IJE TRANSACTIONS B: Applications Vol. 37 No. 08, (August 2024) 1691-1699

International Journal of Engineering

Journal Homepage: www.ije.ir

Enhanced Autoregressive Integrated Moving Average Model for Anomaly Detection in Power Plant Operations

A. T. W. Khalid Fahmi[a], K. R. Kashyzadeh[*b], S. Ghorbani[a]

[a] *Department of Mechanical Engineering, Academy of Engineering, RUDN University, Moscow, Russian Federation*
[b] *Department of Transport Equipment and Technology, Academy of Engineering, RUDN University, Moscow, Russian Federation*

PAPER INFO

Paper history:
Received 11 February 2024
Received in revised form 25 March 2024--
Accepted -4 April 2024-

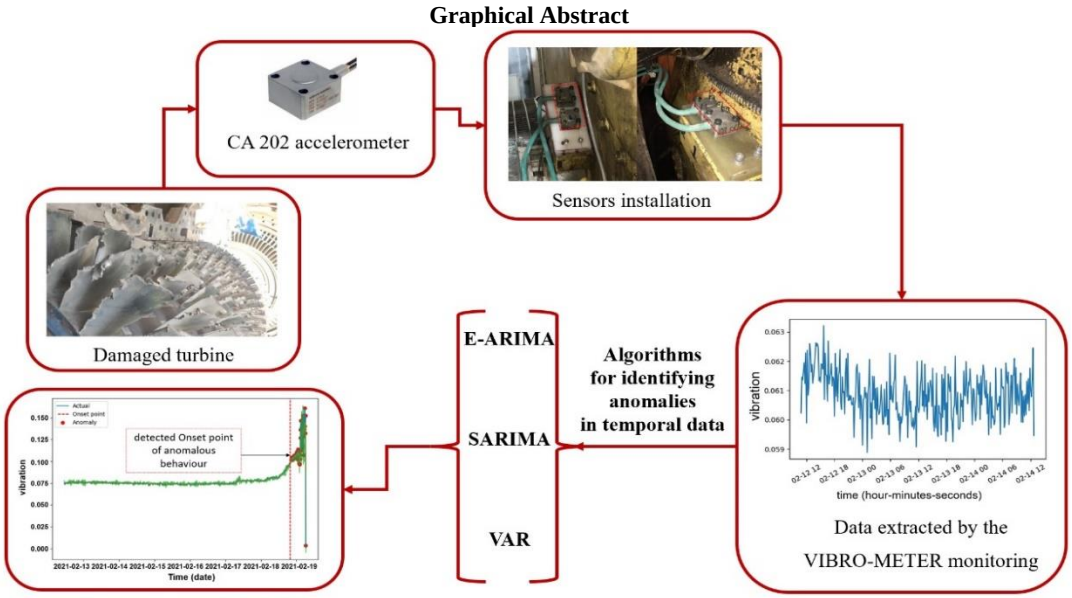
Keywords:
Vibration Monitoring
Time-series Data
Anomaly Detection
Autoregressive Integrated Moving Average
Gas Power Plant

ABSTRACT

This study introduces an Enhanced Autoregressive Integrated Moving Average (E-ARIMA) model for anomaly detection in time-series data, using vibrations monitored by CA 202 accelerometers at the Kirkuk Gas Power Plant as a case study. The objective is to overcome the limitations of traditional ARIMA models in analyzing the non-linear and dynamic nature of industrial sensory data. The novel proposed methodology includes data preparation through linear interpolation to address dataset gaps, stationarity confirmation via the Augmented Dickey-Fuller Test, and ARIMA model optimization against the Akaike Information Criterion, with a specialized time-series cross-validation technique. The results show that E-ARIMA model has superior performance in anomaly detection compared to conventional Seasonal ARIMA (SARIMA) and Vector Autoregressive models. In this regard, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) criteria were utilized for this evaluation. Finally, the most important achievement of this research is that the results highlight the enhanced predictive accuracy of the E-ARIMA model, making it a potent tool for industrial applications such as machinery health monitoring, where early detection of anomalies is crucial to prevent costly downtimes and facilitate maintenance planning.

***doi**: 10.5829/ije.2024.37.08b.19*

Graphical Abstract



*Corresponding Author Email: *reza-kashi-zade-ka@rudn.ru* (K. R. Kashyzadeh)

Please cite this article as: Khalid Fahmi ATW, Kashyzadeh KR, Ghorbani S. Enhanced Autoregressive Integrated Moving Average Model for Anomaly Detection in Power Plant Operations. International Journal of Engineering, Transactions B: Applications. 2024;37(08):1691-9.

1. INTRODUCTION

In the rapidly evolving field of industrial automation and monitoring, the analysis of sensory time-series data plays a fundamental role in ensuring the efficiency and reliability of machinery and systems (1, 2). Meanwhile, one of the important applications of data analysis by various artificial intelligence algorithms that can be seen in daily life is the prediction of transport traffic and routing based on the shortest possible time to reach the destination (3-5). In addition, in the past few years, in order to study the corona virus and its spread in the world, time series analysis methods were also used to analyze the speed of transmission, progress and even treatment time (6). Therefore, it is very important to pay attention to the analysis of time series data in events. Especially if this phenomenon is new, or happens for the first time, or if it does not happen repeatedly, like an earthquake in any region, which does not have the same intensity and duration time (7-9). In industries related to mechanical engineering, one of the most well-known time series data is vibrations. Vibration analysis has emerged as a critical tool in the predictive maintenance of industrial equipment, enabling early detection of anomalies that could lead to equipment failure (10). Meanwhile, the industries related to energy production are one of the most challenging, because over time the equipment wears out and on the other hand, with the growth of the population, the demand for energy production increases. To overcome this issue with the lowest cost (not making fundamental changes in the system and updating it), the industrialists are also forced to receive the maximum load capacity from their power plants. Gas turbines play a vital role in contemporary energy production and offer numerous advantages to satisfy worldwide energy needs. Their key feature is their exceptional efficiency, particularly when utilized in combined cycle setups, which promotes maximum energy utilization. Gas turbines are primarily powered by natural gas but can also run on various fuels, providing operational flexibility. Their compact design allows them to be located near demand centers and reduces energy loss during transmission (11).

Figure 1. Damaged turbine in the Kirkuk gas power plant (severe destruction of blades)

The ability of gas turbines to produce both electricity and heat demonstrates their efficiency and adaptability. Ongoing technological advancements improve their environmental compatibility and operational flexibility. Economically, gas turbines require less time to build and are cheaper than many traditional energy sources, making them an economically viable option for power generation (12). Despite their clear benefits, gas turbines face challenges, including compressor fouling, airflow reversal risks, and combustion issues. Moreover, bearings and seals are critical for turbine operation but can fail due to overheating or malfunctioning, respectively (13). Lubrication and cooling systems are susceptible to contamination and blockages, potentially causing overheating. Control systems depend on accurate sensor data to manage operations, but failures here can disrupt performance (14). Fuel system issues and structural integrity concerns, such as cracks from thermal cycling and vibrational disturbances, also pose risks. Electrical components (e.g., generators) may fail, leading to significant operational and financial consequences, such as downtime, increased maintenance costs, safety hazards, and environmental damage (15). Figure 1 illustrates images of extensive blade destruction in the Kirkuk gas power plant turbine, which requires complete repair to restore its operational capacity.

The ability to accurately detect anomalies in industrial systems is not just a technical challenge, but also a critical aspect of operational management that has important implications for safety, efficiency, and cost. As a result, if anomalies are not detected, they can lead to unexpected downtimes. Therefore, improving anomaly detection techniques in industrial environments, especially in critical facilities such as power plants, is of great importance. In general, to prevent and diagnose these issues, a variety of sensors and diagnostic techniques are employed. Also, vibration analysis, acoustic emissions monitoring, oil analysis, and performance monitoring are key methods used to detect and address potential problems. In this regard, Non-Destructive Testing (NDT) and advanced control systems like SCADA are integrated with these sensors to provide comprehensive diagnostics (16, 17). In addition, data analysis and machine learning are applied to the sensor dtata in order to predict wear and potential failures, and enable the planning of preventive maintenance strategies. Investigations into Artificial Intelligence (AI) methodologies for fault detection and diagnosis highlights the significant promise of cutting-edge AI technologies in enhancing the reliability and operational efficiency of gas turbines (18-20). Despite advances in sensor technology and data analysis, the dynamic and non-linear nature of sensory data poses significant challenges in accurate detection of anomalies. Previous research shows that traditional models such as Auto-Regressive Integrated Moving Average (ARIMA) and its

variants have been widely used for time-series analysis. Traditional ARIMA models exhibit significant limitations when analyzing non-linear and dynamic sensory data, especially in industrial contexts such as power plants. One of the fundamental drawbacks of ARIMA models lies in their inherent design, which is predominantly linear. This linearity assumes a constant relationship between past and future values, an assumption that fails when dealing with non-linear data where such relationships can vary significantly. In industrial environments, sensory data often exhibit non-linear characteristics due to complex interactions between numerous variables and the presence of unpredicted events or faults, which make ARIMA models inadequate to capture such complexities. Furthermore, the dynamic nature of sensory data in industrial environments, characterized by frequent shifts and sudden changes, poses a challenge to the stationary requirement of ARIMA models. ARIMA models necessitate that the data to be stationary, meaning that its statistical properties such as mean, variance, and autocorrelation remain constant over time. However, in power plants, data can fluctuate wildly and be affected by factors such as equipment wear, operational changes, and external disturbances. This non-stationarity in the data leads to poor predictive performance and unreliability in forecasting future values, which significantly limits the applicability of ARIMA models in these contexts (21). The main contributions of this paper are the development and validation of an E-ARIMA model specifically designed for anomaly detection in industrial sensory time-series data, demonstrating the superior performance of the model over traditional ARIMA and other time-series models through rigorous statistical evaluation, and provide a novel methodological framework that increases the accuracy and reliability of anomaly detection in complex industrial environments. This work also extends the application of time-series analysis in predictive maintenance and provides significant practical implications for improving operational efficiency and reducing maintenance costs in industrial settings.

The rest of this paper is organized as follows: Section 2 addresses the methodology, beginning with a detailed description of the data preparation process, including the collection and interpolation of the time-series data from the Kirkuk gas power plant. Section 3 introduces the E-ARIMA model and details its development, the rationale behind its design, and the algorithm structure that facilitates improved anomaly detection in sensory data. Section 4 presents the results of the study and shows the performance of the E-ARIMA model in comparison with traditional time-series models like SARIMA and Vector Autoregressive models. The paper concludes with section 5, which summarizes the main findings, contributions, and potential impact of the E-ARIMA model on the field of time-series analysis and anomaly detection in industrial settings.

2. DATA PREPARATION

This research uses a comprehensive dataset extracted from the Kirkuk gas power plant, which is currently being upgraded to a combined cycle power facility. Data collection occurred throughout February 2021, starting at midnight on the 1st and ending at 11:59 PM on the 28th. This dataset includes readings from four CA 202 accelerometers strategically installed on the four bearings, as shown in Figure 2. These readings provide insight into the plant operations, including machinery activity, load levels, and signs of wear.

The CA 202 accelerometer is purposefully designed with internal insulation and a symmetrical polycrystalline shear-mode measuring element. This transducer is specifically engineered for challenging industrial vibration monitoring and is an excellent choice for heavy-duty gas and steam turbines. The CA 202 piezoelectric accelerometer has a sensitivity rate of 100 pC/g at 120 Hz with ±5% deviation. In addition, it is robust enough to withstand transient overloads of up to 250 g. Also, its linearity is finely tuned and maintains an accuracy of 1% over a peak range of 0.0001 g to 20 g peak, with an average deviation of up to 2% over a peak range of 20g to 200 g. Moreover, its dynamic measurement capability is from 0.0001 grams per minute to a maximum 200 g. Notably, the accelerometer exhibits a transverse sensitivity of 5% and resonates at a nominal frequency of 11 kHz after installation. Its frequency

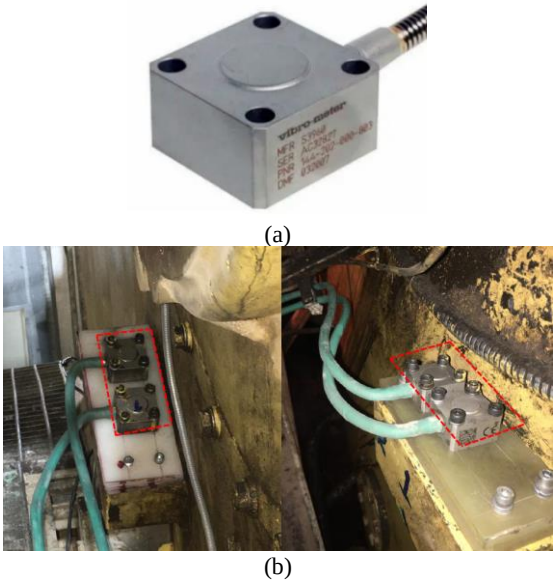


Figure 2. The equipment employed to monitor turbine vibrations in the Kirkuk gas power plant: (a) CA 202 accelerometer and (b) location of sensors installation

response remains constant within 5% from 0.5 Hz to 3000 Hz, with the lower threshold determined by the associated conditioner, and features a 10% response rate between 3 kHz and 4.5 kHz. Figure 3 presents a representative of the data extracted from the VIBRO-METER monitoring system installed in the Kirkuk gas power plant. In fact, this system effectively monitors the vibrations in different parts of the turbin (22).

Ensuring the smooth and accurate continuity of this time-series dataset is of considerable significant importance, as interruptions or gaps can distort its natural temporal sequences. In this research, we have chosen to employ the linear interpolation technique to effectively address and correct such irregularities. This approach estimates missing data points based on nearby values and preserves the integrity of the dataset. Particularly for time-sensitive measurements such as vibrations, linear interpolation assumes a constant progression between adjacent data points, which ensures seamless transition for missing values. However, it is necessary to examine the underlying causes of the data gaps. If they result from specific events such as equipment malfunctions, a more domain-specific treatment may be necessary. After the imputation process, the dataset is thoroughly checked to ensure that the introduced values do not skew the genuine patterns or introduce any biases.

3. E-ARIMA MODEL DEVELOPMENT

The introduction of ARIMA model for anomaly detection in sensory time-series data not only emphasizes its enduring relevance in time-series analysis, but also acknowledges its inherent limitations. While ARIMA excels at capturing and modelling linear relationships within stationary data, its traditional form struggles with the non-linear and dynamic nature of real-world sensory information. Consequently, we have improved the standard ARIMA model to address these shortcomings and ensure rigorous testing of model stationary and optimization, which is crucial for industrial applications like those in the Kirkuk power plant. This proactive measure allows our customized ARIMA model to maintain its structural advantages while effectively adapting to and detecting unusual patterns in volatile time-series data.

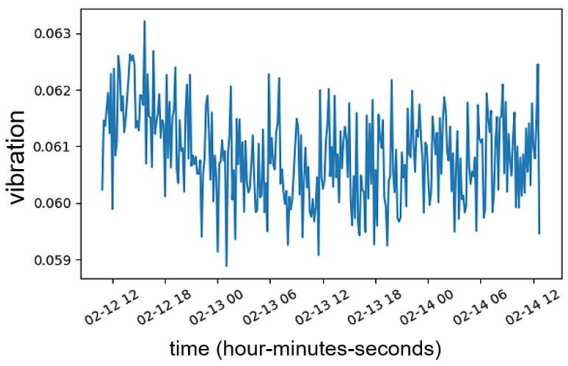


Figure 3. A representative of the data extracted by the VIBRO-METER monitoring system: installed at the Kirkuk gas power plant

Algorithm 1 presents the structure of the E-ARIMA model which we developed to detect anomalies in sensory time-series data (the data collected from CA 202 accelerometers installed in Kirkuk power plant). The initial stage of the algorithm involves preprocessing the data with the Augmented Dickey-Fuller (ADF) (23) test, which is crucial to verify the stationarity of the time series—a requirement for the subsequent ARIMA modelling. The ADF test is a statistical procedure used to test the stationarly of a time series. The test equation can be expressed as:

Algorithm 1 Enhanced ARIMA model for time-series anomaly detection.

```
procedure ADFTest(series, title)
    result ← adfuller(series.dropna(), 'AIC')
    labels ← ['ADF test statistic', 'p-value', '# lags used', '# observations']
    out ← Series(result[0 : 4], index = labels)
    for key, val in result[4] do
        out[f'critical value ({key})'] ← val
    end for
end procedure

for column in data.columns do
    ADFTest(data[column], column)
end for

procedure OptimizeARIMA(order_list, timeseries)
    best_aic ← ∞
    best_order ← None
    best_model ← None
    for order in order_list do
        model ← ARIMA(timeseries, order).fit()
        aic ← model.aic
        if aic < best_aic then
            best_aic ← aic
            best_order ← order
            best_model ← model
        end ifcontinue
    end for
    return best_order, best_model
end procedure

p, d, q ← range(0, 3)
pdq ← all combinations of p, d, q
best_models ← empty dictionary
for column in data.columns do
    order, model ← OptimizeARIMA(pdq, data[column])
    best_models[column] ← (order, model)
end for

procedure TimeSeriesCV(data, n_splits, std_multiplier)
    tscv ← TimeSeriesSplit(n_splits)
    thresholds ← empty list
    scores ← empty list
    for train_index, test_index in tscv.split(data) do
        train ← data.iloc[train_index]
        test ← data.iloc[test_index]
        threshold ← train.mean() + (std_multiplier × train.std())
        thresholds.append(threshold)
        order ← best_models[column][0]
        model ← ARIMA(train, order).fit()
        predictions ← model.predict(start = test.index[0], end = test.index[−1])
        anomalies ← detect_anomalies(test, predictions, threshold)
        scores.append(anomalies.sum())
    end for
end procedure
```

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \quad (1)$$

where, y_t, α, and βt are defined as, the time series at time t, the constant term, and the trend component, respectively. Moreover, γ is the coefficient on y_{t-1}, testing the null hypothesis H_0: $\gamma=0$ (non-stationary) against the alternative H_0: $\gamma<0$ (stationary). The δ_1, δ_2,…, δ_{p-1} and ε_t in Eq. 1 are the coefficients of lagged difference terms and the error term, respectively. The output of the ADF test includes a test statistic, the p-value, the number of lags used, and the number of observations, which provide a detailed description of the static properties of the data. If found to be non-stationarity, the series is differentiated until stationarity is achieved, as shown by the ADF test results. This test was utilized in the current study to check the stationarity of the time-series data. Stationarity is an important assumption in time-series analysis because most statistical models assume that the statistical properties of the data generating process do not change over time. The ADF test helps to confirm whether the data are stationary or need to be covaried to achieve stationarity, which is essential for accurate modelling and forecasting. The existence of non-stationarity in data can lead to misleading models and predictions, especially in the context of industrial sensory data, which are often characterized by dynamic changes and non-linear behaviors (24).

Once the stationarity is confirmed, the algorithm moves to the ARIMA model optimization phase. Here, a range of ARIMA configurations are evaluated (p, d, and q), where 'p' represents the order of the autoregressive terms, 'd' is the degree of differencing, and 'q' is the order of the moving average process. For each ARIMA configuration, the model's fit is assessed using the Akaike Information Criterion (AIC) (25), and the model with the lowest AIC is selected. This step not only identifies the best-fitting model, but also avoids overfitting by penalizing excessive complexity. Model optimization based on the AIC is another critical step in the model development process. It helps to select the best ARIMA model among a set of candidate models by considering the goodness of fit and complexity of the model. AIC penalizes adding more parameters to the model, which helps prevent overfitting. Overfitting is problematic in the context of non-linear and dynamic data, as complex models may capture noise rather than the underlying pattern, leading to poor predictive performance. By optimizing the AIC-based model, this study ensures that the model is sufficiently complex to capture the dynamics in the data while being simple enough to avoid overfitting, thereby enhancing its predictive accuracy and robustness in the face of non-linear and dynamic behaviors in the sensory data.

Unlike standard time-series cross-validation (TimeSeriesCV) methods, TimeSeriesCV respects the temporal order of observations, which is essential for time series analysis. During this step, a dynamic threshold for anomaly detection is established using the standard deviation of the training data multiplied by a predefined 'std_multiplier'. This adaptive threshold is key to the algorithm's ability to maintain high sensitivity to anomalies in the presence of volatile data behaviour. Our approach in developing the proposed E-ARIMA model is novel in its integration of rigorous stationarity testing, AIC-based model optimization, and a specialized cross-validation method for time-series data, which ensures that the model can effectively adapt to the unique characteristics of sensory data. This is especially important in real-world scenarios where such data may exhibit non-linear and non-stationary behaviours.

Figure 4 illustrates the performance of the proposed E-ARIMA algorithm for anomaly detection in vibration data collected from a CA 202 accelerometer installed in the Kirkuk power plant located in Iraq. Figure 4a shows that the vibration levels are relatively stable and a sharp increase indicates an anomalous event. Figure 4b demonstrates a similar scenario but with a different scale of vibration values. In both graphs, three types of lines are shown: the actual vibration data (in green), the points where the onset of an anomaly is detected (in dashed red), and the detected anomalies (in solid red).

In Figure 4a, the actual vibration measurements remain consistently low, with a few small spikes, until a significant increase marks the onset of anomalous behavior. The algorithm successfully detects the starting point shortly before the actual anomaly occurs, as indicated by the proximity of the dashed red line to the solid red anomaly marking. This early detection of potential issues in industrial settings is critical for preventive maintenance and damage reduction.

Figure 4b shows a scenario with higher vibration values, while the onset and anomaly are similar in the real data representation. The onset of anomalous behavior just before a dramatic rise in vibration levels is detected by the E-ARIMA algorithm, which flags a potential issue. This visualization emphasizes the sensitivity and accuracy of the algorithm in predicting deviations from normal operating patterns, which is very important for the proactive maintenance of machinery in the power plant. Overall, the graphs effectively demonstrate the ability of the E-ARIMA algorithm in early anomaly detection, a valuable asset to ensure operational continuity and safety in industrial environments.

4. RESULTS AND DISCUSSION

To elucidate the performance of the proposed E-ARIMA algorithm in anomaly detection, we conducted a thorough

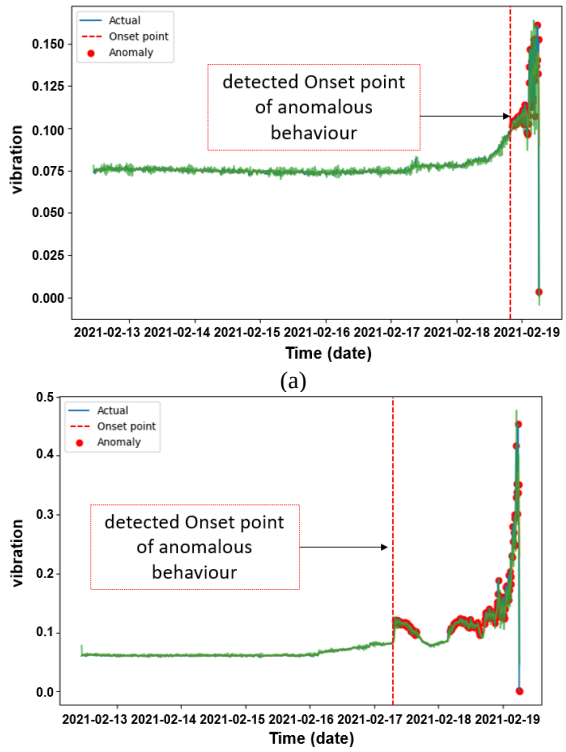


Figure 4. Performance of the proposed E-ARIMA algorithm in anomaly detection on collected sensory data from CA 202 accelerometer installed in Kirkuk power plant

benchmarking analysis against established algorithms in the field, particulary Seasonal Autoregressive Integrated Moving Average (SARIMA) and Vector Autoregressive (VA) models. We first discuss the architectural details of these two models and subsequently evaluate their performance against the proposed E-ARIMA model, comparing execution time, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

SARIMA model:

An extension of the ARIMA model incorporating seasonality, it is also adept at handling complex time series data with seasonal patterns. Its application in anomaly detection uses this capability to identify unusual events by analyzing deviations from seasonal norms and trends within the data.

The pseudocode starts by specifying the split_point to divide the dataset into train and test segments, ensuring a robust model evaluation. In the training loop, for each column, the SARIMA model is instantiated with sarima_order and seasonal_order, then fitted without display messages (disp=False). It then predicts values for the test length and stores the predictions in sarima_forecasts. Algorithm 2 presents SARIMA model. This structured approach enables model training and prediction on multiple data columns and demonstrates the versatility of the SARIMA model in handling various time series patterns within a dataset.

Algorithm 2 SARIMA Model

$split\_point \leftarrow \text{INT}(len(data) \times 0.8)$
$train, test \leftarrow data[: split\_point], data[split\_point :]$
$sarima\_forecasts \leftarrow \{\}$
for each $column$ in $train.columns$ **do**
  $model \leftarrow \text{SARIMAX}(train[column],\ order = sarima\_order,$
    $seasonal\_order = seasonal\_order).fit(disp = False)$
  $predictions \leftarrow model.forecast(steps = len(test))$
  $sarima\_forecasts[column] \leftarrow predictions$
end for
$results \leftarrow \{\}$
for each $column$ in $test.columns$ **do**
  $actual \leftarrow test[column].values$
  $sarima\_pred \leftarrow sarima\_forecasts[column]$
  $results[column] \leftarrow \text{CALCULATE\_METRICS}(actual, sarima\_pred)$
end for

VA model:

A multivariate forecasting algorithm, is adept at capturing the linear interdependencies among multiple time series. In anomaly detection, its strength lies in utilizing the collective information from various related series to identify unusual patterns or deviations that may not be apparent when examining the series independently. This makes the vector autoregressive model particularly useful in complex systems where multiple variables interact, and enables a comprehensive analysis that enhances the detection of anomalies based on the interconnected dynamics of the data.

Algorithm 3 begins by determining `split_point`, a critical step for dividing the dataset into `train` and `test` segments, a standard practice in time-series forecasting to validate model performance. The `train` data is then processed through interpolation (`train_interpolated`), which is an important step to fill in missing values, ensuring data consistency for the vector autoregressive model. Furthermore, this model is fitted with the Akaike Information Criterion ('aic'), which is an efficient method for determining the best lag length in the VAR context. The forecasting process involves using the latest observations (`var_forecast_input`) of the interpolated training data as input, a technique that uses the most recent data trends for prediction. Finally, the pseudocode illustrates a systematic evaluation process for the vector autoregressive model, where each column of the `test` dataset is assessed against the forecasts (`var_forecasts`), highlighting the model's predictive strength and accuracy using a bespoke metric calculation function (`calculate_metrics`).

In the field of time-series forecasting, the MSE, RMSE, and MAE are important criteria for evaluating model performance. These metrics quantitatively measure the difference between values predicted by a model and the observed data. These metrics are crucial for several reasons, (1) MSE provides a global measure

Algorithm 3 VAR Model

```
split_point ← INT(len(data) × 0.8)
train, test ← data.iloc[: split_point], data.iloc[split_point :]
train_interpolated ← train.interpolate()        ▷ Interpolate missing values
var_model ← VAR(train_interpolated).fit(ic =' aic')
var_forecast_input ← train_interpolated.values[−var_model.k_ar :]
var_forecasts ← var_model.forecast(y = var_forecast_input, steps = len(test))
results ← {}
for each column in test.columns do
    actual ← test[column].values
    var_pred ← var_forecasts[:, test.columns.get_loc(column)]
    results[column] ← CALCULATE_METRICS(actual, var_pred)
end for
```

of the quality of an estimator by calculating the mean squared difference between the estimated values and actual value. It gives a clear picture of the model's performance as it squares the errors before averaging, which penalizes more significant errors., (2) RMSE is the square root of MSE and is particularly useful when large errors are undesirable. It is in the same units as the response variable and can be more interpretable than MSE., and (3) MAE measures the average magnitude of errors in a set of predictions, regardless of their direction. It's a linear score, which means all the individual differences are weighted equally. Moreover, it is less sensitive than MSE. Equations (2) to (4) provide relationships to calculate MSE, RMSE, and MAE, respectively (26-28).

$$\mathrm{MSE} = \left(\frac{1}{\mathrm{n}}\right) * \Sigma(\mathrm{y} - \hat{\mathrm{y}})^2 \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \quad (3)$$

$$\mathrm{MAE} = \left(\frac{1}{\mathrm{n}}\right) * \Sigma|\mathrm{y} - \hat{\mathrm{y}}| \quad (4)$$

In the above formulas, the parameters n, y, and ŷ represent the number of data points, the actual values, and predicted values, respectively.

The evaluation metrics clearly indicate that the proposed E-ARIMA model surpasses both SARIMA and VAR models in accuracy. The proposed E-ARIMA model's MSE of 0.0023, RMSE of 0.0395, and MAE of 0.0262 are all significantly lower than those of the SARIMA and VA models. These results are not just statistically desirable; they represent a remarkable improvement in the model's ability to predict and detect anomalies. Lower MSE and RMSE values for the proposed E-ARIMA indicate a better fit to the data with fewer large errors, while the lower MAE reflects its high accuracy in all predictions without undue influence from outliers. The comparison with SARIMA and VA models, which show MSEs of 0.0041 and 0.0034, RMSEs of 0.0486 and 0.0466, and MAEs of 0.0348 and 0.034, respectively, highlights the improvements of proposed E-ARIMA.

It is importnant to notify that: The favorable performance of the E-ARIMA model in the analysis of Kirkuk gas power plant data is mainly due to its specialized design and adaptability to the unique characteristics of industrial time-series data. Unlike traditional ARIMA models, the E-ARIMA framework includes additional preprocessing and optimization steps that allow it to effectively handle the non-linear and dynamic aspects of industrial sensory data. This adaptability ensures that the model can accurately interpret and predict based on the complex data characteristics inherent in such industrial settings.

Ensuring stationarity through the Augmented Dickey-Fuller Test and appropriate differencing is a pivotal component of the E-ARIMA model, which is necessary for the analysis of Kirkuk gas power plant data. This process stabilizes the data, which is vital for generating reliable predictions, especially in environments such as power plants where operational changes can induce non-stationary data behavior. Furthermore, the selection of model parameters, optimized using the Akaike Information Criterion, avoids overfitting and underfitting, thereby enhancing the model's predictive precision and adaptation of the model to the nuances of the dataset.

The E-ARIMA model employs specialized time-series cross-validation and an adaptive thresholding mechanism for anomaly detection, aspects that are particularly beneficial for industrial datasets like those from the Kirkuk plant. The cross-validation of time-series preserves the temporal integrity of the data, which is essential for accurate forecasting in time-dependent scenarios. Meanwhile, adaptive thresholding efficiently identifies potential anomalies, which is a critical feature for monitoring plant operational data. The tailored approach of the model to volatile and non-linear data emphasizes its superior performance and relevance in the fields of industrial data analysis.

5. CONCLUSIONS

This study successfully developed and validated an Enhanced Autoregressive Integrated Moving Average (E-ARIMA) model for anomaly detection in industrial sensory time-series data, focusing on vibration data from CA 202 accelerometers in the Kirkuk gas power plant. The performance of proposed E-ARIMA model was significantly superior, as evidenced by MAE of 0.0262, MSE of 0.0023, and RMSE of 0.0395. These results significantly outperformed those of traditional Seasonal ARIMA (SARIMA) and Vector Autoregressive models, which reported higher MAE, MSE, and RMSE values of 0.0348 and 0.034, 0.0041 and 0.0034, and 0.0486 and 0.0466, respectively. This development emphasizes the ability of the proposed E-ARIMA model to effectively manage the non-linear and dynamic nature of industrial sensory data, thereby increasing the accuracy of anomaly detection in complex environments. Nevertheless, the

scope of the current study focused on a unique industrial context and type of sensor data, may limit the broader applicability of the results. Future research should aim to extend the validity of the E-ARIMA model in different industries and sensory data to strengthen its global applicability.

6. ACKNOWLEDGMENTS

This paper has been supported by RUDN University Strategic Academic Leadership Program.

7. REFERENCES

1. Jahanshahi H, Cevik M, Başar A, editors. Predicting the number of reported bugs in a software repository. Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13–15, 2020, Proceedings 33; 2020: Springer.
2. Ξανθοπούλου ΓΒ. Forecasting power output of photovoltaic systems using machine learning techniques: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης; 2017.
3. Dorrani Z. Traffic Scene Analysis and Classification using Deep Learning. International Journal of Engineering. 2024;37(3):496-502. https://doi.org/10.5829/ije.2024.37.03c.06
4. Al Mallah R, Quintero A, Farooq B. Prediction of traffic flow via connected vehicles. IEEE Transactions on Mobile Computing. 2020;21(1):264-77. https://doi.org/10.1109/TMC.2020.3006713
5. Pan B, Demiryurek U, Shahabi C, editors. Utilizing real-world transportation data for accurate traffic prediction. 2012 ieee 12th international conference on data mining; 2012: IEEE.
6. Ilu SY, Prasad R. Time series analysis and prediction of COVID-19 patients using discrete wavelet transform and auto-regressive integrated moving average model. Multimedia Tools and Applications. 2024:1-19. https://doi.org/10.1007/s11042-024-18528-x
7. Pastén D, Czechowski Z, Toledo B. Time series analysis in earthquake complex networks. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2018;28(8). https://doi.org/10.1063/1.5023923
8. Vogel E, Saravia G, Pastén D, Muñoz V. Time-series analysis of earthquake sequences by means of information recognizer. Tectonophysics. 2017;712:723-8. https://doi.org/10.1016/j.tecto.2017.06.031
9. Moustra M, Avraamides M, Christodoulou C. Artificial neural networks for earthquake prediction using time series magnitude data or seismic electric signals. Expert systems with applications. 2011;38(12):15032-9. https://doi.org/10.1016/j.eswa.2011.05.043
10. Fahmi A-TWK, Kashyzadeh KR, Ghorbani S. A comprehensive review on mechanical failures cause vibration in the gas turbine of combined cycle power plants. Engineering Failure Analysis. 2022;134:106094. https://doi.org/10.1016/j.engfailanal.2022.106094
11. Poullikkas A. An overview of current and future sustainable gas turbine technologies. Renewable and Sustainable Energy Reviews. 2005;9(5):409-43. https://doi.org/10.1016/j.rser.2004.05.009
12. Tanaka K, Cavalett O, Collins WJ, Cherubini F. Asserting the climate benefits of the coal-to-gas shift across temporal and spatial scales. Nature climate change. 2019;9(5):389-96. https://doi.org/10.1038/s41558-019-0457-1
13. Meher-Homji CB, Chaker M, Bromley AF, editors. The fouling of axial flow compressors: Causes, effects, susceptibility, and sensitivity. Turbo Expo: Power for Land, Sea, and Air; 2009.
14. De Michelis C, Rinaldi C, Sampietri C, Vario R. Condition monitoring and assessment of power plant components. Power Plant Life Management and Performance Improvement: Elsevier; 2011. p. 38-109.
15. Mourad A-HI, Almomani A, Sheikh IA, Elsheikh AH. Failure analysis of gas and wind turbine blades: A review. Engineering Failure Analysis. 2023;146:107107. https://doi.org/10.1016/j.engfailanal.2023.107107
16. Mevissen F, Meo M. A review of NDT/structural health monitoring techniques for hot gas components in gas turbines. Sensors. 2019;19(3):711. https://doi.org/10.3390/s19030711
17. Qaiser MT. Data Analysis and Prediction of Turbine Failures Based on Machine Learning and Deep Learning Techniques: NTNU; 2023.
18. Yan W, Yu L. On accurate and reliable anomaly detection for gas turbine combustors: A deep learning approach. arXiv preprint arXiv:190809238. 2019. https://doi.org/10.48550/arXiv.1908.09238
19. Zhong S-s, Fu S, Lin L. A novel gas turbine fault diagnosis method based on transfer learning with CNN. Measurement. 2019;137:435-53. https://doi.org/10.1016/j.measurement.2019.01.022
20. Zhou D, Yao Q, Wu H, Ma S, Zhang H. Fault diagnosis of gas turbine based on partly interpretable convolutional neural networks. Energy. 2020;200:117467. https://doi.org/10.1016/j.energy.2020.117467
21. Kontopoulou VI, Panagopoulos AD, Kakkos I, Matsopoulos GK. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. Future Internet. 2023;15(8):255. https://doi.org/10.3390/fi15080255
22. Fahmi A-TWK, Kashyzadeh KR, Ghorbani S. Fault detection in the gas turbine of the Kirkuk power plant: An anomaly detection approach using DLSTM-Autoencoder. Engineering Failure Analysis. 2024;160:108213. https://doi.org/10.1016/j.engfailanal.2024.108213
23. Paparoditis Efstathios E, Politis DN. The asymptotic size and power of the augmented Dickey–Fuller test for a unit root. 2016. https://doi.org/10.1080/00927872.2016.1178887
24. Mushtaq R. Augmented dickey fuller test. 2011. https://dx.doi.org/10.2139/ssrn.1911068
25. Chodakowska E, Nazarko J, Nazarko Ł. Arima models in electrical load forecasting and their robustness to noise. Energies. 2021;14(23):7952. https://doi.org/10.3390/en14237952
26. Kashyzadeh KR, Ghorbani S. New neural network-based algorithm for predicting fatigue life of aluminum alloys in terms of machining parameters. Engineering Failure Analysis. 2023;146:107128. https://doi.org/10.1016/j.engfailanal.2023.107128
27. Reza Kashyzadeh K, Amiri N, Ghorbani S, Souri K. Prediction of concrete compressive strength using a back-propagation neural network optimized by a genetic algorithm and response surface analysis considering the appearance of aggregates and curing conditions. Buildings. 2022;12(4):438. https://doi.org/10.3390/buildings12040438
28. Rezaeian N, Gurina R, Saltykova OA, Hezla L, Nohurov M, Reza Kashyzadeh K. Novel GA-Based DNN Architecture for Identifying the Failure Mode with High Accuracy and Analyzing Its Effects on the System. Applied Sciences. 2024;14(8):3354. https://doi.org/10.3390/app14083354

Persian Abstract

چکیده

این مطالعه به معرفی یک مدل میانگین متحرک یکپارچه خودبازگشتی پیشرفته (E-ARIMA) را برای تشخیص ناهنجاری در داده‌های سری زمانی، با استفاده از ارتعاشات مانیتور شده توسط شتاب‌سنج‌های CA 202 در نیروگاه گازی کرکوک به عنوان مطالعه موردی می پردازد. هدف غلبه بر محدودیت‌های مدل‌های سنتی ARIMA در تحلیل ماهیت دینامیکی و غیرخطی داده‌های حسی صنعتی است. روش پیشنهادی جدید شامل آماده‌سازی داده‌ها از طریق درونیابی خطی برای پر نمودن به شکاف‌های مجموعه داده، تأیید ایستایی از طریق آزمون دیکی–فولر تقویت‌شده، و بهینه‌سازی مدل ARIMA در برابر معیار اطلاعات آکایک، با تکنیک اعتبارسنجی متقابل سری‌های زمانی تخصصی است. نتایج نشان می‌دهد که مدل E-ARIMA در تشخیص ناهنجاری نسبت به مدل‌های سنتی ARIMA مانند (SARIMA) و اتورگرسیو برداری عملکرد برتری دارد. در این راستا، از معیارهای میانگین خطای مطلق (MAE)، میانگین مربعات خطا (MSE) و ریشه میانگین مربعات خطا (RMSE) برای این ارزیابی استفاده شد. در نهایت، مهمترین دستآورد این تحقیق این است که نتایج دقت پیش‌بینی پیشرفته مدل E-ARIMA را برجسته می‌کند و آن را به ابزاری قوی برای کاربردهای صنعتی مانند پایش سلامت ماشین‌آلات تبدیل می‌کند، جایی که تشخیص زودهنگام ناهنجاری‌ها برای جلوگیری از خرابی‌های پرهزینه بسیار مهم است و برنامه‌ریزی تعمیر و نگهداری را تسهیل می‌نماید.