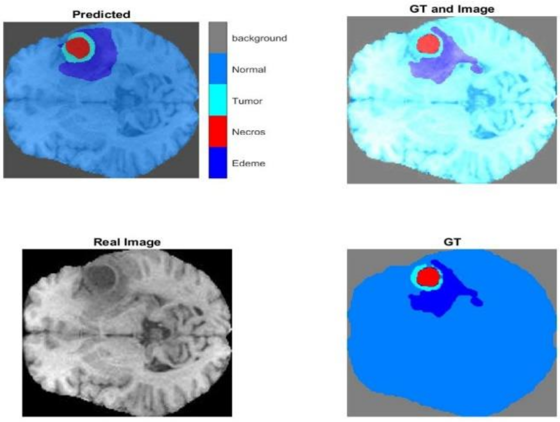


Figure 7. The output image and the segmentation of brain lesion area

5. CONCLUSION

In clinical practice, segmenting the area of a brain tumor in the image still depends on the human operators; nevertheless, manual segmentation is a time-consuming process, and its quality completely depends on the operator's experience. In this field, finding a complete automatic segmentation method is necessary to determine the brain tumor area in measuring the tumor exactly. Some progresses have recently been done in the semi-automatic and fully-automatic algorithms to segment brain tumor. However, there are main challenges for this process due to the many varieties of brain tumors in size, shape, location, and heterogenous appearance. The diagnosis speed presented in the method is much faster than the proposed methods in other studies that use low-level learning methods. In addition, the diagnosis method is performed by a person. This subject can be known as the result of using hierarchy learning of the proposed method that led to deep learning. Further, selecting features with high significance by the method of correlation-based feature selection and decreasing the size of the feature vector were other consequences. The modern world has made it possible to receive and store images digitally. Sometimes, in order to obtain better results, it is necessary to make changes for the purpose of processing, analyzing and understanding the image. In this context, using the science of mathematics and needs assessment in the field of medicine with the use of artificial intelligence, goals such as optimizing deep learning networks through a combined method, considering algorithms to minimize the variance of images with the aim of reducing the avoidable differences in terms of the fact that the network needs less data for training and focusing on optimal feature selection can be considered as a research approach for researchers.

5. REFERENCES

1. Yogananda, C.G.B., Shah, B.R., Vejdani-Jahromi, M., Nalawade, S.S., Murugesan, G.K., Yu, F.F., Pinho, M.C., Wagner, B.C., Emblem, K.E. and Bjørnerud, A., "A fully automated deep learning network for brain tumor segmentation", ***Tomography***, Vol. 6, No. 2, (2020), 186-193. doi: 10.18383/j.tom.2019.00026.

2. Soomro, T.A., Zheng, L., Afifi, A.J., Ali, A., Soomro, S., Yin, M. and Gao, J., "Image segmentation for mr brain tumor detection using machine learning: A review", ***IEEE Reviews in Biomedical Engineering***, (2022). doi: 10.1109/RBME.2022.3185292.

3. Liu, Z., Tong, L., Chen, L., Jiang, Z., Zhou, F., Zhang, Q., Zhang, X., Jin, Y. and Zhou, H., "Deep learning based brain tumor segmentation: A survey", ***Complex & Intelligent Systems***, Vol. 9, No. 1, (2023), 1001-1026. doi: 10.3390/app122311980.

4. Magadza, T. and Viriri, S., "Deep learning for brain tumor segmentation: A survey of state-of-the-art", ***Journal of Imaging***, Vol. 7, No. 2, (2021), 19. doi: 10.3390/jimaging7020019.

5. Somasundaram, S. and Gobinath, R., "Current trends on deep learning models for brain tumor segmentation and detection–a review", in 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE. (2019), 217-221.

6. Biratu, E.S., Schwenker, F., Ayano, Y.M. and Debelee, T.G., "A survey of brain tumor segmentation and classification algorithms", ***Journal of Imaging***, Vol. 7, No. 9, (2021), 179. doi: 10.3390/jimaging7090179.

7. Devunooru, S., Alsadoon, A., Chandana, P. and Beg, A., "Deep learning neural networks for medical image segmentation of brain tumours for diagnosis: A recent review and taxonomy", ***Journal of Ambient Intelligence and Humanized Computing***, Vol. 12, (2021), 455-483. doi: 10.1007/s12652-020-01998-w.

8. Zhang, W., Wu, Y., Yang, B., Hu, S., Wu, L. and Dhelim, S., "Overview of multi-modal brain tumor mr image segmentation", in Healthcare, MDPI. Vol. 9, (2021), 1051.

9. Karimi, D. and Salcudean, S.E., "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks", ***IEEE Transactions on Medical Imaging***, Vol. 39, No. 2, (2019), 499-513. doi: 10.1109/TMI.2019.2930068.

10. Arabi, H., Dowling, J.A., Burgos, N., Han, X., Greer, P.B., Koutsouvelis, N. and Zaidi, H., "Comparative study of algorithms for synthetic ct generation from mri: Consequences for mri-guided radiation planning in the pelvic region", ***Medical Physics***, Vol. 45, No. 11, (2018), 5218-5233. doi: 10.1002/mp.13187.

11. Arabi, H., Zeng, G., Zheng, G. and Zaidi, H., "Novel adversarial semantic structure deep learning for mri-guided attenuation correction in brain pet/mri", ***European Journal of Nuclear Medicine and Molecular Imaging***, Vol. 46, (2019), 2746-2759. doi: 10.1007/s00259-019-04380.

12. Bahrami, A., Karimian, A., Fatemizadeh, E., Arabi, H. and Zaidi, H., "A new deep convolutional neural network design with efficient learning capability: Application to ct image synthesis from mri", ***Medical Physics***, Vol. 47, No. 10, (2020), 5158-5171. doi: 10.1002/mp.14418.

13. Angulakshmi, M. and Deepa, M., "A review on deep learning architecture and methods for mri brain tumour segmentation", ***Current Medical Imaging***, Vol. 17, No. 6, (2021), 695-706. doi: 10.2174/1573405616666210108122048.

14. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M. and Maier-Hein, K.H., "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge", in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3, Springer. (2018), 287-297.

15. Toğaçar, M., Cömert, Z. and Ergen, B., "Classification of brain mri using hyper column technique with convolutional neural network and feature selection method", ***Expert Systems with Applications***, Vol. 149, (2020), 113274. doi: 10.1016/j.eswa.2020.113274Get rights and content.

16. Amin, J., Sharif, M., Gul, N., Raza, M., Anjum, M.A., Nisar, M.W. and Bukhari, S.A.C., "Brain tumor detection by using stacked autoencoders in deep learning", ***Journal of Medical Systems***, Vol. 44, (2020), 1-12. doi: 10.1007/s10916-019-1483.

17. Islam, R., Imran, S., Ashikuzzaman, M. and Khan, M.M.A., "Detection and classification of brain tumor based on multilevel segmentation with convolutional neural network", ***Journal of Biomedical Science and Engineering***, Vol. 13, No. 4, (2020), 45-53. doi: 10.4236/jbise.2020.134004.

18. Zhang, W., Yang, G., Huang, H., Yang, W., Xu, X., Liu, Y. and Lai, X., "Me-net: Multi-encoder net framework for brain tumor segmentation", ***International Journal of Imaging Systems and***

Technology, Vol. 31, No. 4, (2021), 1834-1848. doi: 10.1002/ima.22571.

19. Hasan, S.K. and Linte, C.A., "A modified u-net convolutional network featuring a nearest-neighbor re-sampling-based elastic-transformation for brain tissue characterization and segmentation", in 2018 IEEE Western New York Image and Signal Processing Workshop (WNYISPW), IEEE. (2018), 1-5.
20. Rajan, P. and Sundar, C., "Brain tumor detection and segmentation by intensity adjustment", ***Journal of Medical Systems***, Vol. 43, (2019), 1-13. doi: 10.1007/s10916-019-1368-4&.
21. Begum, S.S. and Lakshmi, D.R., "Combining optimal wavelet statistical texture and recurrent neural network for tumour detection and classification over mri", ***Multimedia Tools and Applications***, Vol. 79, (2020), 14009-14030. doi: 10.1007/s11042-020-08643.
22. Thaha, M.M., Kumar, K.P.M., Murugan, B., Dhanasekeran, S., Vijayakarthick, P. and Selvi, A.S., "Brain tumor segmentation using convolutional neural networks in mri images", ***Journal of Medical Systems***, Vol. 43, (2019), 1-10. doi: 10.1007/s10916-019-1416-0.
23. Gao, X. and Qian, Y., "Segmentation of brain lesions from ct images based on deep learning techniques", in Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, SPIE. Vol. 10578, (2018), 610-615.
24. Emadi, M., Jafarian Dehkordi, Z. and Iranpour Mobarakeh, M., "Improving the accuracy of brain tumor identification in magnetic resonanceaging using super-pixel and fast primal dual algorithm", ***International Journal of Engineering***, ***Transactions C: Aspects***, Vol. 36, No. 3, (2023), 505-512. doi: 10.5829/IJE.2023.36.03C.10.
25. Azimi, B., Rashno, A. and Fadaei, S., "Fully convolutional networks for fluid segmentation in retina images", in 2020 International Conference on Machine Vision and Image Processing (MVIP), IEEE. (2020), 1-7.
26. Khan, A.R., Khan, S., Harouni, M., Abbasi, R., Iqbal, S. and Mehmood, Z., "Brain tumor segmentation using k-means clustering and deep learning with synthetic data augmentation for classification", ***Microscopy Research and Technique***, Vol. 84, No. 7, (2021), 1389-1399. doi: 10.1002/jemt.23694.
27. Rai, H.M., Chatterjee, K. and Dashkevich, S., "Automatic and accurate abnormality detection from brain mr images using a novel hybrid unetresnext-50 deep cnn model", ***Biomedical Signal Processing and Control***, Vol. 66, (2021), 102477. doi: 10.1016/j.bspc.2021.102477.
28. Engineering, J.O.H., "Retracted: Brain tumor detection and classification by mri using biologically inspired orthogonal wavelet transform and deep learning techniques", ***Journal of Healthcare Engineering***, Vol. 2023, (2023), 9845732. doi: 10.1155/2023/9845732.
29. Obeidavi, M.R. and Maghooli, K., "Tumor detection in brain mri using residual convolutional neural networks", in 2022 International Conference on Machine Vision and Image Processing (MVIP), IEEE. (2022), 1-5.
30. Dubey, S.R., Singh, S.K. and Chaudhuri, B.B., "Activation functions in deep learning: A comprehensive survey and benchmark", ***Neurocomputing***, (2022). doi: 10.1016/j.neucom.2022.06.111.
31. Clough, J.R., Byrne, N., Oksuz, I., Zimmer, V.A., Schnabel, J.A. and King, A.P., "A topological loss function for deep-learning based image segmentation using persistent homology", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 44, No. 12, (2020), 8766-8778. doi: 10.1109/TPAMI.2020.3013679.
32. Balasubramanian, K. and Ananthamoorthy, N., "Correlation-based feature selection using bio-inspired algorithms and optimized kelm classifier for glaucoma diagnosis", ***Applied Soft Computing***, Vol. 128, (2022), 109432. doi: 10.1016/j.asoc.2022.109432.
33. Nawaz, M., Mehmood, Z., Nazir, T., Naqvi, R.A., Rehman, A., Iqbal, M. and Saba, T., "Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering", ***Microscopy Research and Technique***, Vol. 85, No. 1, (2022), 339-351. doi: 10.1002/jemt.23908.
34. Musallam, A.S., Sherif, A.S. and Hussein, M.K., "A new convolutional neural network architecture for automatic detection of brain tumors in magnetic resonance imaging images", ***IEEE Access***, Vol. 10, (2022), 2775-2782. doi: 10.1109/ACCESS.2022.3140289.

Persian Abstract

چکیده

قطعه بندی تومور مغزی یکی از مهم‌ترین روش‌های پردازش تصاویر پزشکی به شمار می رود. تقسیم بندی های غیر اتوماتیک به طور گسترده در تشخیص و درمان بالینی مورد استفاده قرار می گیرند، این نوع تقسیم بندی در تصاویر پزشکی بخصوص تصاویر مربوط به تومور مغزی دقت بالایی ندارد و سطح پایینی از قابلیت اعتماد را فراهم می‌کند. هدف اصلی این مقاله توسعه روشی برای تقسیم بندی تومور مغزی است. در این مقاله، ترکیبی از شبکه عصبی کانولوشن و الگوریتم K-means فازی برای تقسیم‌بندی ناحیه ضایعه تومور مغزی ارائه شده است. این پژوهش شامل سه مرحله است، (۱) پیش پردازش تصویر برای کاهش پیچیدگی محاسباتی (۲) استخراج و انتخاب ویژگی (۳) تقسیم بندی. در ابتدا، تصاویر پایگاه داده با استفاده از فیلترهای تطبیقی و تبدیل موجک به منظور بازیابی تصویر از حالت نویز و کاهش پیچیدگی محاسباتی، پیش پردازش می شوند، این فرآیند باعث کاهش اطلاعات غیر ضروری برای شبکه عصبی کانولوشن می‌شود و منجر به استفاده بهینه از شبکه عصبی کانولوشن می‌گردد. سپس استخراج ویژگی توسط شبکه عصبی عمیق پیشنهادی انجام می شود. در نهایت، از طریق الگوریتم فازی K-Means پردازش می‌شود تا ناحیه تومور را به طور جداگانه تقسیم‌بندی کند. نوآوری این مقاله مربوط به اجرای شبکه عصبی عمیق با پارامترهای بهینه، شناسایی ویژگی‌های مرتبط و حذف ویژگی‌های نامرتبط و تکراری با هدف مشاهده زیرمجموعه‌ای از ویژگی‌هایی است که مسئله را به خوبی و با حداقل کاهش کارایی توصیف می‌کنند. این ایده منجر به کاهش مجموعه ویژگی‌ها، ذخیره‌سازی منابع، جمع‌آوری داده‌ها در طول عملیات و کاهش کلی داده‌ها به منظور محدود کردن نیازهای ذخیره‌سازی و متعاقبا منجر به کاهش هزینه‌ها می‌شود. این رویکرد تقسیم‌بندی پیشنهادی بر روی مجموعه داده‌های BRATS انجام گرفته شده و دقت ۹۸/٦٤٪، حساسیت ۱۰۰٪ ویژگی ۹۹٪ را ایجاد می‌کند.

International Journal of Engineering

Journal Homepage: www.ije.ir

Cross-modal Deep Learning-based Clinical Recommendation System for Radiology Report Generation from Chest X-rays

S. Shetty*[a,b], V. S. Ananthanarayana[a], A. Mahale[c]

[a] *Department of Information Technology, National Institute of Technology Karnataka, Mangalore, Karnataka, India*
[b] *Department of Computer Science and Engineering, NMAM Institute of Technology, Nitte (Deemed to be University), Udupi, Karnataka, India*
[c] *Deperatment of Radiology, Kasturba Medical College, Mangalore, Manipal Academy of Higher Education, Manipal, Karnataka, India*

PAPER INFO

Paper history:
Received 01 May 2023
Received in revised form 09 June 2023
Accepted 10 June 2023

Keywords:
Radiology Reports
Deep Learning
Encoder
Decoder
Clinical Recommendation System
Report Generation

ABSTRACT

Radiology report generation is a critical task for radiologists, and automating the process can significantly simplify their workload. However, creating accurate and reliable radiology reports requires radiologists to have sufficient experience and time to review medical images. Unfortunately, many radiology reports end with ambiguous conclusions, resulting in additional testing and diagnostic procedures for patients. To address this, we proposed an encoder-decoder-based deep learning framework that utilizes chest X-ray images to produce diagnostic radiology reports. In our study, we have introduced a novel text modelling and visual feature extraction strategy as part of our proposed encoder-decoder-based deep learning framework. Our approach aims to extract essential visual and textual information from chest X-ray images to generate more accurate and reliable radiology reports. Additionally, we have developed a dynamic web portal that accepts chest X-rays as input and generates a radiology report as output. We conducted an extensive analysis of our model and compared its performance with other state-of-the-art deep learning approaches. Our findings indicate significant improvement achieved by our proposed model compared to existing models, as evidenced by the higher BLEU scores (BLEU1 = 0.588, BLEU2 = 0.4325, BLEU3 = 0.4017, BLEU4 = 0.3860) attained on the Indiana University Dataset. These results underscore the potential of our deep learning framework to enhance the accuracy and reliability of radiology reports, leading to more efficient and effective medical treatment.

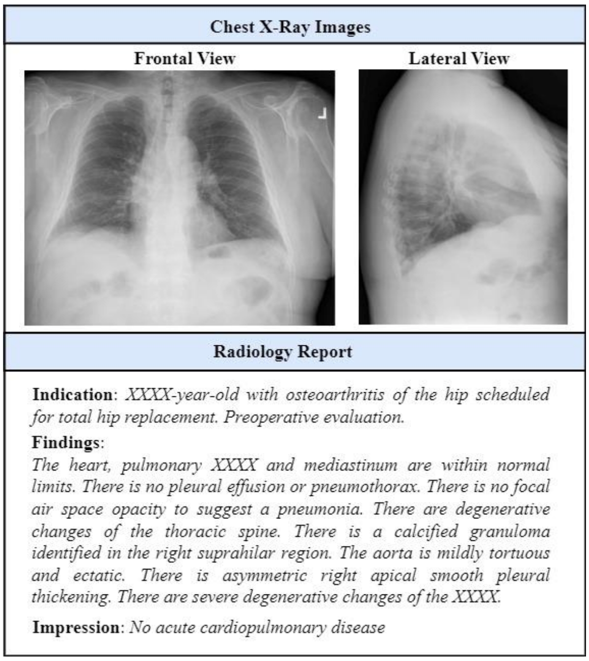
doi: *10.5829/ije.2023.36.08b.16*

1. INTRODUCTION

Hospitals around the world heavily rely on medical imaging, which provides valuable insights for disease diagnosis and treatment planning [1, 2]. However, it is crucial for the radiologist to thoroughly examine the medical images in order to provide comprehensive findings and interpretations [3]. The sample chest x-ray images with the associated unstructured medical text data is shown in Figure 1. The radiologists carefully analyze the chest X-rays, which include both frontal and lateral views. Their meticulous examination leads to the creation of comprehensive reports that confirm the diagnosis by documenting their detailed findings [4]. A diagnostic report is a comprehensive document that includes several sections to convey important information about a patient's condition. One of these sections is the indication, which describes the reason why the diagnostic test was ordered. The findings section presents the results of the test, including any abnormalities or other observations that were made. The impression section summarizes the radiologist's overall interpretation of the test results and may include a diagnosis or recommended next steps. Finally, the report may also include manual annotations, which are additional notes or comments that the radiologist has added to provide more context or clarification.

*Corresponding Author Email: shashankshetty.177it002@nitk.edu.in (S. Shetty)

Please cite this article as: S. Shetty, V. S. Ananthanarayana, A. Mahale, Cross-modal Deep Learning-based Clinical Recommendation System for Radiology Report Generation from Chest X-rays, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1569-1577



Indication: *XXXX-year-old with osteoarthritis of the hip scheduled for total hip replacement. Preoperative evaluation.*

Findings:

The heart, pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia. There are degenerative changes of the thoracic spine. There is a calcified granuloma identified in the right suprahilar region. The aorta is mildly tortuous and ectatic. There is asymmetric right apical smooth pleural thickening. There are severe degenerative changes of the XXXX.

Impression: *No acute cardiopulmonary disease*

Figure 1. Sample chest X-ray images (i.e., frontal and lateral view) with associated unstructured medical report

In order to produce precise and reliable radiology reports, it is necessary for the radiologist to possess ample experience and devote a significant amount of time to scrutinizing the medical images [5]. A large number of radiology reports may end with inconclusive comments, resulting in patients undergoing additional tests, such as advanced imaging or pathology exams. The issue of the time required for a radiologist to create a detailed report is a significant concern, as on average, an experienced radiologist will need approximately 10-20 minutes to produce a thorough report. In situations such as overcrowded hospitals or during a pandemic [6-8], writing radiology reports can become challenging due to the ever-increasing number of cases [9]. These circumstances inspired our research into developing an automated radiology reporting system using a deep learning framework to faciliate Clinical Recommendation System (CRS) [10].

A CRS is a critical component of modern healthcare delivery systems that are necessary for providing high-quality healthcare. CRS is "*a health information system that assists clinicians in making well-informed decisions about patient care by utilizing patient data, including medical history, current medications, and symptoms, to provide enhanced evidence-based recommendations to clinicians in real-time"*. We propose an artificial intelligence (AI)-based CRS framework for generating diagnostic reports from Chest X-Rays (CXR).

The organization of this paper is as follows: Section 2 introduces the problem statement and contribution, followed by section 3, which covers related work on deep learning-based report generation. Section 4 provides a comprehensive methodology, while section 5 focuses on the experimental setup and evaluation. The paper concludes with section 6, which includes the conclusion and discussion, and finally, section 7 presents the references.

2. PROBLEM STATEMENT AND CONTRIBUTION

The increasing demand for accurate and timely radiology reports, coupled with the challenges radiologists face in examining medical images and creating diagnostic notes, has led to burnout, errors, and delays in providing care [11]. While experts have turned to artificial intelligence and deep learning (DL) technologies to automate the generation of radiology reports, implementing and adopting these technologies, face several challenges [12, 13]. These include the need to address concerns regarding the accuracy and reliability of automated notes, integrating these technologies into existing clinical workflows, and ensuring that they are accessible and affordable to all healthcare facilities. Addressing these challenges will be crucial in realizing the potential of AI and DL technologies to improve the speed, accuracy, and efficiency of radiology diagnoses, ultimately enhancing patient care outcomes. The problem statement is defined as follows: "*Considering the multimodal medical cohort containing radiology images with associated diagnostic notes, design and develop an automatic diagnostic report generation by analyzing the visual features from the Chest X-ray scans*".

We propose a solution to the challenge by developing a deep encoder-decoder model that can automatically generate reports from chest X-rays. To achieve this, we have utilized a Multi-channel dilation layer with Depthwise Separable Convolution Neural Network to extract imaging features and knowledge-based text modelling for textual feature extraction. Finally, the Long Short-Term Memory (LSTM) model is used to fine-tune the generated report. We summarize the contributions of this study as follows:

- We propose an encoder-decoder-based deep learning framework to generate diagnostic radiology reports for given chest x-ray images.
- We have developed a dynamic web portal that can efficiently take in chest X-ray images as input and generate radiology reports as output, thereby providing an accessible and user-friendly solution.
- We conduct a comprehensive analysis and compare the performance of the proposed model with the state-of-the-art deep learning approaches.

3. RELATED WORKS

Considerable progress has been made in the field of generating medical descriptions. Yuan et al. [14] introduced an automatic report generation model that utilizes a multiview CNN encoder and a conceptenriched hierarchical LSTM. The model leverages multi-view information in radiology by employing visual attention in a late fusion manner and enriches the semantics in the hierarchical LSTM decoder with medical concepts. Nguyen et al. [15], presented a set of three modules consisting of classification, generation, and interpretation. For the classification module, a multi-view encoder is employed to extract visual features from chest X-rays, while a text encoder converts reports into embeddings. The generation module utilizes both visual and textual features to create text on a word-by-word basis. Finally, the interpretation module fine-tunes the text generated. Tripathy et al. [16] showcased an automatic report generation model with following stages- NLP Pipeline: (Tokenization, Embedding, Removing special characters etc.); CNN: acts as an encoder in our model. A transfer Learning Model: ChexNet is used to extract the features of the image. Hierarchical LSTMs and Co-Attention mechanism: Hierarchical LSTMs are designed to enrich the representation ability of the LSTM, and the co-attention mechanism provides the context. The sentence and word LSTMs then generate the final reports required. Zhou et al. [17] presented a visual-textual attentive semantic model which uses DenseNet201 as a visual encoder model and BioSentVec as a text encoder. The LSTM model is utilized to generate the report.

A Knowledge Graph Auto-Encoder (KGAE) model is proposed by Liu et al. [18], which utilizes independent sets of Chest X-ray images and their associated reports during the training phase. KGAE consists of a pre-constructed knowledge graph, a knowledge-driven encoder and a knowledge-driven decoder. They have used the knowledge-driven encoder to project medical images and reports to the corresponding coordinates in latent space and the knowledge-driven decoder to generate a medical report on a given coordinate in that space. Sirshar et al. [19] propose an encoder-decoder model with CNN used as a visual encoder and an RNN decoder with attention used to produce the radiology reports. Nicolson et al. [20] presented the report generation framework, where the DenseNet pretrained on imageNet is used as an encoder for imaging feature extraction, and the Bidirectional Encoder Representations from Transformers (BERT) NLP encoder is utilized for textual feature extraction. The decoder model with attention is incorporated for report generation. The various general domain and domain-specific pre-trained checkpoints are evaluated and the best checkpoints are chosen for warm starting the encoder-decoder of a CXR report generator. These warm starting helps generate a diagnostically accurate report that can be used in a clinical setting. From the literature it is observed that there is a significant need for improving performance and the quality of the generated report.

The research paper introduced a deep learning model called CDGPT2, which aimed to generate radiology reports based on chest X-rays sourced from the Indiana University dataset [21]. To extract both visual and textual features, the model incorporated pre-trained Chexnet and ChatGPT2. However, the study identified limitations in the model's performance attributed to the small size of the available data. In a separate investigation, Babar et al. [22] introduced a novel metric known as Diagnostic Content Score (DCS). They initially created Diagnostic Tags for each report, leveraging them as external knowledge. By utilizing these tags, they developed a probabilistic model based on the training data. Subsequently, the model was employed to assess the diagnostic quality of the generated reports from the test data. However, the approach exhibited limitations, as indicated by a reduced Bleu4 score of 0.12 on the Indiana University dataset.

The accurate interpretation and summary of medical images, particularly those generated by radiology tests such as X-rays, CT scans, and MRIs, are crucial components of clinical diagnosis. Generating a diagnosis report from radiology images is an essential step in clinical diagnosis, and highly skilled radiologists are required for this task. However, the process can be time-consuming and mentally taxing for radiologists, especially in busy and overcrowded situations. To alleviate this burden and speed up the diagnosis process, there is a growing need for automated and reliable diagnostic report generation frameworks. Existing deep learning techniques for report generation have shown promise, but there is still room for improvement, particularly in terms of the BLEU score [14-22]. One promising approach is to develop a cross-modal framework that combines textual and imaging features to assist radiologists in automatically generating accurate reports from medical images. The proposed cross-modal framework leverages the knowledge base to extract textual features and incorporates multi-scale feature extraction from chest X-ray images. This approach facilitates the extraction of highly discriminative features, resulting in enhanced performance compared to existing methodologies. By using such frameworks, healthcare providers can reduce the workload on radiologists, speed up the diagnosis process, and provide better patient care. Additionally, these frameworks can ensure consistency and accuracy in diagnosis reports, minimizing the risk of errors and improving the overall quality of patient care.

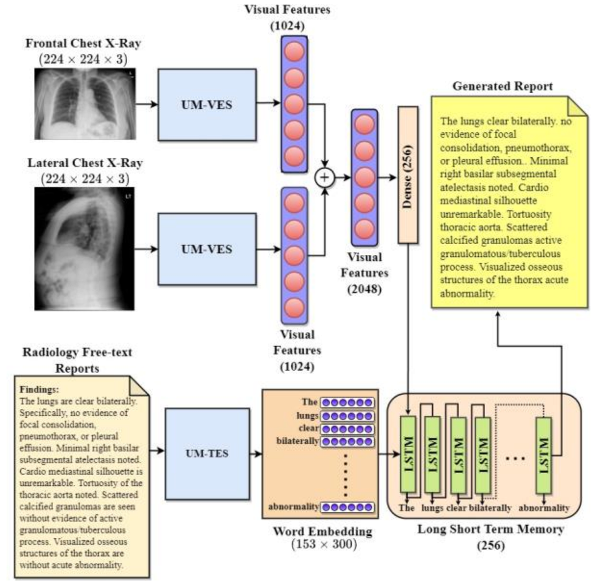
4. METHODOLOGY

The proposed encoder-decoder framework aims to generate radiology reports from chest X-rays, which include both frontal and lateral images. During the training process, both the chest X-rays and the corresponding reports are provided as input to the encoder. The encoder consists of two components: the Unimodal Medical Visual Encoding Subnetwork (UM-

VES) for extracting visual features and the Unimodal Medical Text Embedding Subnetwork (UM-TES) for extracting textual features. These features are then used by the LSTM-based decoder to generate the reports. The encoder operates by processing each item in the input sequence and aggregating the captured information into a context vector. Once the entire input sequence has been processed, the encoder transfers the context vector to the decoder, which generates the output sequence item by item. This process allows the model to effectively combine the visual and textual information and generate contextually relevant reports.

The detailed architecture of the proposed cross-modal retrieval is shown in Figure 2. During the training phase, the model aims to establish connections between the textual information in the reports and the visual features extracted from the chest X-ray images. The UM-TES approach is employed to encode the textual information, while the UM-VES technique is used to extract visual features. These modalities are then integrated into a joint representation, enabling the model to learn the correlations between the input chest X-ray images and the associated textual information in the reports. By iteratively optimizing the model's parameters, it gradually acquires the capability to generate coherent and contextually relevant reports.

During the testing phase, only the chest X-ray images are provided as input to the trained model. Drawing upon the learned associations between the image and the textual information from the training phase, the model generates a report based solely on the input image. This process is achieved by utilizing the decoding mechanism of the trained model, such as a Long Short-Term Memory (LSTM), to generate the text-based output.



The lungs clear bilaterally. no evidence of focal consolidation, pneumothorax, or pleural effusion.. Minimal right basilar subsegmental atelectasis noted. Cardio mediastinal silhouette unremarkable. Tortuosity thoracic aorta. Scattered calcified granulomas active granulomatous/tuberculous process. Visualized osseous structures of the thorax acute abnormality.

Findings:
The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion. Minimal right basilar subsegmental atelectasis noted. Cardio mediastinal silhouette is unremarkable. Tortuosity of the thoracic aorta noted. Scattered calcified granulomas are seen without evidence of active granulomatous/tuberculous process. Visualized osseous structures of the thorax are without acute abnormality.

Figure 2. Overall architecture of the proposed cross-modal deep learning-based model for automatic report generation

4. 1. Unimodal Medical Visual Encoding Subnetwork (UM-VES) The UM-VES technique suggested employs a depthwise separable convolution neural network with a multichannel dilation layer to extract imaging features. The suggested multichannel dilation convolution layer provides more comprehensive imaging data by generating a larger receptive field, while keeping the network parameters constant, in contrast to the traditional convolutional layer. Moreover, to ensure an even distribution of computational workload across each layer, the Depthwise Separable convolution network is utilized instead of the conventional convolution network [23]. The UM-VES framework is used to extract visual features from both the frontal and lateral CXR images independently, and the resulting features are combined by concatenation. The overall architecture of the proposed UM-VES model is shown in Figure 3. The UM-VES model is composed of three parallel dilation channels that capture imaging features with a wider receptive field. The resulting features are then concatenated and passed through 13 depthwise separable layers to learn and extract additional features. For a more comprehensive understanding of the model, readers can refer to our previous paper [23], where we provide a detailed overview and description of the UM-VES model's architecture and components.

4. 2. Unimodal Medical Text Embedding Subnetwork (UM-TES) The radiology findings are subjected to pre-processing, which involves removing stop words and punctuation, as well as performing stemming to retain root words. Additionally, tokenization is applied to extract important latent medical concepts. Customized Clinical Knowledge-based Text Modelling is utilized to learn word embeddings from medical terminology. During the training of the text model, the glove word embeddings [24] are combined with the word embeddings obtained from a knowledge base of 4.5 million Stanford reports [25, 26]. This combination enhances the effectiveness of the text model by leveraging the information contained in the knowledge base. The dense word embeddings generated are then mapped to medical terms from the findings using the Embedding Layer. The detailed architecture of the proposed UM-TES model is shown in Figure 4. To gain a more thorough understanding of the UM-TES model's architecture and components, readers can refer to our previous paper [23], where we offer a detailed overview and description.

4. 3. Long Short-term Memory-based Report Generation The fundamental concept of utilizing LSTM for report generation centers around the memory cell, denoted as c, which primarily stores the information on the input received at any given moment. The function of these cells is controlled by layers or gates that are inserted in a multiplicative manner and can maintain values of either 0 or 1 which are determined by the gates.

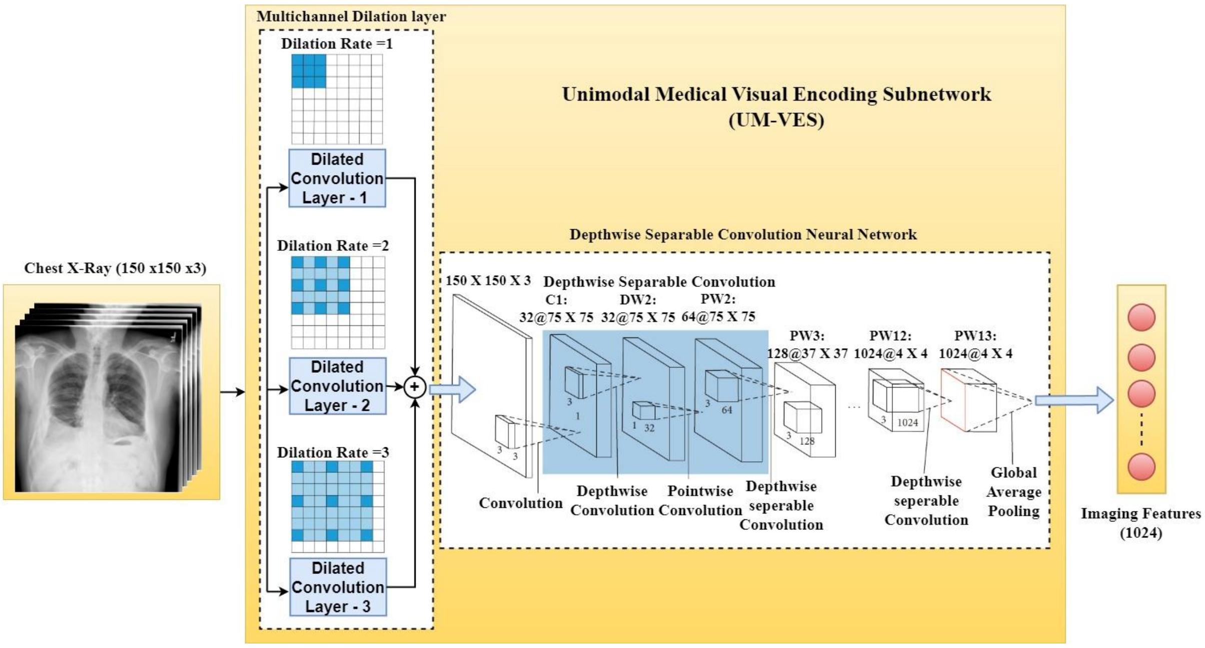

Multichannel Dilation layer
Dilation Rate =1
Dilated Convolution Layer - 1
Dilation Rate =2
Dilated Convolution Layer - 2
Dilation Rate =3
Dilated Convolution Layer - 3
Unimodal Medical Visual Encoding Subnetwork (UM-VES)
Chest X-Ray (150 x150 x3)
Depthwise Separable Convolution Neural Network
150 X 150 X 3
Depthwise Separable Convolution
C1: 32@75 X 75
DW2: 32@75 X 75
PW2: 64@75 X 75
PW3: 128@37 X 37
PW12: 1024@4 X 4
PW13: 1024@4 X 4
Convolution
Depthwise Convolution
Pointwise Convolution
Depthwise seperable Convolution
Depthwise seperable Convolution
Global Average Pooling
Imaging Features (1024)

Figure 3. Overall architecture of the UM-VES model

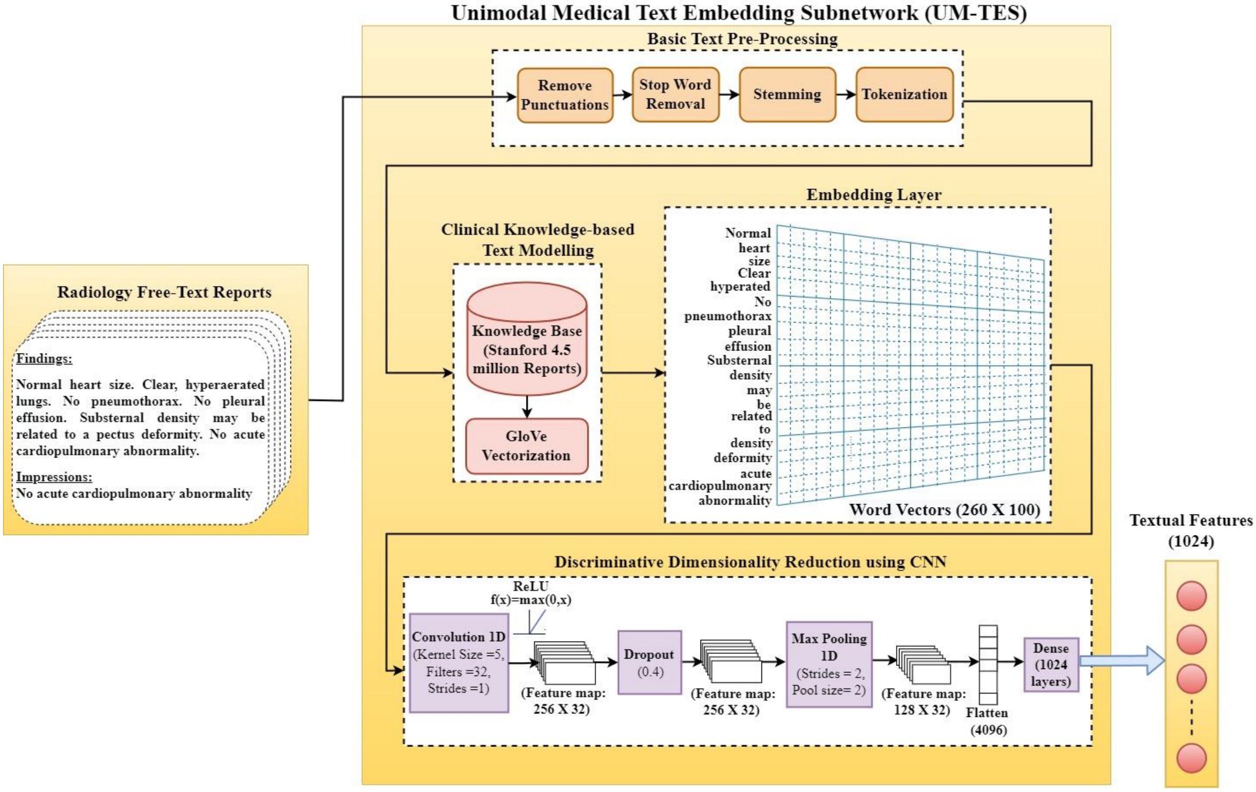

Unimodal Medical Text Embedding Subnetwork (UM-TES)
Basic Text Pre-Processing
Remove Punctuations
Stop Word Removal
Stemming
Tokenization
Radiology Free-Text Reports
Findings:
Normal heart size. Clear, hyperaerated lungs. No pneumothorax. No pleural effusion. Substernal density may be related to a pectus deformity. No acute cardiopulmonary abnormality.
Impressions:
No acute cardiopulmonary abnormality
Clinical Knowledge-based Text Modelling
Knowledge Base (Stanford 4.5 million Reports)
GloVe Vectorization
Embedding Layer
Normal heart size Clear hyperated No pneumothorax pleural effusion Substernal density may be related to density deformity acute cardiopulmonary abnormality
Word Vectors (260 X 100)
Discriminative Dimensionality Reduction using CNN
ReLU f(x)=max(0,x)
Convolution 1D (Kernel Size =5, Filters =32, Strides =1)
(Feature map: 256 X 32)
Dropout (0.4)
(Feature map: 256 X 32)
Max Pooling 1D (Strides = 2, Pool size= 2)
(Feature map: 128 X 32)
Flatten (4096)
Dense (1024 layers)
Textual Features (1024)

Figure 4. Overall architecture of the UM-TES model

Specifically, three gates are employed to monitor whether the present value of the cell should be disregarded, if the new cell value should be generated (output gate 0), or if it should be interpreted as input, as illustrated in Figure 5.

Equations (1), (2) and (3) depicts the input, forget, and output layers, respectively.

$$input_t = \sigma(W_{iy}y_t + W_{im}m_{t-1}) \quad (1)$$

$$forget_t = \sigma(W_{fy}y_t + W_{fm}m_{t-1}) \quad (2)$$

$$output_t = \sigma(W_{oy}y_t + W_{om}m_{t-1}) \quad (3)$$

Equations (4), (5) and (6) represent the other operation of the LSTM model.

$$cell_t = forget_t \odot cell_{t-1} + input_t \odot h(W_{cy}y_t + W_{cm}m_{t-1}) \quad (4)$$

$$cell_t = output_t \odot cell_t \quad (5)$$

$$P_{t+1} = Softmax(m_t) \quad (6)$$

where, $input_t$, $forget_t$ and $output_t$ denotes the output of the input, forget and output gates, respectively at time t; y_t represents the input vector at time t; $m_t - 1$ is the hidden state of the LSTM at time $t-1$; W_{iy}, W_{fy}, W_{oy}, W_{cy}, W_{im}, W_{fm}, W_{om} and W_{cm} indicate the weight matrices that manage that manage the input and hidden connections between the input, forget and output gates and cells; $cell_t$ represents the state of the cell at time t and P_{t+1} represents a probability distribution over a set of possible outcomes at time $t+1$.

4. 4. Web-based Framework for Report Generation

We utilized the Flask web framework to create a user-friendly web interface for our model. By uploading both frontal and lateral X-ray images through this interface, users can obtain reports with ease. To streamline the user experience, we implemented Ajax, a technique that enables data to be sent and retrieved asynchronously in the background of the application without requiring the entire page to be reloaded. This approach is particularly useful when we want to update specific portions of an existing page without redirecting or reloading the page for the user. As depicted in Figure 6, in order to obtain a report, users are required to upload both frontal and lateral X-ray images. After clicking on the 'Generate Report' button, an Ajax request is sent to the Flask App hosted on the server. The Flask application utilizes the uploaded images to generate predictions for the report, which are then transmitted back to the client side. Upon receipt, the predicted report is displayed to the users.

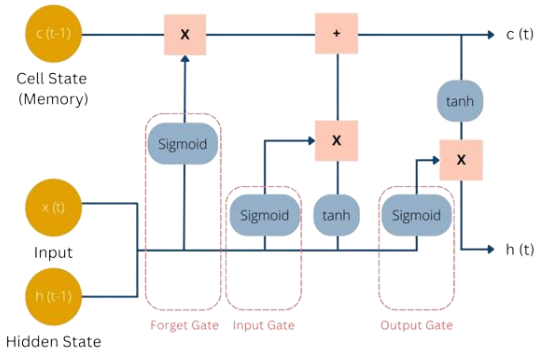


Figure 5. Long Short-term Memory Architecture

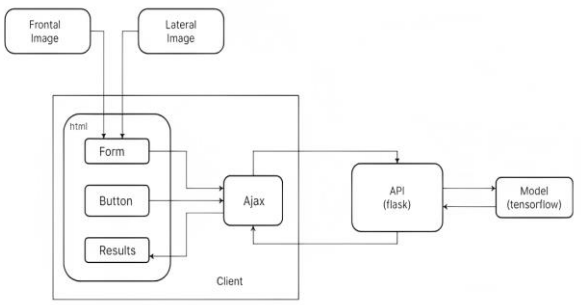


Figure 6. Client-Server interaction used for predicting reports

5. EXPERIMENTAL SETUP AND EVALUATION

For model training, we utilized the IU Chest X-Ray collection [27], which includes a comprehensive set of chest x-ray images accompanied by their corresponding diagnostic reports. The cohort of multimodal medical data consisted of 7,470 pairs of images and reports, with a total of 3,996 cases. The reports contained two main sections, impressions and findings. In our investigation, we selected frontal and lateral images and the content of the findings section as the target captions to be generated. To conduct our experiment, we removed cases without reports and frontal/lateral images, ultimately working with 3,638 cases. Two methods were used to generate text reports: greedy search [28] and beam search [29]. Greedy search is an algorithmic approach that incrementally constructs a solution by selecting the next piece that seems to provide the most immediate benefit. In contrast, beam search expands on the greedy search technique by generating a list of the most likely output sequences, each with its own score. The sequence with the highest score is then chosen as the final result.

To evaluate the performance of the generated reports, we incorporated the BLEU score [30]. The Bilingual Evaluation Understudy (BLEU) Score is a method used to evaluate the similarity between a generated sentence and a reference sentence. The score ranges from 0.0, indicating a total mismatch, to 1.0, indicating a perfect match. This approach involves tallying the number of matching n-grams in the candidate text with those in the reference text. For instance, a uni-gram or 1-gram would correspond to each token, whereas a bi-gram comparison would correspond to each pair of words. Achieving a perfect score is not practical, as it necessitates an exact match with the reference, which even human translators cannot achieve. Furthermore, comparing scores across datasets can be difficult due to the number and quality of the references used to determine the BLEU score.

We computed the BLEU score for an automatic report generated using beam and greedy search. It is observed that beam search produces a superior BLEU score compared to the greedy search algorithm. The qualitative analysis of the proposed deep learning-based model using a beam and greedy search algorithm is shown in the Table 1. The BLEU score of 0.5459, 0.4131, 0.386 and 0.3552 is obtained for different n-grams in the greedy search approach. The beam search approach produces a BLEU score of 0.5881, 0.4325, 0.4017 and 0.3860. We have also compared the results with the existing automatic diagnostic report generation work. Most of the existing work has shown lesser BLEU4 as it compares the four words together with the ground truth. Our proposed model outperforms the existing models while generating robust diagnostic reports. This may be due to the multi-channel visual features and knowledge-based discriminate text features extracted in the encoder of the proposed network. The detailed analysis with the various existing model is summarized in Table 2.

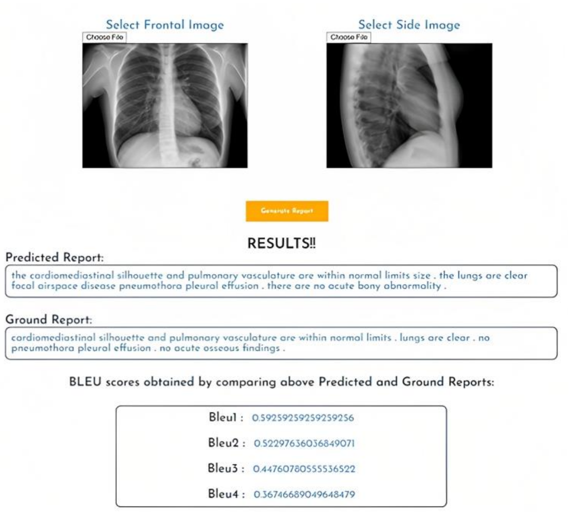
We designed and developed a flask web application interface for quantitative analysis of the model. Figure 7 shows the web interface to upload the chest x-ray images and produce the diagnostic report. The user has to input frontal and lateral chest X-ray images to the web interface. When the user clicks the "generate report", an Ajax request will be sent to the Flask App on the server, where the Flask application uses the uploaded images to predict reports. The predicted reports will be sent back to the client, where they are displayed to the users with the BLEU score. Figures 8 and 9 present two samples of reports generated using the proposed framework.

TABLE 1. Performance analysis of the proposed model

Method	Bleu1	Bleu2	Bleu3	Bleu4
Greedy Search	0.5459	0.4131	0.3864	0.3552
Beam Search	0.5881	0.4325	0.4017	0.3860

TABLE 2. Performance analysis compared with existing work of report generation

Method	Bleu1	Bleu2	Bleu3	Bleu4
Tripathy et al. [16], 2021	0.213	0.258	0.325	0.381
Nguyen et al. [15], 2021	0.515	0.378	0.293	0.235
Liu et al. [18], 2021	0.417	0.263	0.181	0.126
Zhou et al. [17], 2021	0.536	0.392	0.314	0.339
Sirshar et al. [19], 2022	0.58	0.342	0.263	0.155
Nicolson et al. [20], 2022	0.4777	0.308	0.2274	0.1773
Proposed Model	**0.5881**	**0.4325**	**0.4017**	**0.3860**



RESULTS!!

Predicted Report:

the cardiomediastinal silhouette and pulmonary vasculature are within normal limits size . the lungs are clear focal airspace disease pneumothora pleural effusion . there are no acute bony abnormality .

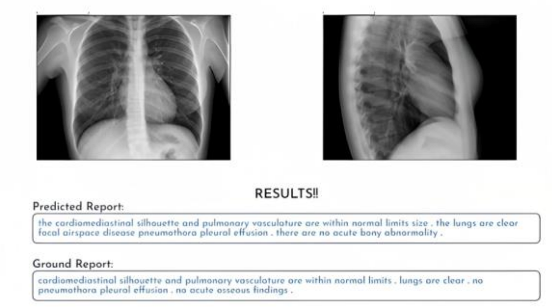
Ground Report:

cardiomediastinal silhouette and pulmonary vasculature are within normal limits . lungs are clear . no pneumothora pleural effusion . no acute osseous findings .

BLEU scores obtained by comparing above Predicted and Ground Reports:

Bleu1 : 0.59259259259259256
Bleu2 : 0.52297636036849071
Bleu3 : 0.44760780555536522
Bleu4 : 0.36746689049648479

Figure 7. The dynamic web portal for automatic diagnostic report generation



RESULTS!!

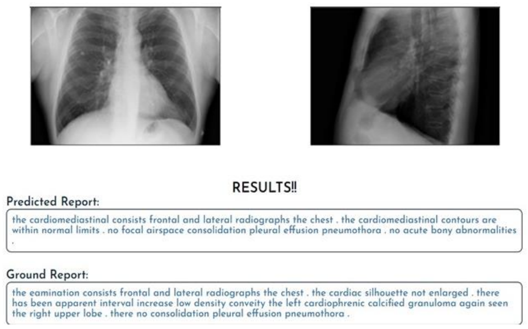
Predicted Report:

the cardiomediastinal silhouette and pulmonary vasculature are within normal limits size . the lungs are clear focal airspace disease pneumothora pleural effusion . there are no acute bony abnormality .

Ground Report:

cardiomediastinal silhouette and pulmonary vasculature are within normal limits . lungs are clear . no pneumothora pleural effusion . no acute osseous findings .

Figure 8. Generated Report (Sample 1)



RESULTS!!

Predicted Report:

the cardiomediastinal consists frontal and lateral radiographs the chest . the cardiomediastinal contours are within normal limits . no focal airspace consolidation pleural effusion pneumothora . no acute bony abnormalities .

Ground Report:

the eamination consists frontal and lateral radiographs the chest . the cardiac silhouette not enlarged . there has been apparent interval increase low density conveity the left cardiophrenic calcified granuloma again seen the right upper lobe . there no consolidation pleural effusion pneumothora .

Figure 9. Generated Report (Sample 2)

In summary, an automated framework that employs a deep learning-based encoder-decoder approach to generate reports from chest X-ray scans. The modules used in the framework, such as UM-VES, UM-TES, and LSTM, are discussed in detail. In addition, a dynamic web framework was developed and implemented that accepts chest X-ray images as input and generates diagnostic reports as output. To evaluate the proposed framework, a comprehensive set of experiments was conducted, and the results were compared with those of state-of-the-art report generation frameworks. The proposed framework yielded better performance, as evidenced by an improved BLEU score compared to existing models.

6. CONCLUSION AND FUTURE WORK

In this paper, we aimed to develop a deep learning-based model that can accurately and automatically generate diagnostic reports from CXR images. To achieve this, we employed a cross-modal retrieval technique that retrieves radiology reports from the image. Our approach, which utilized the beam search method, outperformed existing models in generating robust diagnostic reports. This can be attributed to the encoder of our proposed network, which extracted multi-channel visual features and discriminative text features based on knowledge. Compared to existing models, our approach showed superior results in terms of BLEU4 scores, which is a standard metric used to compare the accuracy of generated text to the ground truth. In addition, we created a dynamic web portal that allows for the easy uploading of frontal and lateral CXR images, and provides the corresponding diagnostic reports as output. This feature greatly simplifies the report writing process for radiologists, as it automates the process and saves time.

One potential limitation of the proposed work is the need to assess the generalizability of the model. While the deep learning framework exhibited notable enhancements in generating accurate and reliable radiology reports compared to existing models, it is crucial to recognize that the evaluation and analysis were restricted to the Indiana University Dataset. The performance of the model may differ when applied to alternative datasets or diverse clinical settings. Therefore, additional research and validation on varied datasets and real-world scenarios are imperative to determine the generalizability and robustness of the proposed approach.

Furthermore, we plan to expand the scope of our model by applying it to other types of diagnostic images such as MRI, ultrasound and CT scans. This will allow us to further evaluate the effectiveness and robustness of our proposed model across different modalities and ultimately improve its overall utility in clinical settings.

7. REFERENCES

1. Shetty, S. and Mahale, A., "Comprehensive review of multimodal medical data analysis: Open issues and future research directions", ***Acta Informatica Pragensia***, Vol. 11, No. 3, (2022), 423-457. doi: 10.18267/j.aip.202.
2. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G. and Murphy, K., "Deep learning for chest x-ray analysis: A survey", ***Medical Image Analysis***, Vol. 72, (2021), 102125. https://doi.org10.1016/j.media.2021.102125
3. Ramirez-Alonso, G., Prieto-Ordaz, O., López-Santillan, R. and Montes-Y-Gómez, M., "Medical report generation through radiology images: An overview", ***IEEE Latin America Transactions***, Vol. 20, No. 6, (2022), 986-999. doi: 10.1109/TLA.2022.9757742.
4. Monshi, M.M.A., Poon, J. and Chung, V., "Deep learning in generating radiology reports: A survey", ***Artificial Intelligence in Medicine***, Vol. 106, (2020), 101878. https://doi.org/10.1016/j.artmed.2020.101878
5. Jing, B., Xie, P. and Xing, E., "On the automatic generation of medical imaging reports", ***arXiv preprint arXiv:1711.08195***, (2017). doi: 10.18653/v1/P18-1240.
6. Dey, A., "Cov-xdcnn: Deep learning model with external filter for detecting covid-19 on chest x-rays", in Computer, Communication, and Signal Processing: 6th IFIP TC 5 International Conference, ICCCSP 2022, Chennai, India, February 24–25, 2022, Revised Selected Papers, Springer. (2022), 174-189.
7. Shetty, S. and Mahale, A., "Ms-chexnet: An explainable and lightweight multi-scale dilated network with depthwise separable convolution for prediction of pulmonary abnormalities in chest radiographs", ***Mathematics***, Vol. 10, No. 19, (2022), 3646. https://doi.org/10.3390/math10193646
8. Alahmari, S.S., Altazi, B., Hwang, J., Hawkins, S. and Salem, T., "A comprehensive review of deep learning-based methods for covid-19 detection using chest x-ray images", ***IEEE Access***, (2022). doi: 10.1109/ACCESS.2022.3208138.
9. Yang, S., Wu, X., Ge, S., Zhou, S.K. and Xiao, L., "Knowledge matters: Chest radiology report generation with general and specific knowledge", ***Medical Image Analysis***, Vol. 80, (2022), 102510. https://doi.org/10.1016/j.media.2022.102510
10. Carson, E.R., Cramp, D.G., Morgan, A. and Roudsari, A.V., "Clinical decision support, systems methodology, and telemedicine: Their role in the management of chronic disease", ***IEEE Transactions on Information Technology in Biomedicine***, Vol. 2, No. 2, (1998), 80-88. doi: 10.1109/4233.720526.
11. Abtahi, Z., Sahraeian, R. and Rahmani, D., "A stochastic model for prioritized outpatient scheduling in a radiology center", ***International Journal of Engineering Transactions A: Basics***, Vol. 33, No. 4, (2020). doi: 10.5829/ije.2020.33.04a.11.
12. Khatami, A., Babaie, M., Tizhoosh, H., Nazari, A., Khosravi, A. and Nahavandi, S., "A radon-based convolutional neural network for medical image retrieval", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 31, No. 6, (2018), 910-915. doi: 10.5829/ije.2018.31.06c.07.
13. Gheitasi, A., Farsi, H. and Mohamadzadeh, S., "Estimation of hand skeletal postures by using deep convolutional neural networks", ***International Journal of Engineering, Transactions A: Basics***, Vol. 33, No. 4, (2020), 552-559. doi: 10.5829/ije.2020.33.04a.06.
14. Yuan, J., Liao, H., Luo, R. and Luo, J., "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment", in Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, Springer. (2019), 721-729.
15. Nguyen, H.T., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J. and Cheng, L., "Automated generation of accurate\& fluent medical x-ray reports", arXiv preprint arXiv:2108.12126, (2021). doi: 10.18653/v1/2021.emnlp-main.288.
16. Tripathy, B., Sai, R.R. and Banu, K.S., Automated medical report generation on chest x-ray: Images using co-attention mechanism,

in Hybrid computational intelligent systems. 2023, CRC Press.111-122.

17. Zhou, X., Li, Y. and Liang, W., "Cnn-rnn based intelligent recommendation for online medical pre-diagnosis support", ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, Vol. 18, No. 3, (2020), 912-921. doi: 10.1109/TCBB.2020.2994780.
18. Liu, F., You, C., Wu, X., Ge, S. and Sun, X., "Auto-encoding knowledge graph for unsupervised medical report generation", ***Advances in Neural Information Processing Systems***, Vol. 34, (2021), 16266-16279. https://arxiv.org/abs/2111.04318
19. Sirshar, M., Paracha, M.F.K., Akram, M.U., Alghamdi, N.S., Zaidi, S.Z.Y. and Fatima, T., "Attention based automated radiology report generation using cnn and lstm", ***Plos one***, Vol. 17, No. 1, (2022), e0262209. doi: 10.1371/journal.pone.0262209.
20. Nicolson, A., Dowling, J. and Koopman, B., "Improving chest x-ray report generation by leveraging warm-starting", arXiv preprint arXiv:2201.09405, (2022). https://arxiv.org/abs/2201.09405
21. Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M. and Fahmy, A., "Automated radiology report generation using conditioned transformers", ***Informatics in Medicine Unlocked***, Vol. 24, (2021), 100557. doi: 10.1016/j.imu.2021.100557.
22. Babar, Z., van Laarhoven, T., Zanzotto, F.M. and Marchiori, E., "Evaluating diagnostic content of ai-generated radiology reports of chest x-rays", ***Artificial Intelligence in Medicine***, Vol. 116, (2021), 102075. https://doi.org10.1016/j.artmed.2021.102075
23. Shetty, S. and Mahale, A., "Multimodal medical tensor fusion network-based dl framework for abnormality prediction from the radiology cxrs and clinical text reports", ***Multimedia Tools and Applications***, (2023), 1-48. https://doi.org/10.1007/s11042-023-14940-x.
24. Pennington, J., Socher, R. and Manning, C.D., "Glove: Global vectors for word representation", in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). (2014), 1532-1543.
25. Zhang, Y., Ding, D.Y., Qian, T., Manning, C.D. and Langlotz, C.P., "Learning to summarize radiology findings", arXiv preprint arXiv:1809.04698, (2018). http://arxiv.org/abs/1809.04698
26. Shetty, S., Ananthanarayana, V. and Mahale, A., "Medical knowledge-based deep learning framework for disease prediction on unstructured radiology free-text reports under low data condition", in Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference: Proceedings of the EANN 2020 21, Springer. (2020), 352-364.
27. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R. and McDonald, C.J., "Preparing a collection of radiology examinations for distribution and retrieval", ***Journal of the American Medical Informatics Association***, Vol. 23, No. 2, (2016), 304-310. doi: 10.1093/jamia/ocv080.
28. Gu, J., Cho, K. and Li, V.O., "Trainable greedy decoding for neural machine translation", arXiv preprint arXiv:1702.02429, (2017). doi: 10.18653/v1/D17-1210.
29. Freitag, M. and Al-Onaizan, Y., "Beam search strategies for neural machine translation", arXiv preprint arXiv:1702.01806, (2017). https://doi.org/10.48550/arXiv.1702.01806
30. Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J., "Bleu: A method for automatic evaluation of machine translation", in Proceedings of the 40th annual meeting of the Association for Computational Linguistics. (2002), 311-318.

Persian Abstract

چکیده

تولید گزارش رادیولوژی یک وظیفه حیاتی برای رادیولوژیست ها است و خودکار کردن این فرآیند می تواند حجم کار آنها را به میزان قابل توجهی ساده کند. با این حال، ایجاد گزارش های رادیولوژی دقیق و قابل اعتماد مستلزم آن است که رادیولوژیست ها تجربه و زمان کافی برای بررسی تصاویر پزشکی داشته باشند. متأسفانه، بسیاری از گزارش‌های رادیولوژی با نتیجه‌گیری‌های مبهم خاتمه می‌یابند که منجر به آزمایش‌های اضافی و روش‌های تشخیصی برای بیماران می‌شود. برای پرداختن به این موضوع، ما یک چارچوب یادگیری عمیق مبتنی بر رمزگذار-رمزگشا پیشنهاد کردیم که از تصاویر اشعه ایکس قفسه سینه برای تولید گزارش‌های رادیولوژی تشخیصی استفاده می‌کند. در مطالعه خود، یک مدل سازی متن جدید و استراتژی استخراج ویژگی بصری را به عنوان بخشی از چارچوب یادگیری عمیق مبتنی بر رمزگذار-رمزگشای پیشنهادی خود معرفی کرده‌ایم. هدف رویکرد ما استخراج اطلاعات بصری و متنی ضروری از تصاویر اشعه ایکس قفسه سینه برای تولید گزارش‌های رادیولوژی دقیق‌تر و قابل اعتمادتر است. علاوه بر این، ما یک پورتال وب پویا ایجاد کرده ایم که اشعه ایکس قفسه سینه را به عنوان ورودی می پذیرد و گزارش رادیولوژی را به عنوان خروجی تولید می کند. ما تجزیه و تحلیل گسترده ای از مدل خود انجام دادیم و عملکرد آن را با دیگر رویکردهای پیشرفته یادگیری عمیق مقایسه کردیم. یافته‌های ما نشان‌دهنده بهبود قابل‌توجهی است که توسط مدل پیشنهادی ما در مقایسه با مدل‌های موجود به دست آمده است، همانطور که با نمرات BLEU بالاتر (BLEU1 = 0.588، BLEU2 = 0.4325، BLEU3 = 0.4017، BLEU4 = 0.3860) به دست آمده در مجموعه داده‌های دانشگاه ایندیانا مشهود است. این نتایج بر پتانسیل چارچوب یادگیری عمیق ما برای افزایش دقت و قابلیت اطمینان گزارش های رادیولوژی تاکید می کند که منجر به درمان پزشکی کارآمدتر و موثرتر می شود.

International Journal of Engineering

Journal Homepage: www.ije.ir

A Novel Approach to Modular Control of Highway and Arterial Networks using Petri Nets Modeling

M. Mohammadi[a], A. Dideban*[a], B. Moshiri[b]

[a] *Control Engineering, ECE Faculty, Semnan University, Semnan, Iran*
[b] *School of Electrical and Computer Engineering, University College of Engineering, University of Tehran, Tehran, Iran*

PAPER INFO

Paper history:
Received 13 March 2023
Received in revised form 32 May 2023
Accepted 20 June 2023

Keywords:
Smart Optimization
Modular Model
Petri Nets
Integrated Highway
Arterial Control

ABSTRACT

In this paper, integrated control of highways and intersections is investigated. A modular Petri-Net-based framework is implemented to model the traffic flow of highway and arterial traffic network systems. In this framework, arterial intersection traffic lights are modeled by Timed Petri Nets (TPN). The timing of traffic lights and variable speed limits on the highway is managed to be optimized using an intelligent algorithm. This algorithm provides a trade-off between the length of the queue of vehicles on the highway and the entrance ramp and the length of the queue at the intersection after each time cycle. The performance of the optimized traffic controller and the fixed control were compared. The simulation results verify that the use of optimization methods to manage the timing of traffic lights in intersections and speed limitation in highways can considerably improve traffic flow in special conditions such as rainy weather and accidents. Additionally, this method can considerably enhance traffic flow in normal hours, while in rush hours and midnight, such improvement is negligible.

***doi**: 10.5829/ije.2023.36.08b.17*

1. INTRODUCTION

Nowadays, the urban transportation sector is one of the major emitters of greenhouse gases. The number of vehicles traveling in urban areas is increasing, especially in metropolitan areas, due to the rapid growth of urban areas. This incremental growth in the number of on-road vehicles has led to an increasing number of undesirable drawbacks such as heavy traffic, air and noise pollution, road congestion, and waste of citizens' useful time which will indeed lead to a higher citizen dissatisfaction rate. A sign of these drawbacks may be seen in the growing number of research in this field [1].

Traffic control is considered to be a very beneficial and cost-effective approach to overcome the aforementioned problems in comparison to infrastructural development. Designing a solution for improved traffic control performance is conducted through two different approaches in the literature: Highway Networks Control and Arterial Control, with the latter tending to manage the control of intersections and urban arteries. The highway network control approach imposes restrictions on highway entrances. Speed limits on the upstream will reduce congestion and facilitate traffic [2]. Furthermore, much research has been conducted to investigate the effectiveness of the input control method. The performance of these methods is evaluated using the comparison with local scheduling strategies [3]. These experiments demonstrate the superiority of this approach over local and no-control strategies. Different input control strategies have been implemented and tested in Paris and Amsterdam [4]. Taheri et al. [5] used queueing system analysis to provide an analytical method for calculating the fixed-time control system's average waiting time at a single isolated intersection.

Furthermore, the same methodology was applied to control the input as well as the speed limit simultaneously for one ramp and two-speed control panels. The coordinated feedback control of the input and the control of the mainstream traffic flow on the highways was applied using speed limitations [6]. Furthermore, a Resilliance Control for multi-lane highways in the presence of vehicle counting systems and prediction

*Corresponding Author Email: adideban@semnan.ac.ir (A. Dideban)

Please cite this article as: M. Mohammadi, A. Dideban, B. Moshiri, A Novel Approach to Modular Control of Highway and Arterial Networks using Petri Nets Modeling, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1578-1588

engine has been proposed by Mohammadi et al. [7].

In the arterial control approach, the most common way to regulate and manage urban traffic is to control the timing signal of traffic lights. Regarding work introduced by Fu and Chen [8], these functions are categorized into two groups; Fixed Time Control Systems and Traffic Responsive Control Systems. In the first category, pre-scheduled timing scenarios are executed using the offline optimization method. In the second category, the traffic control strategy is executed using optimized timing scenarios obtained from stimulated signals from traffic sensors and an online optimization method.

The origin-destination (O-D) pairs that colored Petri networks consume were analyzed by Fu and Chen [8], along with the latency that could be computed between Petri net nodes. Each node holds data about network subregional congestion, and colored tokens represent automobiles that move through the graph over time, following an O-D pair.

In addition to these studies, the utilization of Petri nets (PNs) in modeling, performance analysis, and control of traffic systems has been used for several years [9]. Additionally, a timed Petri net is utilized to model the timing maps of traffic lights controlling the intersections [5], implemented deterministic-timed PN (DTPN) for a microscopic model of a signalized traffic urban area, including signalized intersections and roads. This paper presents a Traffic Responsive Control System based on the colored timed Petri net (CTPN) model and Macroscopic models of traffic flow. Microscopic models are very detailed, and consequently, the computational effort can be exceptionally high when modeling large road networks. On the other hand, macroscopic models are less computationally intensive and can be used to model large road networks with an acceptable computational load. However, this computational advantage is balanced by their inability to capture some specific traffic phenomena related to the behavior of individual drivers. CTPN [3, 9, 10] describes traffic more finely: vehicles are individually shown by considering the interactions between them. By contrast, the macroscopic models represent the traffic flow with general variables such as the flow rate, flow density, and average flow speed. In the past literature, modeling is either considered microscopically for intersections or macroscopically for highways. At the same time, both models have been considered in the model proposed in this paper. The proposed control system aims to control the timing of traffic lights to reduce traffic jams in a certain part of the city. This urban area consists of three intersections and a highway [11], traffic light schemes with two extra traffic signals were defined using a Petri nets method. A warning on a particular road will be announced on one, and a road closure will be announced on the other. They made assumptions about the start and end dates of the strategy's operational period.

With respect to Fu et al. [12] Multi-Regional MFD Systems with Boundary Queues have been designed using Colored Petri Nets for both Perimeter Control and Route Guidance. The intersections and road segments that separate each pair of adjacent subregions are modeled as a buffer zone in this Accumulation-based traffic model using Petri Nets, which is introduced here as a reference for perimeter control. The proposed control framework incorporates both perimeter control and route guiding, taking into account the waiting cars in the buffer zones.

A large number of studies have been carried out to develop different signal timing plans, which are mainly classified into three classifications fixed-timed, predictive control strategies, and traffic responsive [13]. The first one is widely adopted in most current urban traffic systems due to their inexpensive and easy implementation. However, its disadvantage is that the time plan is fixed even in abnormal conditions. The second one is based on an optimization control strategy that can predict the future traffic behavior of the network based on traffic-forecasting models. The third one is based on those measured current traffic states and has been effectively used in many cities around the world. In this paper, the signal timing plan of the intersections and speed limitation in highways are determined so that traffic congestion is minimized. In other words, we deal with an optimization problem aiming to maximize the traffic flow using a signal timing plan of the intersections and speed limitation in different sections of the highway. The superiority of this method in comparison with other studies is modular traffic network modeling. In other words, any possible structure of the traffic network containing several intersections and highways can be created using this modular system. In contrast, other studies just considered either intersections or highways while, in a modular framework, the dimension of the traffic network becomes huge, and therefore, the computational cost is heavy. Additionally, in this modular framework, the performance of the proposed control system can be evaluated under various types of abnormal conditions such as rainy weather, accident occurrence, temporary obstruction, and any stochastic fluctuation in demand.

Bargegol et al. [13], investigated the correlation between speed, density, and flow in various conditions by analyzing the conduct of pedestrians while crossing both within and outside of crosswalks, and considering the duration of crossing during both pedestrian green and red times. The study employs linear and non-linear regression analysis to obtain these findings. In the work that follows, Petri nets and evolutionary algorithms are combined. Srivastava and Sahana [14] attempted to optimize waiting times in a city network model using this technique. In the study, the authors employed a hybrid technique that included ACO, GA, and the Stackelberg competition model to reduce the average waiting time in the tested network by 4% more than they could have with an evolutionary algorithm that had found the best solution on its own. Luo et al. [15] investigated the utilization of CTM for the purpose of devising control tactics to address traffic congestion resulting from dispersion accidents. Additionally, the effectiveness of said strategies is evaluated. The utilization of deep learning methodology for the identification of traffic congestion was suggested by Perez-Murueta and colleagues in 2019. The utilization of deep learning and k-shortest path algorithm was suggested by the authors for the purpose of identifying traffic congestion [16].

The paper is organized as follows: The next section describes the urban area in a modular representation. Section III represents the formulation of the problem as an optimization problem. In section IV, various traffic scenarios used to evaluate the proposed control system are described, and simulation results are presented under different scenarios. The conclusion of the paper is presented in the last section.

2. MODULAR MODELING OF THE HIGHWAY, INTERSECTIONS AND LINKS

Traffic flow can be modeled as a hybrid system characterized by continuous and discrete behaviors. The continuous behaviors of traffic flow use continuous Petri nets with variable speed (VCPN) to model and analyze the highways from a macroscopic point of view, whereas in discrete ones, colored timed Petri nets (TPN) are used to describe traffic flow in intersections from a microscopic point of view. VCPN [17, 18] describes traffic flow by global variables such as the flow rate, the flow density, and the average flow velocity. In contrast, TPN focuses on the individual vehicle's behaviors in the streets .

As mentioned before, VCPN can model the traffic flow on the highway, and TPN can be used to represent how the vehicles move in the intersections; therefore, in this section, the modular modeling of the highway and intersections using VCPN and TPN are presented, respectively.

2. 1. Highways Modelling by VCPNs Continuous Petri Nets with Variable Speed (VCPN) are proven to be suitable to model traffic flow with modular spatial discretization. VCPN model for a segment of highway is shown in Figure 1.

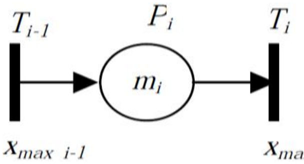


Figure 1. VCPN model for a segment of highway

Each place P_i corresponds to the segment $[x_{i-1}, x_i]$ whose length can be variable. The transition T_i represents the stream from one segment to another. The marking $m_i(t)$ of the place P_i stands for the number of vehicles $n_i(t)$ in the named segment. The transition firing speed $x_i(t)$ stands for the flow rate $q_i(t)$. C_i and Δ_i denote the capacity and length of segment i respectively. $v_{free\,i}$ and α_i are maximum velocity and firing rate of each segment. ρ is a continuous function denoting density. The relation between different variables in VCPN can be formulated as follows [19]:

$$\rho_i(t) = \frac{m_i(t)}{\Delta_i} \tag{1}$$

$$v_i(t) = \frac{x_i(t).\Delta_i}{m_i(t)} \tag{2}$$

$$\frac{d\rho_i(t)}{dt} = \frac{1}{\Delta_i}\frac{dm_i(t)}{dt} \tag{3}$$

$$\frac{dm_i(t)}{dt} = x_{i-1}(t) - x_i(t) \tag{4}$$

$$x_i(t) = x_{\max i}.\min\left(\alpha_i, m_i(t), C_{i+1} - m_{i+1}(t)\right) \tag{5}$$

$$x_{\max i} = \frac{v_{free\,i}}{\Delta_i} \tag{6}$$

$$\alpha_i = \frac{q_{\max i}.\Delta_i}{v_{free\,i}} \tag{7}$$

$$C_i = \rho_{\max i}.\Delta_i \tag{8}$$

Thus, a huge highway can be modeled. Every segment is characterized by its own density, velocity, capacity, and speed limitation. Therefore, traffic flow can be represented by VCPNs on highways.

2. 2. Intersection Modelling by Colored Timed Petri Nets In this colored Petri net modeling, four basic components are taken into consideration: signalized intersections, links, vehicles, and traffic lights. Each link shows the space between two adjacent intersections and can contain one or several lanes. That is, a signalized urban area consists of several intersections controlled by planned traffic lights and has a number of links gathered in the set $L=\{L_i | i=1,...,I\}$. The links are divided into three main categories: input, intermediate, and output links. Each general link with a pertained length has a limited capacity of vehicles $C_i > 0$. which denotes the number of Passenger Car Units that the link can handle at the same time. Therefore, each link may be divided into C_i cells based on unit capacity. It is also necessary to consider the physical space which each vehicle occupies while crossing the intersection. It is assumed that each cell's capacity is 1.

Traffic Networks (TNs) are modeled using two types of Petri nets in the simulation. The intersections of the traffic network are represented by a Colored Timed Petri Net (CTPN), and traffic lights are represented by a Timed Petri Net (TPN). Traffic lights are considered to pertain to a common signal timing plan. A TPN is defined as a digraph $TPN=\{P,T,Pre,Post,FT\}$ in which P represents a set of places, T represents a set of transmissions, Pre and Post are the pre-incidence and post-incidence matrices and FT is the firing time vector. FT specifies the deterministic duration of the firing of each transition. Colored Timed Petri Nets (CTPNs) are defined as a digraph $CTPN=\{P,T,C_0,Pre,Post,FT\}$, in which we can consider all of the same elements with the TPN and C_0 as different colors. Furthermore, T is a set of timed transitions representing the flow of vehicles between successive cells.

The value of FT_j pertaining to each transition is equal to the average time interval at which each vehicle moves from one cell to another or occupies it. This value depends on the average speed of the vehicle. The firing times are equal to the times when vehicles can enter the network. In addition, a colored token is indicative of a vehicle. The color of each token is equal to the routing assigned to each vehicle. This routing indicates the different paths that a vehicle can travel, starting from a particular position.

The Petri net model of the three links' traffic lights is shown in Figure 2.

In the initial state, the token is located in m_0, then the token is transited to m_1 and the traffic light color is changed to the color of the corresponding place. The successive steps take place similarly until a cycle completes and reaches the initial state.

Moreover, the TPN model of the traffic light controller is depicted in Figure 3 in accordance with Table 1. Also, the controller operating process is shown in Figure 4.

2. 3. Timed Petri Net Model of Traffic Lights

The traffic lights of a traffic network must be defined in accordance to a signal timing plan. This signal timing

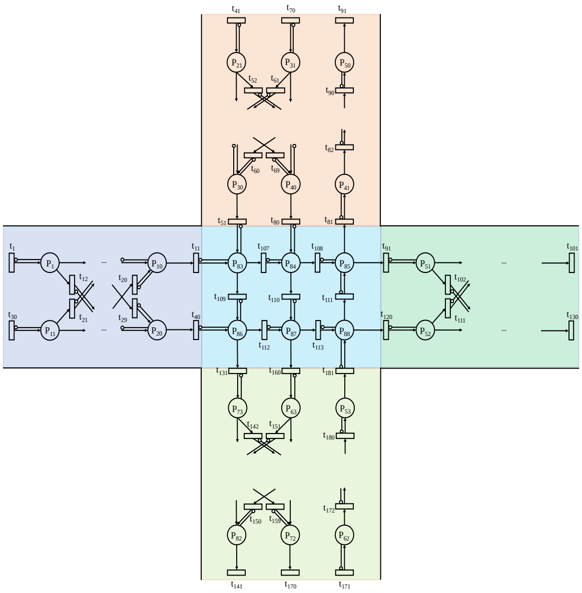


Figure 2. CTPN modelling the simulated intersection

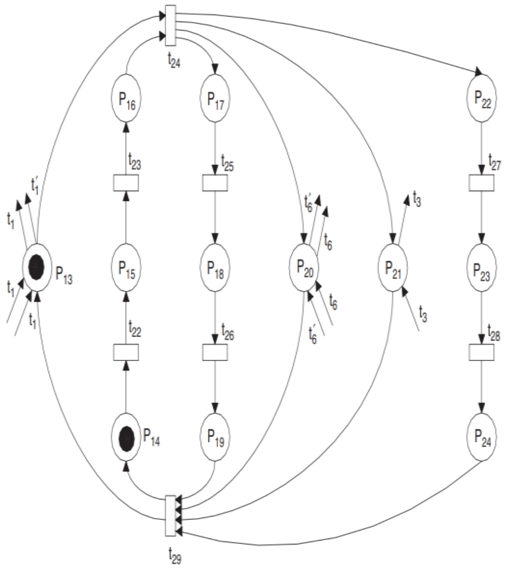


Figure 3. The TPN Modeling of the Traffic Light Controller of Intersection

TABLE 1. Signal Timing Plan of Represented Intersection

			Phases					
		Links	1	2	3	4	5	6
	1, 2	1	🟥	🟥	🟥	🟥	🟩	🟨
Streams	5	3	🟥	🟥	🟩	🟨	🟥	🟥
	3, 4	6	🟩	🟨	🟥	🟥	🟥	🟥
Phase Duration [s]			τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Cycle Duration [s]			CT					

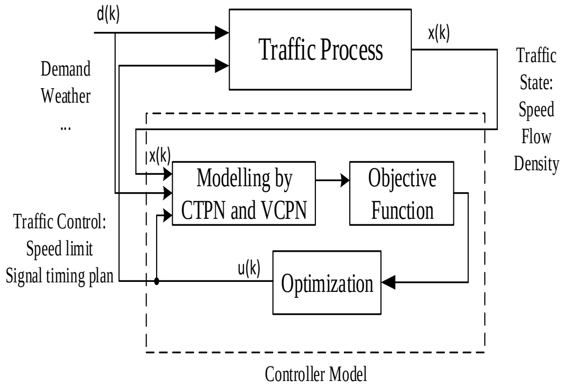


Figure 4. The way of applying the control signal to the system

plan has three phases including red, yellow and green used in Iran. Green, red and amber splits are the decision variables in the timing plan.

Cycle Time is defined as the duration of time from the center of red phase to the center of the next red phase. Furthermore, Green split for a signal is the fraction of the cycle time when the light is green in a certain direction.

The phase is defined as the time interval during which a given combination of traffic signals in the area is unchanged. The definition of the $TPN = \{P, T, \Pr e, Post, FT\}$ is utilized to model the traffic light controller. The places P and transitions, T, demonstrate the green, yellow, and red phases and their succession.

In order to obviously clarify the proposed method to model a generic signal timing plan, Table 1 shows a sample timing plan of the intersection's traffic lights. According to the table, the time of green phase for the traffic lights of link 1, 3 and 6 can be tuned by intelligent algorithms discussed in this paper. For the sample intersection, the streams allowed to proceed during the phases of the signal timing plan are depicted in Figure 3. These streams are numbered from 1 to 6. Moreover, the fixed signal timing plan consists of 6 phases as depicted in Figure 5.

2. 4. Connection Modelling by VCPN

As mentioned before, in this paper a highway and the intersections are modeled by VCPN and CTPN, respectively. In fact, they can be considered as modules and can be connected together in order to constitute a specific part of traffic network in a city. Connections are used to join highways and intersections. The highway exits connect to intersections via paths called off-ramp while intersections connect to highway entrances via road junctions called on-ramp. Additionally, the connection between two intersections is created by a street called junction. It should be noticed that off-ramp, on-ramp and road junction are modelled by VCPN.

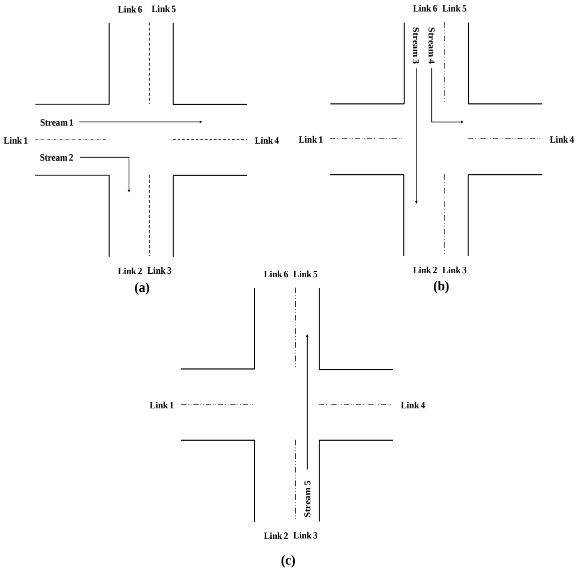


Figure 5. The Streams of the signal timing plan controlling the intersection (a) The Streams of Link 1 (b) The Streams of Link 6 (c) The Streams of Link 3

3. TRAFFIC NETWORK DESCRIPTION

Based on the modeling described before, a modular Petri nets model is proposed to describe a traffic network. To be more precise, the TN can be divided into the following sections: Highway and intersections. These sub-sections are then interconnected to create the model of the sample traffic urban network.

In this section, an urban network used to implement the proposed method is described. The sample traffic urban network is shown in Figure 6. It consists of a part of a six segments highway having a 3 km length and three lanes in aggregate. It includes three junctions considered to be juxtaposed to each other and two ramps with two lanes connecting the highway to the junctions as well. The distance between junctions and the length of both ramps is equal and about 500 m. Although cars and buses can enter this urban section via the north and south intersections, the main entrance of this urban network is located at the beginning of the highway. The entering cars can take two routes; the highway and the off-ramp. It is assumed that about 80 percent of total input cars pass the highway while only 20 percent take the off-ramp.

The traffic control of each segment is implemented using speed limitation. As mentioned before, the traffic flow in the highway can be modeled by VCPN in which various segments of the highway can be characterized by their own characteristics such as density, capacity, speed limitation, and flow rate. Additionally, the highway comprises an off-ramp road for transit vehicles to the first intersection. The input cars to the off-ramp are a fraction of the total input to the highway. On the other hand, the on-ramp road connects the third intersection to the end of

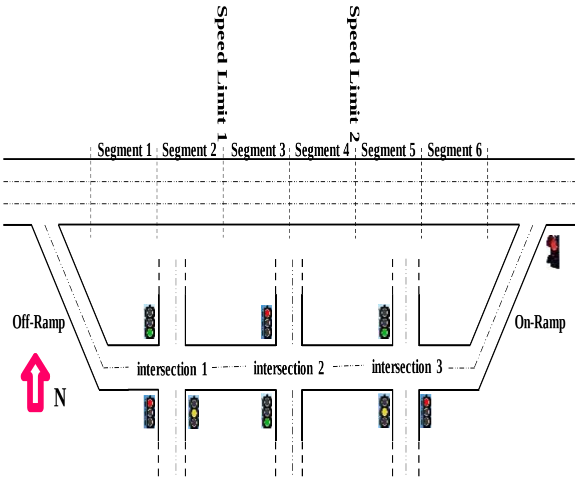


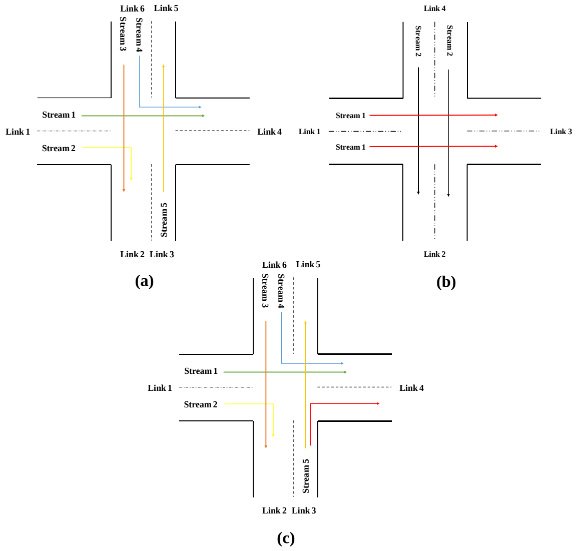
Figure 6. The Sample Traffic network

the highway. Both ramps are modeled by VCPN. The street located between the intersections can be considered as one or more sections and modeled by VCPN. Therefore, the urban network possesses six segments, two ramps, and two streets in the aggregate, and the traffic flow in these areas can be analyzed by VCPN. In contrast, the traffic flow in all three intersections is modeled using TPN and controlled by setting a signal timing plan. In the first intersection, there are three input links called links 1, 3, and 6. Links 1 and 3 possess two streams, while link 6 has one stream used for only BRT. The buses just take Link 5 to depart from the intersection. Links 2,4, and 6 are considered as the intersection output. Figure 7(a) illustrates the intersection associated with streams of cars. The second intersection, located in the middle of the urban area, is an ordinary one that personal cars enter via link 1, as shown in Figure 7(b). The third intersection resembles the first intersection, except that there is no special route for BRT. In other words, the car moving in-stream five can keep right or straight, as shown in Figure 7(c).

4. OPTIMIZATION ALGORITHM

In this section, the proposed traffic system control flowchart is presented. Such a structure must be able to optimize traffic flow by setting the time of traffic light phases in intersections and speed limitation in highways under different normal hours and abnormal conditions.

As stated before, this optimization method aims to minimize the number of occupancy in the considered urban area after a specific period of time by setting signal timing plans of traffic lights in all three intersections and speed limitations in different segments of the highway. In fact, we face an optimization problem in which the objective function is the number of occupancies in the urban area, and decision variables are the time of each phase in all of the three intersections and speed limitation in every segment of the highway. In this section, the optimization problem will be formulated. The value of the objective function can be defined as follows:



(a)

(b)

(c)

Figure 7. The Streams and links of intersections (a) intersection 1 (b) intersection 2 (c) intersection 3

$$OF = \sum_{i=1}^{N\_Seg} Nc\_seg(i) + Nc\_\mathrm{junc}12 + Nc\_\mathrm{junc}23 + Nc\_\mathrm{on\_ramp} + Nc\_\mathrm{off\_ramp} + \sum_{i=1}^{N\_\mathrm{Junc}} Nc\_\mathrm{junc}(i) + VarOnRramp + VarSpeedSegL\mathrm{im}2 + VarSpeedSegL\mathrm{im}4 \tag{9}$$

$$VarSpeedSegLim2 = \alpha_{Speed\_\mathrm{lim}\_2} \times \left(r_{speed\_\mathrm{lim}\_2}(t) - r_{speed\_\mathrm{lim}\_2}(t-1)\right) \tag{10}$$

$$VarSpeedSegLim4 = \alpha_{Speed\_\mathrm{lim}\_4} \times \left(r_{speed\_\mathrm{lim}\_4}(t) - r_{speed\_\mathrm{lim}\_4}(t-1)\right) \tag{11}$$

$$VarOnRamp = \alpha_{on\_ramp} \times \left(r_{on-ramp}(t) - r_{on-ramp}(t-1)\right)^2 \tag{12}$$

where $Nc\_seg(i)$ and $Nc\_\mathrm{junc}(i)$ denotes the number of vehicles in the highway segments and junctions at the end of the time horizon (t_{end}) respectively and the index i denotes the number of segments and intersections. Additionally, since variation of speed limitation in each step time can disrupt urban order, two terms are considered as penalty for changing the speed limitation on segments 2 and 4, denoted by $VarSpeedLimSeg2$,

$VarSpeedLimSeg4$. They are calculated in Equations (10)-(11). On the other hand, the traffic light installed on the on-ramp must rarely change the ratio of green and red phases in every step time. Therefore, when this ratio changes, a penalty must be imposed on the objective function. This penalty is denoted by $VarOnRamp$ and calculated according to Equation (12). $\alpha_{Speed\_\lim}, \alpha_{on\_ramp}$ are considered as weight factors. Also, the sum of occupancy between junctions 1, 2 and 3 and both ramps must be added to the objective function. According to Equation (9), it is obvious that the optimization problem aims to minimize the number of remaining cars in the urban area without changing the speed limit in the highway and the control signal in the on-ramp traffic light.

The various steps of the flowchart are given as follows:

Algorithm1:

Input: Length, Free Speed, Maximum Speed, Max Density, Critical Density, Maximum Flow Rate, Maximum Capacity and Number of Highway Segments and Set of Places, Transitions and Colors and Pre-incidence and Post-incidence matrices.

Output: Speed limitation on segments 2 and 4, signal timing plan of the intersections.

Step 1: Initialize the preliminary data including Length, Free Speed, Maximum Speed, Max Density, Critical Density, Maximum Flow Rate, Maximum Capacity and Number of Highway Segments and Set of Places, Transitions and Colors and Pre-incidence and the Post-incidence matrices.

Step 2: Initialize the number of input cars in various scenarios and periods of time

Step 3: Assign the initial population of decision variable containing signal timing plan of traffic lights in three intersections and speed limitations in segment 2 and 4 in the highway randomly for first step time.

Step 4: Run PSO algorithm to find the most optimized solution in the first step time

Step 5: Run the PSO algorithm again to find the best solution for the next step time. Consider the final status of the best solution in the previous step time as the initial status of urban network for the next step time.

Step 6: Go to step 5 if there is a next step time, otherwise go to 7.

Step 7: Save the best solution and end.

5. SIMULATION RESULTS

In order to evaluate the control system efficiency proposed in this paper, a comparison study is conducted between the proposed model traffic controller and no control condition. In the proposed traffic controller model, the Signal Timing Plan of intersections and on-ramp and speed limit in the highway are set using a Particle swarm (PSO) algorithm aiming to minimize the number of occupancies in the urban area. In contrast, it is supposed that the Signal Timing Plan in no control condition is fixed and unchanged. In other words, Signal Timing Plan and the speed limit are optimized according to the traffic volume. Five different types of scenarios are considered in order to evaluate the control system's performance. They are called normal hour, rush hour, midnight hour, rainy weather conditions, and accident hours. MATLAB software is utilized for simulation. The different scenarios mentioned above will be described in the following section in detail, and simulation results will be presented under these scenarios.

5. 1. Rush Hours Scenario Traffic flow during rush hours is one of the most challenging issues in the urban traffic system. In this period of time, as illustrated in Table 2, most intersections and highways in the city are occupied and congested by a large number of vehicles. This causes considerable air pollution and waste of time. For this reason, traffic flow control is much more important in these hours .

Traffic flow in the sample urban area is considered in this subsection, and the traffic control system is optimized by the proposed method. The number of cars arriving at different parts of urban areas within 15 minutes has a Poisson distribution with parameters tabulated in Table 2. As mentioned before, the total simulation time is 2 hours; consequently, this period of time is divided into 8 equal periods lasting 15 minutes. As expected, this table shows that the rate of input cars is significantly high in this period of time. Figure 8 illustrates the average number of cars in different areas in the traffic network. Figure 9 illustrates the status of the traffic network at 11:15 for both cases of optimization control method and fixed signal timing plan. The results clarify the advantage of applying the optimization method to set the timing signal of traffic lights in the intersection and on-ramp and the speed limit on the highway. The comparison between the two cases shows that when the number of input cars grows during rush hours, the proposed method can considerably enhance the traffic flow.

5. 2. Normal Hours Scenario In this scenario, it is assumed that the number of input cars and, consequently, the traffic volume is mediocre. This period of time refers to the midday hours, typically between 11 and 13. The average number of input cars in this period of time is reported in Table 3. The simulation result shows that the number of cars in the optimized control in all parts is lower than in the no-control condition, and the optimization control system can considerably increase the traffic flow rate of the network. Generally speaking, the results verify that the optimization control system can reduce the number of remaining cars rather than the no-control conditions in normal hours.

TABLE 2. Real traffic data in rush hours

	Highway Entrance	Intersection I Link 6	Intersection I BRT (Link 3)	Intersection II Link 4	Intersection III Link 6	Intersection III Link 3
19:00 - 19:15	1100	800	15	700	500	250
19:15 - 19:30	1300	600	12	700	600	200
19:30 -19:45	1450	700	14	650	400	300
19:45 - 20:00	1500	680	16	680	480	280
20:00 - 20:15	1500	630	18	500	550	290
20:15 - 20:30	1300	650	13	400	580	310
20:30 -20:45	1200	670	11	400	480	340
20:45 - 21:00	1000	600	9	300	380	200

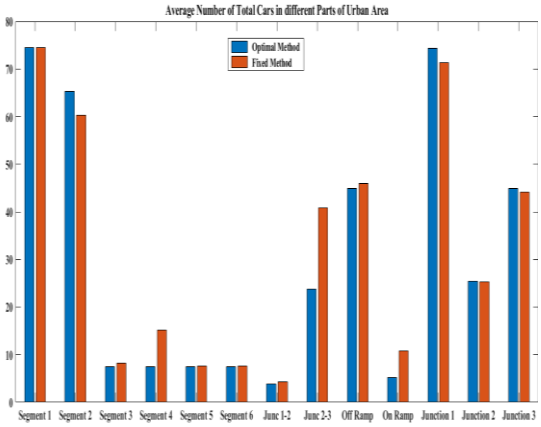
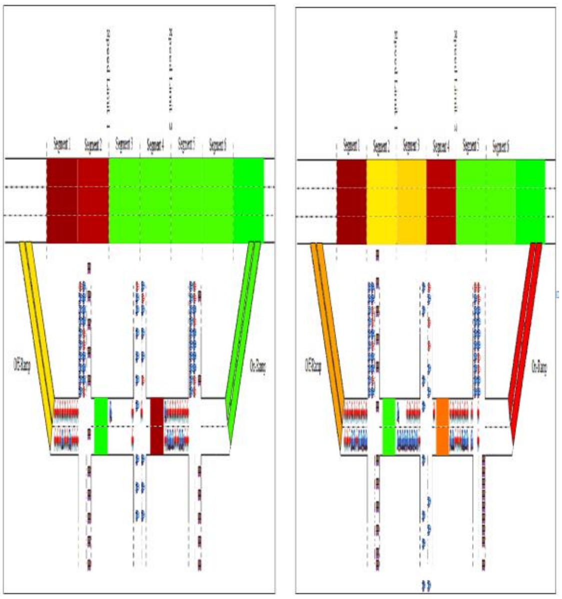


Figure 8. Average Number of Cars in Different Areas in Traffic Network



(b) (a)

Figure 9. Status of traffic network at 12:00 (a) Optimal Method (b) Fixed Method

5. 3. Midnight Hour Scenario Despite the low number of cars in these hours, the control system performance for both cases is compared in this scenario. The simulation results show that because of the low traffic flow at midnight hours, the optimization of the

TABLE 3. The average number of input cars in normal hours

	Highway Entrance	Intersection I Link 6	Intersection I BRT (Link 3)	Intersection II Link 4	Intersection III Link 6	Intersection III Link 3
11:00 - 11:15	700	300	9	240	140	100
11:15 - 11:30	800	250	7	250	180	120
11:30 -11:45	900	280	6	270	220	140
11:45 - 12:00	850	330	5	210	210	125
12:00 - 12:15	780	310	6	265	265	145
12:15 - 12:30	800	275	7	200	200	170
12:30 -12:45	750	240	7	180	180	135
12:45 - 13:00	810	250	8	150	150	155

control system has a negligible effect on the amount of traffic in the urban area. In other words, whether the timing signal traffic light and the speed limit are optimized or not, the traffic flow will be acceptable. Therefore, there is no necessity to use the proposed method in this period of time. The average number of input cars in this period of time is reported in Table 4.

5. 4. Rainy Weather Condition Scenario

Rain can slip the surface of streets. Hence, drivers have to drive much more carefully and reduce their speed. As a result, traffic flow increases in the rainy weather condition significantly. Under these conditions, the control system performance is evaluated when both cases are applied. It is supposed that in a specific period of time, the number of input cars increases suddenly, and the traffic network is congested. The data of input cars in this scenario is given in Table 5.

The comparison between Tables 5 and 2 shows that the number of input cars between 11:30 and 12:15 increases by 15%. Similarly, the simulation results show that the optimization method can untie the traffic node under this condition. This means that the proposed method can be used even in critical situations like rainy weather.

5. 5. Accident Scenario

In this scenario, it is supposed that an accident happens in segment five on the highway at 11:45 and lasts for 15 minutes. It is obvious that the flow rate of vehicles decreases because at least two main streams are congested on the highway. Since this happens on the highway, only the traffic flow in this part of the urban area is affected. As mentioned before, traffic flow on the highway is controlled by the speed limit on segments. Therefore, the signal timing plan of traffic lights is not important in this scenario. As the simulation results show in Figure 10 When speed limits in segments 2 and 4 are set appropriately, congestion caused by accidents can be reduced. In fact, speed limitation in the segments before the accident must be reduced while speed limitation in the next segments must be increased.

TABLE 4. The average number of input cars at midnight hours

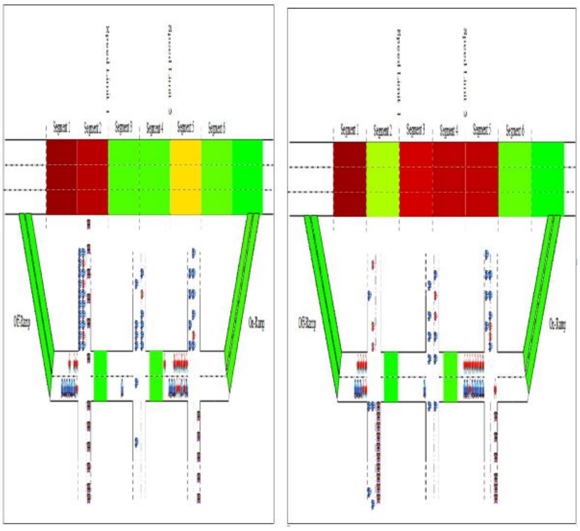
	Highway Entrance	Intersection I Link 6	Intersection I BRT (Link 3)	Intersection II Link 4	Intersection III Link 6	Intersection III Link 3
11:00 - 11:15	260	80	1	70	80	80
11:15 - 11:30	250	60	0	70	60	60
11:30 -11:45	230	70	1	65	70	70
11:45 - 12:00	220	68	0	68	68	68
12:00 - 12:15	210	63	1	50	63	63
12:15 - 12:30	200	65	0	40	65	65
12:30 -12:45	180	67	1	40	67	67
12:45 - 13:00	190	60	0	30	60	60

TABLE 5. The data of input cars in Rainy weather condition

	Highway Entrance	Intersection I Link 6	Intersection I BRT (Link 3)	Intersection II Link 4	Intersection III Link 6	Intersection III Link 3
11:00 - 11:15	700	300	9	240	140	100
11:15 - 11:30	800	250	7	250	180	120
11:30 -11:45	1200	400	6	400	410	250
11:45 - 12:00	1100	450	5	350	400	270
12:00 - 12:15	1000	420	6	410	430	240
12:15 - 12:30	800	275	7	200	200	170
12:30 -12:45	750	240	7	180	180	135
12:45 - 13:00	810	250	8	150	150	155

TABLE 6. Performance Improvement of Optimization Control System

Percentage of performance improvement (%)	**Normal Hours**	**Rush Hours**	**Midnight Hours**	**Rainy**	**Accident**
No control	10.54	5.8233	0.1599	40.8599	20.8193
Timing signal planningoptimization (Our proposed Method)	5.23	1.24	0.004	28.15	10.54
Highway Optimozation Previous paper [19]	8.37	3.23	0.089	32.189	15.12



(b) (a)

Figure 10. Status of traffic network at 12:00 (a) Optimal Method (b) Fixed Method

6. DISCUSSION

In order to examine the proposed optimization method more, a comparison study is conducted. Four different control system methods are considered. In the first method, it is supposed that there is no control for signal timing plan in the intersections and speed limitation in the highway, and they are determined based on historical data in the intersection and highway. In the third method, it is supposed that the urban traffic network model proposed by Dotoli and Fanti [10] is used to optimize only timing signal planning in the intersections. In this paper, a modular framework based on colored timed Petri nets (CTPNs) is presented to represent the dynamics of signalized TN systems: places modeling link cells intersections, tokens are vehicles and token colors .

Demonstrate the routing of the corresponding vehicle. In addition, ordinary timed Petri nets model the signal timing plans of the traffic lights controlling the area. The proposed modeling framework is applied to a real intersection. In contrast, in the third method, continuous Petri nets models proposed by Tolba et al. [19] for the analysis of highways are used. This model is proposed for the analysis and control design in highways. Under this condition, speed limitation on highways is just optimized. In the last control system, the traffic control flow is fully optimized using a signal timing plan in the intersections and speed limitations on the highways.

Table 6 tabulates the percentage of performance improvement of these four optimization control system methods. As expected, the simulation results show that when the optimization process is implemented for the whole of the system, the percentage of performance improvement is considerably higher. Additionally, when the signal timing plan is just optimized, the traffic flow is conducted more effectively in comparison with speed limitation optimization

7. CONCLUSION

In this paper, a modular framework is proposed to model the traffic flow of the Highway and intersections. The modeling procedure for the intersections is based on Timed Petri Nets, while Continuous Petri Nets with Variable speeds are used to model the highway networks. The timing of traffic lights and variable speed limits on the Highway were optimized by a PSO, and the status of the urban network area was compared in both optimization and fixed control conditions. Six various scenarios, including rush hours, normal hours, midnight hours, rainy weather, and accident occurrence, were considered to completely examine both conditions. The simulation results showed that when traffic control is optimized, the number of occupancies can be reduced in special conditions such as rainy weather and accident occurrence. In contrast, when traffic flow is extremely high or low, like rush and midnight hours, control flow optimization is not able to improve traffic flow in the city.

8. REFERENCES

1. Hafram, S.M., Valery, S. and Hasim, A.H., "Calibrating and validation microscopic traffic simulation models vissim for enhanced highway capacity planning", ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 8, (2023), 1509-1519. doi: 10.5829/IJE.2023.36.08B.11

2. Skovajsa, J., Přibyl, O., Přibyl, P., Ščerba, M. and Janota, A., "Evaluation of a mobile highway management system at roadwork zones", ***International Journal of Engineering, Transactions B: Applications***, Vol. 35, No. 5, (2022), 900-907. https://doi.org/10.5829/IJE.2022.35.05B.06

3. Di Febbraro, A., Giglio, D. and Sacco, N., "Urban traffic control structure based on hybrid petri nets", ***IEEE Transactions on***

Intelligent Transportation Systems, Vol. 5, No. 4, (2004), 224-237. https://doi.org/10.1109/TITS.2004.838180

4. Papageorgiou, M., Hadj-Salem, H. and Middelham, F., "Alinea local ramp metering: Summary of field results", ***Transportation Research Record***, Vol. 1603, No. 1, (1997), 90-98. https://doi.org/10.3141/1603-12

5. Taheri, M., Arkat, J., Farughi, H. and Pirayesh, M., "Modeling the traffic signal control system at an isolated intersection using queuing systems", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 34, No. 9, (2021), 2077-2086. https://doi.org/10.5829/IJE.2021.34.09C.05

6. Keyvan-Ekbatani, M., Papageorgiou, M. and Knoop, V.L., "Controller design for gating traffic control in presence of time-delay in urban road networks", ***Transportation Research Procedia***, Vol. 7, (2015), 651-668. https://doi.org/10.1016/j.trpro.2015.06.034

7. Mohammadi, M., Dideban, A., Lesani, A. and Moshiri, B., "An implementation of the ai-based traffic flow prediction in the resilience control scheme", ***International Journal of Transportation Engineering***, Vol. 8, No. 2, (2020), 185-198. https://doi.org/10.22119/ijte.2020.218863.1509

8. Fu, H. and Chen, K., "Macroscopic traffic modeling of heterogeneous road networks using coloured petri nets", in 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), IEEE. (2018), 1-6. https://doi.org/10.1109/ICNSC.2018.8361304

9. Huang, Y.-S., Weng, Y.-S. and Zhou, M., "Modular design of urban traffic-light control systems based on synchronized timed petri nets", ***IEEE Transactions on Intelligent Transportation Systems***, Vol. 15, No. 2, (2013), 530-539. https://doi.org/10.1109/TITS.2013.2283034

10. Dotoli, M. and Fanti, M.P., "An urban traffic network model via coloured timed petri nets", ***Control Engineering Practice***, Vol. 14, No. 10, (2006), 1213-1229. https://doi.org/10.1016/j.conengprac.2006.02.005

11. Qi, L., Zhou, M. and Luan, W., "A two-level traffic light control strategy for preventing incident-based urban traffic congestion", ***IEEE Transactions on Intelligent Transportation Systems***, Vol. 19, No. 1, (2016), 13-24. https://doi.org/10.1109/TITS.2016.2625324

12. Fu, H., Chen, S., Chen, K., Kouvelas, A. and Geroliminis, N., "Perimeter control and route guidance of multi-region mfd systems with boundary queues using colored petri nets", ***IEEE Transactions on Intelligent Transportation Systems***, Vol. 23, No. 8, (2021), 12977-12999. https://doi.org/10.1109/TITS.2021.3119017

13. Bargegol, I., Najafi Moghaddam Gilani, V. and Jamshidpour, F., "Relationship between pedestrians’ speed, density and flow rate of crossings through urban intersections (case study: Rasht metropolis)(research note)", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 30, No. 12, (2017), 1814-1821. doi: 10.5829/ije.2017.30.12c.01

14. Srivastava, S. and Sahana, S.K., "Nested hybrid evolutionary model for traffic signal optimization", ***Applied Intelligence***, Vol. 46, No. 1, (2017), 113-123. https://doi.org/10.1007/s10489-016-0827-6

15. Luo, J., Huang, Y.-S. and Weng, Y.-S., "Design of variable traffic light control systems for preventing two-way grid network traffic jams using timed petri nets", ***IEEE Transactions on Intelligent Transportation Systems***, Vol. 21, No. 7, (2019), 3117-3127. https://doi.org/10.1109/TITS.2019.2925824

16. Perez-Murueta, P., Gómez-Espinosa, A., Cardenas, C. and Gonzalez-Mendoza Jr, M., "Deep learning system for vehicular re-routing and congestion avoidance", ***Applied Sciences***, Vol. 9, No. 13, (2019), 2717. https://doi.org/10.3390/app9132717

17. Mercan, M.C., Kayalica, M.Ö., Kayakutlu, G. and Ercan, S., "Economic model for an electric vehicle charging station with v ehicle-to-grid functionality", ***International Journal of Energy Research***, Vol. 44, No. 8, (2020), 6697-6708. https://doi.org/10.1002/er.5407

18. Novikov, A., Novikov, I. and Shevtsova, A., "Modeling of traffic-light signalization depending on the quality of traffic flow in the city", ***Journal of Applied Engineering Science***, Vol. 17, No. 2, (2019), 175-181. https://doi.org/10.5937/jaes17-18117

19. Tolba, C., Lefebvre, D., Thomas, P. and El Moudni, A., "Continuous petri nets models for the analysis of traffic urban networks", in 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236), IEEE. Vol. 2, (2001), 1323-1328. https://doi.org/10.1109/ICSMC.2001.973104

Persian Abstract

چکیده

در این مقاله کنترل یکپارچه بزرگراه ها و تقاطع ها مورد بررسی قرار گرفته است. یک چارچوب مبتنی بر شبکه پتری مدولار برای مدل‌سازی جریان ترافیک بزرگراه‌ها و سیستم‌های شبکه ترافیک شریانی پیاده‌سازی شده است. در این چارچوب، چراغ‌های راهنمایی تقاطع شریانی توسط شبکه‌های پتری زمان‌بندی شده (TPN)مدل‌سازی می‌شوند. زمان‌بندی چراغ‌های راهنمایی و محدودیت‌های سرعت متغیر در بزرگراه با استفاده از یک الگوریتم هوشمند بهینه‌سازی می‌شود. این الگوریتم مبادله ای بین طول صف وسایل نقلیه در بزرگراه و رمپ ورودی و طول صف در تقاطع پس از هر چرخه زمانی ارائه می دهد. عملکرد کنترل کننده ترافیک بهینه و کنترل ثابت مقایسه شد. نتایج شبیه‌سازی تأیید می‌کند که استفاده از روش‌های بهینه‌سازی برای مدیریت زمان‌بندی چراغ‌های راهنمایی در تقاطع‌ها و محدودیت سرعت در بزرگراه‌ها می‌تواند به طور قابل‌توجهی جریان ترافیک را در شرایط خاص مانند هوای بارانی و تصادفات بهبود بخشد. علاوه بر این، این روش می تواند به طور قابل توجهی جریان ترافیک را در ساعات عادی افزایش دهد، در حالی که در ساعات شلوغی و نیمه شب، چنین بهبودی ناچیز است

AIMS AND SCOPE

The objective of the International Journal of Engineering is to provide a forum for communication of information among the world's scientific and technological community and Iranian scientists and engineers. This journal intends to be of interest and utility to researchers and practitioners in the academic, industrial and governmental sectors. All original research contributions of significant value focused on basics, applications and aspects areas of engineering discipline are welcome.

This journal is published in three quarterly transactions: Transactions A (Basics) deal with the engineering fundamentals, Transactions B (Applications) are concerned with the application of the engineering knowledge in the daily life of the human being and Transactions C (Aspects) - starting from January 2012 - emphasize on the main engineering aspects whose elaboration can yield knowledge and expertise that can equally serve all branches of engineering discipline.

This journal will publish authoritative papers on theoretical and experimental researches and advanced applications embodying the results of extensive field, plant, laboratory or theoretical investigation or new interpretations of existing problems. It may also feature - when appropriate - research notes, technical notes, state-of-the-art survey type papers, short communications, letters to the editor, meeting schedules and conference announcements. The language of publication is English. Each paper should contain an abstract both in English and in Persian. However, for the authors who are not familiar with Persian, the publisher will prepare the latter. The abstracts should not exceed 250 words.

All manuscripts will be peer-reviewed by qualified reviewers. The material should be presented clearly and concisely:

- *Full papers* must be based on completed original works of significant novelty. The papers are not strictly limited in length. However, lengthy contributions may be delayed due to limited space. It is advised to keep papers limited to 7500 words.
- *Research* notes are considered as short items that include theoretical or experimental results of immediate current interest.
- *Technical notes* are also considered as short items of enough technical acceptability with more rapid publication appeal. The length of a research or technical note is recommended not to exceed 2500 words or 4 journal pages (including figures and tables).

Review papers are only considered from highly qualified well-known authors generally assigned by the editorial board or editor in chief. Short communications and letters to the editor should contain a text of about 1000 words and whatever figures and tables that may be required to support the text. They include discussion of full papers and short items and should contribute to the original article by providing confirmation or additional interpretation. Discussion of papers will be referred to author(s) for reply and will concurrently be published with reply of author(s).

INSTRUCTIONS FOR AUTHORS

Submission of a manuscript represents that it has neither been published nor submitted for publication elsewhere and is result of research carried out by author(s). Presentation in a conference and appearance in a symposium proceeding is not considered prior publication.

Authors are required to include a list describing all the symbols and abbreviations in the paper. Use of the international system of measurement units is mandatory.

- On-line submission of manuscripts results in faster publication process and is recommended. Instructions are given in the IJE web sites: www.ije.ir-www.ijeir.info
- Hardcopy submissions must include MS Word and jpg files.
- Manuscripts should be typewritten on one side of A4 paper, double-spaced, with adequate margins.
- References should be numbered in brackets and appear in sequence through the text. List of references should be given at the end of the paper.
- Figure captions are to be indicated under the illustrations. They should sufficiently explain the figures.
- Illustrations should appear in their appropriate places in the text.
- Tables and diagrams should be submitted in a form suitable for reproduction.
- Photographs should be of high quality saved as jpg files.
- Tables, Illustrations, Figures and Diagrams will be normally printed in single column width (8cm). Exceptionally large ones may be printed across two columns (17cm).

PAGE CHARGES AND REPRINTS

The papers are strictly limited in length, maximum 8 journal pages (including figures and tables). For the additional to 8 journal pages, there will be page charges. It is advised to keep papers limited to 3500 words.

Page Charges for Papers <u>More Than 8 Pages (Including Abstract)</u>	
For International Author ***	**$55 / per page**
For Local Author	**100,000 Toman / per page**

AUTHOR CHECKLIST

- Author(s), bio-data including affiliation(s) and mail and e-mail addresses).
- Manuscript including abstracts, key words, illustrations, tables, figures with figure captions and list of references.
- MS Word files of the paper.

International **J**ournal of **E**ngineering, P.O. Box 13145-1659, Tehran, I.R. Iran
Materials and Energy Research Center, 5 Ahuramazda St., Alvand Ave., Argentina Sq., Tehran, Islamic Republic of Iran
www.merc.ac.ir - www.emerc.ac.ir