IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2583-2601

International Journal of Engineering

Journal Homepage: www.ije.ir

An Ensemble Deep Learning Approach for Automated Bone Fracture Detection in Medical Imaging

M. Mahtabi[a], S. Asadi Amiri*[a], S. Mavaddati[b]

[a] *Department of Computer Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran*
[b] *Department of Electronic Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran*

PAPER INFO

Paper history:
Received 02 October 2025
Received in revised form 02 November 2025
Accepted 31 Januray 2026

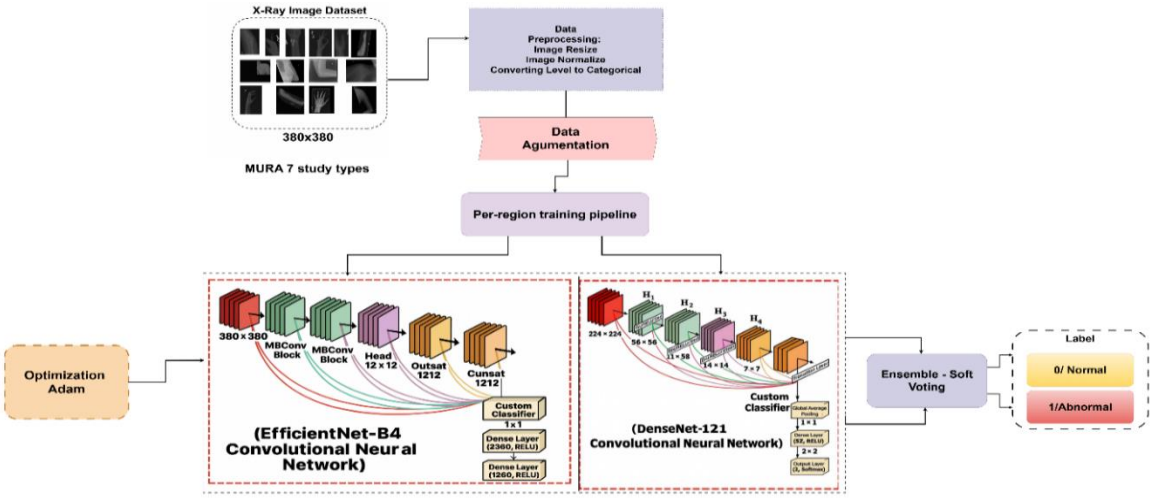
Keywords:
Deep Learning
Bone Abnormality Detection
Medical Image Analysis
Radiographic Images
EfficientNet-B4
DenseNet-121
Ensemble Learning

ABSTRACT

Recent advances in deep neural networks have significantly improved medical image analysis; however, detecting bone fractures in radiographic images remains challenging due to the complex skeletal structure, subtle fracture patterns, and limited annotated data. This study proposes an advanced ensemble deep learning framework that integrates two optimized architectures, EfficientNet-B4 and DenseNet-121, through a soft-voting fusion strategy to enable automated and precise bone fracture detection. The hybrid framework introduces adaptive weighting and optimized dense layers, which enhance feature discrimination and strengthen the network's capacity to distinguish fine-grained fracture details. Moreover, transfer learning and fine-tuning techniques are employed to address data imbalance and improve model generalization across multiple anatomical regions. Comprehensive experiments conducted on the MURA dataset, consisting of radiographs from seven distinct anatomical regions, demonstrate that the proposed model achieves superior performance with 83.52% accuracy and 90.76% sensitivity, outperforming each individual baseline. The model's robustness under different training configurations confirms its reliability and stability for clinical deployment. Overall, this research presents a novel ensemble-based diagnostic system that leverages complementary architectural strengths and adaptive feature fusion to achieve high diagnostic precision. The proposed method contributes not only to improving classification accuracy but also to establishing a scalable and interpretable framework for computer-aided fracture diagnosis, offering a practical step toward intelligent and reliable radiological decision support.

***doi**: 10.5829/ije.2026.39.10a.19*

Graphical Abstract



*Corresponding Author Email: s.asadi@umz.ac.ir (S. Asadi Amiri)

Please cite this article as: Mahtabi M, Asadi Amiri S, Mavaddati S. An Ensemble Deep Learning Approach for Automated Bone Fracture Detection in Medical Imaging. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2583-601.

1. INTRODUCTION

Bone fractures are among the most common orthopedic injuries, affecting millions of people worldwide each year. Increasing life expectancy and changing lifestyle patterns in industrialized societies have led to a considerable rise in the incidence of these injuries. Fractures typically occur when forces exceed the load-bearing capacity of bone and may result in from causes such as traffic accidents, falls, sports injuries, and underlying conditions like osteoporosis (1). Timely diagnosis is vital because delays in treatment can lead to serious consequences such as infection, permanent bone deformity, and loss of function (1).

Medical imaging modalities play a key role in the identification and assessment of fractures. X-ray radiography, computed tomography (CT), and magnetic resonance imaging (MRI) are among the primary options, chosen based on the type and location of the injury (1-3). Due to its speed, accuracy, and ability to reveal fine structural details, X-ray imaging is widely used to diagnose osseous problems and certain soft-tissue abnormalities. Figure 1 shows examples of radiographic images that illustrate bone abnormalities and fractures. Beyond its role in initial diagnosis, this modality is an important tool for the treatment planning and monitoring of recovery (2). Although X-rays are relatively safe, adherence to radiation safety principles remains necessary to avoid unnecessary exposure.

Musculoskeletal disorders, as a prevalent group of conditions, have a substantial impact on quality of life (4). These disorders, including compression fractures, carpal tunnel syndrome, and ligament tears, often arise from occupational incidents, repetitive movements, sports injuries, and poor ergonomic conditions. Occupations such as banking, software development, or hairdressing, which involve prolonged non-neutral postures, are associated with increased prevalence, particularly in the neck and lower back. Research indicates that both physical and psychosocial workplace factors play important roles in the emergence of these disorders (3, 4). Early diagnosis is especially important, since timely interventions such as physiotherapy and shock-wave therapy can effectively relieve mild-to-moderate pain and prevent disease progression. Nevertheless, accurate and timely diagnosis is often challenging and demands precise, sensitive diagnostic approaches.

To date, the detection of musculoskeletal abnormalities has largely been performed manually by radiologists. However, the complexity of shapes and the variability in lesion size, especially when patterns remain hidden in images, make the diagnostic process difficult and error-prone (5). In addition, the large volume of imaging data, together with factors such as fatigue and human error, can negatively affect accuracy and increase the likelihood of delays or mistakes in diagnosis (6, 7). Consequently, conventional approaches face serious limitations, and the need for modern, automated solutions is increasingly felt. In this context, advances in image processing and computer vision have provided powerful tools for automatic abnormality detection, which, by improving diagnostic speed and accuracy, can enhance clinical outcomes and reduce healthcare costs (8, 9).

In recent years, deep neural networks have been widely applied to bone-fracture detection. By learning complex patterns, these models can automatically highlight suspicious regions and, while reducing human error, improve both the accuracy and efficiency of diagnosis. They can also detect subtle fractures and provide quantitative analyses related to severity and displacement. Despite challenges such as the need for high-quality data and the risk of false results, for example, false positives, the application of deep learning in this area holds great potential for improving the quality of medical services and accelerating diagnosis.

In this paper, we present an innovative approach for more accurate detection of musculoskeletal abnormalities in radiographic images by combining two advanced models, EfficientNet-B4 and DenseNet-121, within an ensemble framework based on soft voting. Manual detection of these abnormalities is consistently challenging due to structural complexity, variability in lesion morphology, and fluctuating image quality, and thus requires approaches with higher accuracy and generalizability. Accordingly, our method leverages the strengths of deep models and intelligently fuses their outputs via soft voting to yield decisions that are more accurate, more stable, and more robust to noise and imaging variability.

To comprehensively evaluate performance, the models were trained separately on the seven anatomical regions of the MURA dataset. To address data imbalance, between 7,000 and 10,000 augmented images were generated per region, depending on its size. The outputs of the individual models for each region were then combined using soft voting, and the final accuracy for that region was obtained. In the final step, aggregating the results across all seven regions yielded an overall evaluation of the ensemble's performance, showing that our approach outperforms previous methods in terms of accuracy. These results corroborate the high efficacy of our hybrid model for automatic skeletal-lesion detection and its suitability for use in intelligent computer-aided diagnostic systems (10, 11). The main findings of this study are as follows:

- Improved performance over individual models: The proposed ensemble of EfficientNet-B4 and DenseNet-121 with soft voting outperformed either model used in isolation in detecting bone abnormalities.
- Notable accuracy: The model achieved an accuracy of 83.52%, demonstrating the strong capability of the

proposed framework in correctly classifying radiographs.

- High sensitivity: The sensitivity was 90.76%, indicating strong ability to detect positive cases (presence of abnormality) with a low miss rate.
- Enhanced high-level feature extraction: Adding customized dense layers at the end of each network enabled extraction of higher-level, more discriminative features, increasing classification power.
- Strong performance on MURA: Evaluation on the large MURA dataset, radiographs from seven anatomical regions, showed that the proposed model effectively detected skeletal abnormalities across regions.
- Clinical potential: The findings suggest that this framework can serve as a clinical decision-support tool, helping radiologists improve diagnostic accuracy and speed.

The novelty of this work lies in designing a hybrid ensemble framework that combines EfficientNet-B4 and DenseNet-121 through a soft-voting strategy to leverage their complementary feature extraction capabilities. In addition, customized dense layers were incorporated at the end of each network to improve high-level feature discrimination. This design provides a more generalizable and accurate model for bone abnormality detection than existing single-network approaches.

In the remainder of this paper, Section 2 reviews the literature and prior work on bone-abnormality detection using machine-learning and deep-learning methods. Section 3 details the proposed method, including the ensemble architecture based on EfficientNet-B4 and DenseNet-121, the customized classification layers, the dataset description, and the data-augmentation strategies. Section 4 presents experimental results and performance metrics and compares the proposed method with individual baselines. Finally, Section 5 offers discussion and conclusions, summarizing the findings and suggesting directions for future research.

2. LITERATURE REVIEW

Image processing and computer vision are among the most important branches of computer science in the analysis of medical images. The application of machine learning algorithms, intensive learning, enables the recognition of complex patterns in radiological images and the automatic detection of abnormalities. Research has shown that these approaches can increase diagnostic accuracy and speed, while supporting radiologists in clinical decision-making. Among modern techniques, convolutional neural networks (CNNs) hold a special place in medical image analysis due to their ability to extract local and complex features from images (12). Different CNN architectures, such as AlexNet, VGG, and ResNet, have been applied to various medical tasks.

Nevertheless, the lack of annotated datasets remains a major challenge in developing these models. To overcome this limitation, techniques such as data augmentation (rotation, cropping, color transformation), transfer learning from pre-trained models, and semi-supervised approaches are widely used. Furthermore, issues such as class imbalance, image noise, and the difficulty of result interpretation remain significant challenges. Despite these obstacles, CNNs have demonstrated remarkable performance in diagnosing diseases such as cancer, cardiovascular conditions, and neurological disorders. Karthik et al. (4) and Krizhevsky et al. (13) by using a deep CNN achieved groundbreaking success in the ImageNet competition, which paved the way for widespread CNN applications across domains, including medicine.

Ebsim et al. (14) introduced an innovative approach was that combined feature fusion with ensemble random forests for wrist fracture detection. By integrating diverse image features, this method significantly improved classification accuracy.

Chada (15) applied transfer learning with DenseNet and InceptionResNetV2 on the MURA dataset (finger and humerus, binary image-level). The best result on the humerus reached 88.2% accuracy, while on the finger, the performance was lower at only 77.7%. This shows the variability of CNN performance across different body parts.

Yahalom. et al. (16) used the Faster R-CNN model to detect distal radius fractures in posteroanterior wrist radiographs. Optimized with ImageNet pre-trained weights, the model achieved performance superior to that of radiologists in identifying fracture regions, despite being trained on a relatively small dataset.

The MSDNet model was introduced by Karthik et al. (4) combines multiple CNNs to enhance the detection and classification accuracy of radiographic abnormalities. In addition to identifying the presence or absence of abnormalities, it was also capable of localizing their exact position in the image and generating automated diagnostic reports with high accuracy.

A deep neural network was developed by Yadav and Rathor (17) to classify healthy and fractured bones. To improve performance and prevent overfitting, data augmentation techniques were employed. The results showed that this model achieved an accuracy above 92%, outperforming earlier approaches and demonstrating high capability in fracture detection.

Several methods for fracture detection and classification were reviewed by Meena and Roy (18), concluding that CNN-based architectures, particularly InceptionNet and XceptionNet, perform strongly in fracture detection. However, the lack of sufficiently large annotated datasets continues to be a barrier to the development of high-performance automated algorithms.

Spahr et al. (19) compared interpretability techniques such as Grad-CAM, attention maps, and Chefer across four medical datasets. The results showed that Chefer provided more accurate explanations, underscoring the importance of selecting appropriate interpretability methods in sensitive medical applications. Dahal et al. (20) proposed a hybrid model combining VGG-16 and Vision Transformer (ViT) for musculoskeletal abnormality detection on the full MURA dataset, achieving 82.88% accuracy and 0.88 sensitivity. The study highlights that integrating local and global feature extractors can improve automated radiographic diagnosis, though performance is still constrained by dataset limitations. Hardalaç et al. (21) used models such as Faster R-CNN, RetinaNet, and ResNet for automatic wrist fracture detection in X-ray images. By combining five models, the ensemble model WFD-C was introduced, which achieved strong performance with AP50 = 0.8639. Such models can serve as supportive tools in clinical diagnosis.

Beyaz et al. (22) combined three models, Xception, EfficientNet, and NfNet, using majority voting for hip fracture detection. Additionally, DenseNet-169, DenseNet-201, and InceptionResNetV2 were tested on the MURA dataset for detecting arm and finger abnormalities. The accuracy for fingers was lower due to disagreement among radiologists. Ghosh et al. (23) developed shallow neural networks and convolutional neural networks within an AdaBoost framework for humerus data, demonstrating that with an appropriate number of epochs and the use of ensemble models, performance close to that of radiologists and the 169-layer deep networks can be achieved, while significantly reducing training time and computational resources. In recent studies Lu et al. (24) have focused on CNN-based and hybrid models for fracture detection, with Inception/Xception showing strong performance despite data limitations. Hybrid CNN–Transformer models improved accuracy on MURA by combining local and global features. Building on this, HAMIL-Net used hierarchical attention and multiple-instance learning, achieving robust study-level results (AUC ≈ 0.87) for foot and ankle radiographs.

The main objective of the present study is to develop a CNN-based model for the automatic detection of musculoskeletal radiographic abnormalities with improved performance. To achieve this, an ensemble framework was designed, combining two advanced architectures, EfficientNet-B4 and DenseNet-121, in a jointly optimized manner to enhance both feature extraction and classification accuracy. Furthermore, the use of a soft-voting mechanism to integrate the results of individual models increases system stability and improves final detection accuracy. Customized dense layers were added to the end of each network to enhance learning capacity and extraction of high-level, discriminative features, thereby boosting model performance in abnormality detection. A comprehensive evaluation of the proposed framework was performed on the MURA dataset, which covers seven different anatomical regions, demonstrating the model's applicability in real-world, multi-purpose medical contexts. Ultimately, this approach not only maintains high accuracy but also offers the capability to be integrated as a clinical decision-support tool to accelerate diagnostic processes in radiology.

In recent years, several studies have explored the application of deep learning models for detecting bone fractures and abnormalities. A summary of the most relevant works is presented in Table 1.

3. DETECTION OF BONE FRACTURES AND ABNORMALITIES

In this section, the detailed procedure of the proposed system for detecting musculoskeletal abnormalities in radiographic images is described. The overall framework is based on ensemble learning, and by combining two powerful deep models, EfficientNet-B4 and DenseNet-121, it seeks to simultaneously leverage their high accuracy and generalization capability.

The overall process consists of the following stages: data preparation, design and customization of the model architectures, independent training of the models, and, finally, fusion of the models' outputs via soft voting to increase the accuracy of the final prediction.

3. 1. Dataset In studies related to detecting musculoskeletal disorders through image processing and computer vision, the use of benchmark datasets is very common and practical (25). The MURA dataset was developed specifically for training and evaluating deep learning models, particularly convolutional neural networks, for detecting musculoskeletal disorders from radiographic images. This dataset is one of the largest radiographic image datasets for musculoskeletal disorder detection. It was prepared by Stanford University in the United States and is widely used in computer vision and deep learning research on medical image analysis. The dataset contains more than 40,000 radiographic images from approximately 14,000 studies. These studies include radiographs from different body regions such as wrist, arm, elbow, shoulder, finger, forearm, and others. All images in this dataset are radiographs collected to investigate musculoskeletal problems and disorders. Each image in the dataset is categorized into one of two classes: "normal" or "abnormal" (presence of an abnormality). The labeling was performed by specialist radiologists, and these labels serve as the main basis for training machine learning models.

The dataset is publicly and freely available to researchers and can be downloaded from reputable platforms such as GitHub or websites related to Stanford University (25). Using this large dataset, together with the accuracy of deep learning models, improves the detection of musculoskeletal abnormalities and enables comparison of results across studies. The dataset includes 9,045 normal images and 5,818 abnormal images labeled by radiologists, and with diverse images from different body parts, poses challenges for automatic detection algorithms. The distribution of studies in the MURA dataset for normal and abnormal classes is summarized in Table 2.

3. 2. Preprocessing and Data Augmentation in the Proposed Method Deep learning models require diverse data for accurate detection of bone fractures. Although the MURA dataset is large, it does not provide sufficient diversity to cover all fracture cases and inter-patient variations. Incorporating additional data can prevent overfitting and improve the generalizability of the model (26-30). This is particularly important due to the inherent imbalance between "Normal" (healthy) and "Abnormal" or fractured samples in medical datasets. In fact, when the number of samples across classes is imbalanced, the model may become biased toward the majority class and show weakness in correctly

TABLE 1. Recent studies on bone fracture and abnormality detection using deep learning models.

Reference	Classes	Dataset	Classifier	Year
(13)	1000 classes (animals, objects, scenes, etc.)	ImageNet (ILSVRC)	AlexNet, ResNet, VGG	2012
(12)	Normal vs. abnormal (2 classes)	MURA V1.1	DenseNet-169	2018
(14)	Normal vs. abnormal (2 classes)	Wrist dataset	RF-CLM	2018
[15]	Normal vs. abnormal (2 classes)	MURA V1.1	DenseNet-169 ،DenseNet-201 ، InceptionResNetV2	2019
(16)	Normal vs. abnormal (2 classes)	Distal Radius Wrist X-ray dataset	MSDNet (Ensemble CNN), Faster R-CNN	2020
(23)	Normal vs. abnormal (2 classes)	MURA V1.1	Shallow NN, CNN + AdaBoost Ensemble	2021
(4)	Normal vs. abnormal (2 classes)	MURA + Indiana	MSDNet (Ensemble CNN)	2022
(24)	Normal vs. abnormal (2 classes)	Foot & Ankle Radiographs	HAMIL-Net (Hierarchical Attention + MIL)	2023
(22)	Normal vs. abnormal (2 classes: hip fracture vs. healthy)	Hip Fracture X-ray dataset (10,849 images, 5 hospitals)	Xception, EfficientNet, NfNet + Majority Voting	2023
(20)	Normal vs. abnormal (2 classes)	MURA V1.1	VGG-16 and Vision Transformer (ViT)	2024
(18)	Normal vs. abnormal (2 classes)	Hip Fracture X-ray (10,849 images), MURA, CP-Child, Duke Breast Cancer, Kvasir GI Disease (Z-line subset)	Vision Transformers	2025
Proposed Method	Normal vs. abnormal (2 classes)	MURA V1.1	EfficientNet-B4, DenseNet-121	–

TABLE 2. Distribution of studies in the MURA dataset with 9045 cases in the normal class and 5818 cases in the abnormal class

Region (Body Part)	Total Images	Validation (Normal)	Validation (Abnormal)	Training (Normal)	Training (Abnormal)
Elbow	1,912	92	66	1094	660
Finger	2,110	92	83	1280	655
Hand	2,185	101	66	1,497	521
Arm (Humerus)	727	68	67	321	271
Forearm	1,010	69	64	590	267
Shoulder	3,015	99	95	1364	1457
Wrist	3,697	140	97	2134	1326
Total	14,656	661	538	8280	5177

Figure 1. Examples of radiographic images from the MURA dataset illustrate the variety of abnormalities in different regions of the upper limb. Each image pair represents a single abnormal case. These images highlight challenges such as fractures and other anomalies, as well as the presence of medical devices that the model needs to handle

TABLE 3. Number of images in the dataset before and after augmentation

Anatomical Region	# Before Augmentation	Target per Class	# After Augmentation
Elbow	4437	4000	8062
Finger	4595	20000	39319
Forearm	1641	5000	10060
Hand	4988	10000	19591
Arm (Humerus)	1144	10000	19382
Shoulder	7540	10000	20030
Wrist	9441	7000	14057
Total	**33786**	–	**130501**

identifying the minority class (which is often the more clinically significant one).

In this study, data augmentation was applied online and simultaneously during training. The augmentation operations included a set of geometric and standard compression transformations, described as follows:

- Random rotation: images were rotated at random angles up to ±180°.
- Horizontal and vertical translation: up to 10% of the image width and height.
- Zooming: up to 10% to simulate scale variations.
- Flipping: both horizontally and vertically to increase orientation diversity.

This procedure was performed independently for each of the seven anatomical regions, increasing the number of training samples per class in each region to approximately 5,000 to 20,000 images. This process not only helped balance the classes but also forced the model to learn invariant features that are robust against these transformations. The number of images in the dataset before and after data augmentation is reported in Table 3. After augmentation, two main preprocessing steps were carried out to match the requirements of pre-trained neural networks:

- Pixel normalization using the standard functions of DenseNet or EfficientNet.
- Image resizing to 380×380 pixels to fit the architecture input.

These preprocessing steps enabled the model to learn features independent of position and shape, providing more stable and accurate performance when dealing with diverse real-world radiographs.

3. 3. Designed Architecture of EfficientNet-B4

The EfficientNet-B4 model is one of the advanced members of the EfficientNet family, a series of convolutional neural network architectures designed using the compound scaling approach. Unlike traditional methods that scale only one dimension of the network, for example, depth, width, or resolution, compound scaling expands all three dimensions simultaneously and in a balanced manner. This strategy increases efficiency and enhances model performance in visual classification tasks. EfficientNet-B4 is pre-trained on the large-scale ImageNet dataset, providing substantial transfer learning capability for visual pattern recognition. Its architecture is composed of hierarchical MBConv blocks, which reduce computational complexity and the number of parameters while maintaining the ability to extract high-level features. In this study, to adapt EfficientNet-B4 to the specific task of bone abnormality detection, the base structure of the model was modified in the classifier head. As shown in Figure 2, after the global average pooling layer, two dense layers with 2,560 and 1,200 nodes were added. The inclusion of these layers allows the extraction of more complex and specialized patterns from high-level features.

This structural adaptation enables the model to more effectively distinguish subtle differences between "Normal" and "Abnormal" images, thereby improving the overall classification performance (26).

3. 4. Designed Architecture of DenseNet-121

DenseNet-121 is one of the prominent models in the family of Dense Convolutional Networks, which revolutionized neural network design by introducing direct connections among all layers within a block. In the original DenseNet architecture, each layer passes its output not only to the subsequent layer but also to all the following layers. This dense connectivity mechanism enhances feature reuse, strengthens gradient flow during training, and significantly reduces the number of parameters compared to conventional architectures such as ResNet. The structure of DenseNet consists of dense blocks and transition layers. Within each dense block, the feature maps of the layers are concatenated along the channel dimension, a method that, unlike the additive operation in ResNet, preserves information and expands feature diversity. Transition layers, through convolution

and down-sampling operations, compress the feature map volume and prevent excessive growth of dimensions. In this study, the DenseNet-121 model, pre-trained on ImageNet, has been purposefully customized for optimizing radiographic image classification. As illustrated in Figure 3, following the global average pooling layer, two fully connected layers with 1024 and 512 nodes were added, serving as intermediaries between the feature extractor and the final output. These layers enhance the model's capability to learn complex nonlinear relationships between image features and classification labels (27).

This modified design is not only aligned with the specific objectives of the research but also preserves the ability of DenseNet's deep and rich structure to extract multi-level features from bone structures. Consequently, the optimized DenseNet-121, as part of the proposed hybrid framework, makes a substantial contribution to improving the accuracy of abnormality detection (27).

As mentioned, Figure 3 presents the detailed architecture of the customized DenseNet-121 model used in the proposed ensemble framework. In this design, the standard DenseNet-121 backbone is enhanced with additional dense layers at the end to improve the extraction of high-level and discriminative features from radiographic images.

The model processes images from various anatomical regions of the upper limb and outputs class probabilities for Normal and Abnormal categories, which are later combined in the ensemble model. This Figure provides a clear visualization of how the DenseNet-121 architecture is adapted for musculoskeletal abnormality detection.

Table 4 illustrates the stage-wise process of image processing in both architectures. Both models start with the same input layer and, through feature extraction blocks, progressively reduce the spatial dimensions of the feature maps while increasing their depth. The main difference lies in the structure and number of blocks in each architecture. The critical stage is the customized classifier, which includes a global average pooling layer and two fully connected layers for extracting high-level features. In EfficientNet-B4, the sizes of these layers are 2560 and 1200, whereas in DenseNet-121, they are 1024 and 512, respectively. Finally, the output layer with Softmax predicts the probability of an image belonging to the "normal" or "abnormal" class, and both models are optimized for bone abnormality detection.

3. 5. Soft Voting in the Proposed Method

Although EfficientNet-B4 and DenseNet-121 are well-established pretrained architectures, in this work, they were individually fine-tuned on the MURA dataset and enhanced with customized dense layers to improve high-level feature extraction. The outputs of the two optimized models were then combined through a soft-voting ensemble strategy to leverage their complementary strengths and achieve higher detection accuracy.

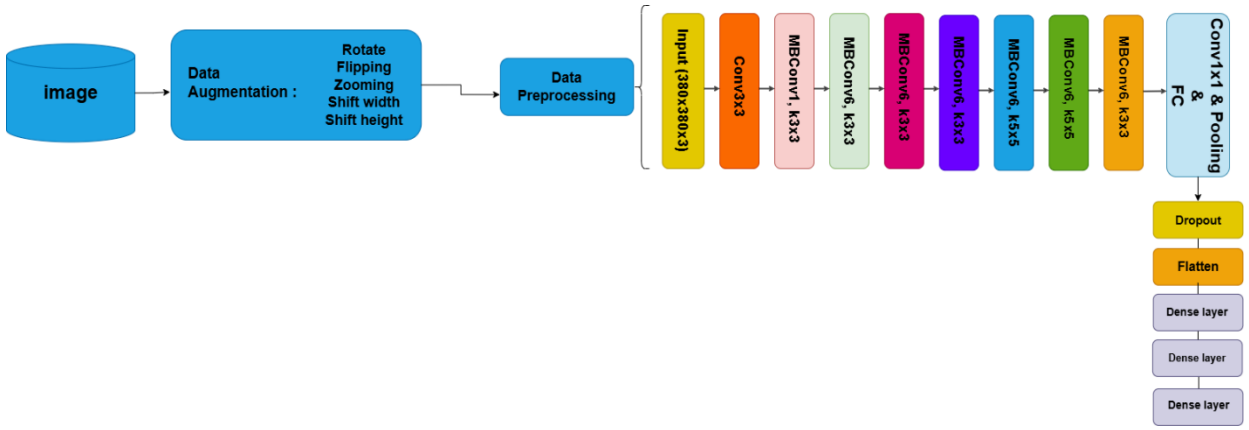


Figure 2. Schematic of the EfficientNet-B4 architecture as designed in the proposed method

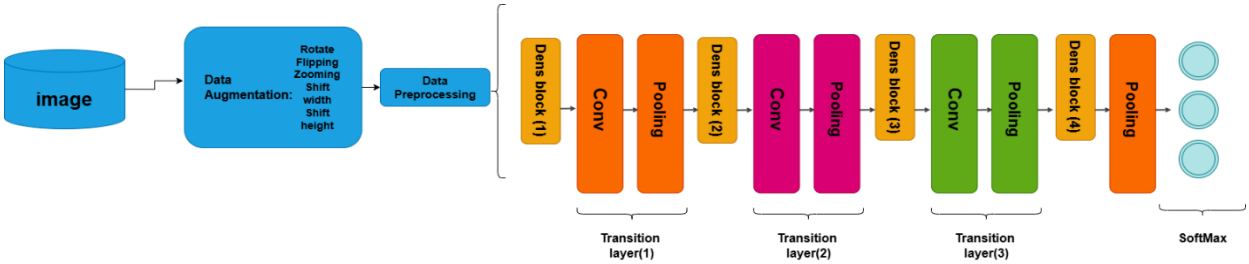


Figure 3. Overview of the architecture of the customized DenseNet-121 model in the proposed method.

TABLE 4. Comparative overview of the designed architectures of EfficientNet-B4 and DenseNet-121. This table details the stage-wise structure of each model, from the input layer to the final classification head, which has been optimized with customized fully connected layers

Stage	Designed EfficientNet-B4 Architecture	Designed DenseNet-121 Architecture
Input	Size: 380×380 Details: Image input	Size: 380×380 Details: Image input
Initial Conv/Pool	Size: 190×190 Details: Stem (Conv, BN, Swish)	Size: 95×95 Details: Conv1, Pool1
Feature Extraction 1	Size: 95×95 Details: 4 × MBConv Block	Size: 47×47 Details: Dense Block (1), Transition Layer (1)
Feature Extraction 2	Size: 48×48 Details: 4 × MBConv Block	Size: 23×23 Details: Dense Block (2), Transition Layer (2)
Feature Extraction 3	Size: 24×24 Details: 6 × MBConv Block	Size: 11×11 Details: Dense Block (3), Transition Layer (3)
Feature Extraction 4	Size: 12×12 Details: 10 × MBConv Block + Head	Size: 11×11 Details: Dense Block (4) + Classification Layer
Custom Classifier	Size: 1×1 Details: Global Average Pooling Dense Layer (2560, ReLU) Dense Layer (1200, ReLU) Output Layer (2, Softmax)	Size: 1×1 Details: Global Average Pooling Dense Layer (1024, ReLU) Dense Layer (512, ReLU) Output Layer (2, Softmax)

Soft voting is one of the common techniques in the field of ensemble learning, where multiple machine learning models, known as base learners, are trained in parallel, and their outputs are combined to improve the overall performance of the system (29).

Unlike individual models that make decisions solely based on their internal structure, ensemble models benefit from the diversity of architectures, perspectives, and capabilities of each model to enhance prediction accuracy (30).

Among the methods for combining model outputs, hard voting and soft voting are the two main approaches. Hard voting operates purely based on the majority vote (the class selected by each model) and assigns equal importance to all models. However, this approach has limitations, including the disregard for the confidence level of the models in their predictions. In contrast, soft voting takes into account the probability distributions produced by each model, providing a more precise and sensitive approach. In this method, the final prediction is determined by averaging the predicted probabilities for each class across the base models, which leads to more informed decision-making. This approach reduces the likelihood of errors when handling inconsistent outputs among models and results in greater stability and higher classification accuracy (31).

In the proposed framework of this study, the EfficientNet-B4 and DenseNet-121 models were independently trained on the dataset, and their outputs were extracted as probability vectors corresponding to the two target classes (normal and abnormal). By performing element-wise averaging of these vectors, the final output of the combined system was generated. This ensemble strategy, based on soft voting, leverages the distinctive features of each model to enhance overall performance while increasing robustness against noise, data inconsistencies, and structural variability in radiographic images. Ultimately, the application of soft voting in this research not only improved the accuracy of the hybrid model but also demonstrated its potential as a practical and reliable approach for integration into intelligent computer-aided medical diagnosis systems.

The block diagram of the proposed method is shown in the Graphical Abstract Section for automated bone fracture and abnormality detection. This diagram shows the step-by-step process starting from input radiographic images, followed by preprocessing, augmentation, and feature extraction using the customized DenseNet-121 and EfficientNet-B4 models. The outputs of these base models are then combined through a soft-voting ensemble mechanism to generate the final classification for each anatomical region as Normal or Abnormal. This Figure provides a concise visual summary of the methodology, highlighting how the ensemble framework integrates multiple deep learning architectures for robust detection of musculoskeletal abnormalities.

Algorithm 1 illustrates the proposed ensemble deep learning framework for the automatic detection of musculoskeletal abnormalities in radiographic images. The algorithm takes as input radiographs from different anatomical regions, including the elbow, finger, hand, arm, forearm, shoulder, and wrist. Each image is first preprocessed by resizing, normalization, and optional augmentation to enhance model generalization. The dataset is then divided into training and validation sets for each body part. Two optimized deep learning

architectures, EfficientNet-B4 and DenseNet-121, serve as base models. Their final layers are customized with dense layers to improve feature extraction and discrimination. Both models are trained separately on the training set and evaluated on the validation set. The ensemble model is then constructed using a soft-voting mechanism that combines the predictions of the base models to produce a final probability score for each class (Normal or Abnormal). For each validation image, the ensemble model predicts the likelihood of abnormality, which is then thresholded to assign a class label. The performance metrics, including accuracy, sensitivity, precision, and F1-score, are computed separately for each anatomical region, allowing detailed evaluation of the model's effectiveness across different parts of the upper limb.

3. 6. Relations in the Proposed Method This framework provides a robust and clinically relevant approach for automated musculoskeletal abnormality detection, facilitating faster and more accurate radiological assessments.

In the training process of the proposed model for classifying shoulder X-ray images, the categorical cross-entropy loss function was employed.

This function is one of the most common and effective loss functions in deep learning for binary or multi-class classification tasks, particularly when labels are represented using one-hot encoding. Its primary purpose is to measure the discrepancy between the probability distribution predicted by the model and the actual label distribution. In other words, it evaluates how well the model's predictions match the ground truth. The categorical cross-entropy loss is mathematically formulated as:

$$L = -\sum_{i=1}^{C} t_i \log(P_i) \quad (1)$$

where t_i denotes the true label for class i, P_i represents the probability predicted by the model's Softmax output, and C is the number of classes. In this study, given the two classes "Normal" and "Abnormal," the loss function encourages the model to assign the highest probability to the correct class. Incorrect predictions with high confidence result in higher loss, guiding the model toward improved accuracy.

For optimization, the Adaptive Moment Estimation (ADAM) algorithm was utilized. By leveraging both momentum and adaptive learning rates, ADAM updates the network weights to minimize the loss efficiently (28).

Let $\tilde{X}_i$ denote the preprocessed input image. In the proposed method, two optimized deep learning architectures, DenseNet-121 and EfficientNet-B4, are employed as base models to extract discriminative features from the input image. The outputs of the two base models, DenseNet-121 and EfficientNet-B4, for image i are given by:

$$\begin{aligned} y_i^{(1)} &= f_{DensNet}(\tilde{X}_i; \Theta_1), \\ y_i^{(2)} &= f_{EfficientNet}(\tilde{X}_i; \Theta_2) \end{aligned} \quad (2)$$

where Θ_1 and Θ_2 represent the learnable parameters of DenseNet-121 and EfficientNet-B4, respectively. These outputs correspond to the predicted probabilities for the "Normal" and "Abnormal" classes and form the basis for the subsequent ensemble prediction. The ensemble output combines the predictions of the base models using a soft-voting mechanism:

$$y_i^{Ensemble} = \alpha y_i^{(1)} + (1-\alpha) y_i^{(2)} \quad (3)$$

where $\alpha \in [0,1]$ is the weight assigned to DenseNet-121 and $(1-\alpha)$ is the weight for EfficientNet-B4.

In this paper, the weight parameter α in the soft-voting ensemble was empirically determined using a validation subset of the MURA dataset. Several candidate values were tested, and the value that maximized overall accuracy and sensitivity was selected to optimally balance the contributions of DenseNet-121 and EfficientNet-B4.

The predicted class for image iii is determined by selecting the class with the highest probability:

$$\hat{C}_i = arg \max_{c \in \{\text{Normal,Abnormal}\}} y_i^{Ensemble}[c] \quad (4)$$

The weights of both models are updated during training using the ADAM optimizer, which combines momentum

Algorithm 1: Ensemble Deep Learning Framework for Musculoskeletal Abnormality Detection

Input: Radiographic images from MURA dataset (Elbow, Finger, Hand, Arm, Forearm, Shoulder, Wrist)
Output: Classification result for each region (Normal / Abnormal)

1: Load MURA dataset
2: For each image I in the dataset do
3: Preprocess I:
- Resize to fixed dimensions
- Normalize pixel intensities
- Apply augmentation (rotation, flipping, zoom, etc.)

4: Divide the dataset into a Training set and a Validation set for each body part
5: Initialize base models: EfficientNet-B4 and DenseNet-121
6: For each base model M do
7: Replace final layers with customized dense layers
8: Train M on the Training set of each body part
9: Validate M on Validation set
10: End For
11: Construct an Ensemble model using a soft-voting mechanism:
- Input: Predictions from EfficientNet-B4 and DenseNet-121
- Output: Weighted probability scores for each class (Normal / Abnormal)

12: For each image I in the Validation set do
13: Predict abnormality probability using an Ensemble model
14: Assign class label:
If probability ≥ threshold, then Abnormal
Else Normal
15: End For
16: Compute performance metrics (Accuracy, Sensitivity, Precision, F1-score) separately for each body part
17: Return classification results for all anatomical regions

and adaptive learning rates to minimize the categorical cross-entropy loss efficiently.

4. EXPERIMENTS AND RESULTS

The effectiveness of the proposed method was evaluated through a comprehensive experimental study conducted on the MURA dataset. In the following sections, we first describe the performance metrics used for evaluating the method. Next, a detailed explanation of the dataset employed in this study is provided. Finally, we present an in-depth analysis of the experimental results, offering insights into the performance and potential implications of the proposed approach.

4. 1. Performance Metrics Metric rationale. Given the class imbalance and the clinical cost of missed fractures, we report Accuracy together with Sensitivity (Recall), Precision, F1, Specificity, and Cohen's κ. Sensitivity captures clinical safety, F1 balances Precision–Recall under imbalance, and κ reflects agreement beyond chance across regions with varying prevalence. We also report ROC-AUC (and scope-matched PR-AUC where noted) to assess threshold-free discrimination. These metrics are derived from the confusion matrix and provide a comprehensive view of the model's reliability and correctness in image classification (32, 33). A brief description of each metric is provided below:

$$Acc = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

$$Spe = TN/(TN+FP) \quad (6)$$

$$Ppr = TP/(TP+FP) \quad (7)$$

$$Sen = TP/(TP+FN) \quad (8)$$

$$F\text{-}Measure=2\times(Ppr\times Sen)/(Ppr+Sen) \quad (9)$$

In these relations, TP denotes the number of positive samples correctly classified as positive, FN represents the number of positive samples incorrectly classified as negative, FP indicates the number of negative samples incorrectly classified as positive, and TN refers to the number of negative samples correctly classified as negative.

Accuracy reflects the proportion of samples correctly classified (both positive and negative) out of the total number of samples. This metric represents the overall percentage of correct predictions made by the model compared to the ground truth labels.

Sensitivity (or recall) measures the percentage of actual positive samples correctly identified as positive. This metric indicates the model's ability to detect positive instances, such as abnormal images, correctly.

Specificity measures the percentage of actual negative samples correctly identified as negative. This metric reflects the model's capability to avoid misclassifying negative instances as positive (34, 35).

Additionally, Cohen's Kappa coefficient is a measure of agreement between the model's predictions and the actual results, taking into account the effect of random agreement. This metric is particularly important in binary classification problems, such as "normal" versus "abnormal" cases. The Kappa value ranges from -1 to +1, where 1 indicates perfect agreement, 0 corresponds to agreement by chance, and negative values signify agreement worse than chance (30).

$$Kappa = \frac{\Pr(a)-\Pr(c)}{1-\Pr(c)} \quad (7)$$

Here, $\Pr(a)$ represents the observed agreement between the model's predictions and the ground truth labels, while $\Pr(c)$ denotes the expected agreement purely by chance.

In this study, in addition to the aforementioned metrics, the confusion matrix was employed for a more detailed analysis of the model's performance. This matrix provides precise information regarding the number of correctly and incorrectly classified samples, offering a deeper understanding of the model's strengths and weaknesses in image classification.

Furthermore, the ROC curve was introduced to illustrate the model's capability to distinguish between positive and negative samples across different thresholds. The area under the curve (AUC) serves as a comprehensive metric for evaluating the overall performance of the model in musculoskeletal radiographs classification.

These metrics collectively provide a precise and thorough assessment of the system's ability to identify and differentiate various types of bone fractures, offering an overall view of the model's effectiveness in managing the complexity and inherent variability of bone fracture images.

4. 2. Experimental Results of the Proposed Method In this section, the results of evaluating the proposed method on the MURA dataset are presented. The proposed model, designed for detecting abnormalities in bone radiographs, was developed based on the combination of two advanced and customized neural networks: EfficientNet-B4 and DenseNet-121. The models were trained using the Adam optimizer and the categorical cross-entropy loss function. The training process was conducted independently for each of the seven anatomical regions, and the best weights were stored according to validation performance.

The final architecture employed an ensemble model, in which the predictions of the two networks were integrated through a soft-voting mechanism. In this approach, for each anatomical region, the final output

was calculated as the weighted average of the class probabilities predicted by both models. Table 5 summarizes the hyperparameters applied for training the EfficientNet-B4 and DenseNet-121 models on musculoskeletal radiographs from the MURA dataset. This includes settings such as learning rate, batch size, optimizer choice, number of epochs, and any regularization techniques, providing a clear overview of the configurations used to achieve optimal model performance. The hyperparameters used for training EfficientNet-B4 and DenseNet-121 on the MURA musculoskeletal radiographs are reported in Table 6. Both models were trained with a learning rate of 0.001, a batch size of 8, and for 15 epochs. The Adam optimizer with an epsilon of 1.0 was employed, and transfer learning from ImageNet was used for weight initialization. Softmax was chosen as the activation function, and categorical cross-entropy served as the loss function.

These hyperparameters were selected empirically based on stability, convergence behavior, and prior experience with similar medical imaging tasks, ensuring balanced learning and optimal performance across the models. Table 6 presents the classification performance of the proposed ensemble model with soft voting across different anatomical regions in the MURA dataset. The results show that the model achieves robust overall accuracy (83.52%) with a balanced precision (75.62%), high sensitivity (90.76%), and reasonable specificity (78.66%). Among individual regions, the elbow and humerus (arm) achieved the highest accuracies of 87.31% and 87.50%, respectively, indicating strong reliability in these categories. In contrast, hand and finger regions obtained relatively lower precision values, which may reflect the higher structural variability and complexity of these areas. Notably, the consistently high sensitivity across all regions highlights the model's strong capability in correctly identifying abnormal cases, while specificity values suggest some challenges in distinguishing normal samples, especially in smaller anatomical regions.

These findings confirm the effectiveness of the ensemble model while also pointing to potential areas for further refinement in clinical applications.

TABLE 5. Hyperparameters used for training EfficientNet-B4 and DenseNet-121 on musculoskeletal radiographs from the MURA dataset.

Hyperparameters	EfficientNet-B4	DenseNet-121
Learning rate	0.001	0.001
Batch size	8	8
Epochs	15	15
Optimizer	Adam(epsilon = 1.0)	Adam(epsilon = 1.0)
Weight initialization	Transfer learning (ImageNet)	Transfer learning (ImageNet)
Activation function	Softmax	Softmax
Loss function	Categorical Cross Entropy	Categorical Cross Entropy

TABLE 6. Classification results of the ensemble model (with soft voting) for each region in the MURA dataset.

Anatomical Region	Accuracy	Precision	Sensitivity	Specificity
Elbow	87.31	80.87	93.62	74.59
Finger	80.69	77.33	84.58	61.47
Forearm	82.06	72.19	92.00	64.14
Hand	80.65	61.90	93.73	58.19
Arm (Humerus)	87.50	88.57	86.49	75.00
Shoulder	81.17	76.26	85.96	62.30
Wrist	85.74	73.90	95.33	70.59
Overall Average	**83.52**	**75.62**	**90.76**	**78.66**

4. 3. Analysis of Results by Anatomical Region

In this chart, the performance results of the proposed ensemble model for each of the seven anatomical regions are presented. These regions include the elbow, finger, forearm, hand, arm, shoulder, and wrist. Each region was evaluated separately using the metrics of accuracy, sensitivity, and F1-score.

The overall performance of the model indicates that the proposed ensemble demonstrated stable and strong results across most regions. In particular, in six out of the seven regions, the F1-score was higher than 0.80, which reflects the high reliability and generalizability of the model.

The best performance was achieved in the elbow and arm regions, with F1-scores of 0.87 and 0.86, respectively. This result highlights the strong ability of the model to detect abnormalities in these regions, which may be due to the simpler bone structures and clearer fracture patterns found there.

The hand region, however, proved to be the most challenging, with an F1-score of 0.72. This performance resulted from a significant imbalance between high precision (0.87) and low sensitivity (0.62). This finding suggests that the model acts very cautiously when identifying fractures in the hand region. The likely reason is the structural complexity of the hand, with its multiple small bones and subtle fractures, which makes accurate detection more difficult.

In some cases, misclassifications occurred due to inherent challenges in the dataset. These include the complexity of small bone structures, subtle fractures that are visually similar to normal anatomy, and image

artifacts or noise present in radiographs. Understanding these factors can guide future improvements in model design and data preprocessing to further enhance classification accuracy.

4. 4. Experimental Results and Performance Analysis In this section, the performance of the proposed and baseline models on the MURA dataset is thoroughly analyzed and reported. This analysis aims to carefully examine the effectiveness of the proposed framework in detecting skeletal abnormalities in radiographic images across seven anatomical regions. As explained earlier, the two base architectures, EfficientNet-B4 and DenseNet-121, were trained independently, and their predictions were then combined into an ensemble model using the soft-voting method.

While some prior papers report higher Cohen's κ, cross-study κ comparisons are fragile because κ is sensitive to class prevalence, labeling variability, and evaluation scope (image- vs study-level) and subset coverage (single region vs all seven). Under our all-regions, image-level protocol, $\kappa = 0.68$ coexists with high Sensitivity (90.76%) and robust AUC (0.89), reflecting a recall-oriented operating point desirable in screening. Regions with small, complex anatomy (e.g., hand/finger) naturally depress κ despite strong discrimination, whereas simpler regions (e.g., elbow, humerus) yield higher per-region scores, consistent with our accuracy peaks (87.31% elbow; 87.50% humerus).

To evaluate performance, the following metrics were employed: accuracy, precision, sensitivity (recall), F1-score, and area under the ROC curve (AUC). In the next step, the overall performance of the models was assessed based on all test data from the seven anatomical regions. The comparison of precision, recall, and F1-score for the ensemble model across different upper-limb regions is presented in Figure 4.

Figure 5 presents the confusion matrices for the three models. Analysis of these matrices reveals the following:

- The DenseNet-121 model shows higher accuracy in predicting the *normal* class compared to the others, but` it is weaker in detecting the *abnormal* class.
- The EfficientNet-B4 model performs better in detecting abnormal samples, though it has a slightly higher error rate in the normal class.
- The ensemble model achieves the best balance between true positives (TP) and true negatives (TN), resulting in the lowest error rate across both classes.

In this section, the ROC curves for the ensemble model across each anatomical region are presented. As shown in Figure 6, the ROC curves illustrate the relationship between true positives (TP) and false positives (FP) at different classification thresholds.

A strong discriminative power is observed: as the chart indicates, the ROC curves for all regions lie significantly above the diagonal chance line (AUC = 0.5). This demonstrates the high capability of the model in distinguishing between normal and abnormal classes across all anatomical regions.

As shown in Figure 7, the ROC curves comparing DenseNet-121, EfficientNet-B4, and the Soft Voting Ensemble further highlight that while each model performs well, the ensemble consistently achieves a more balanced and superior discriminative power across test data.

AUC analysis: The area under the curve values for different regions show that the elbow achieved the highest AUC of 0.93, representing the best performance, whereas the hand region, with an AUC of 0.85, was identified as the most challenging. These results align well with the F1-score outcomes, confirming that simpler regions, such as the elbow, yield considerably better performance.

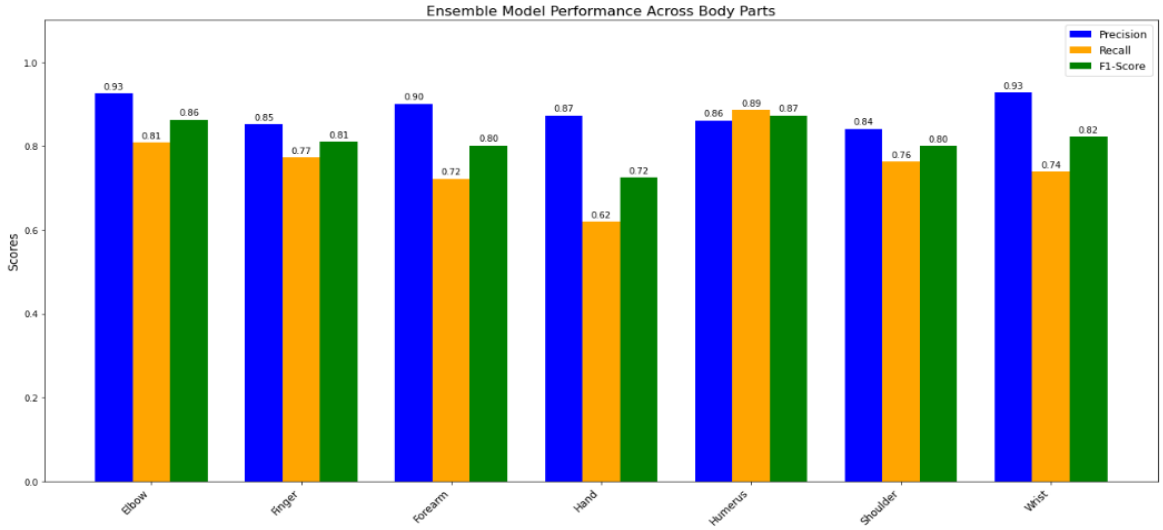
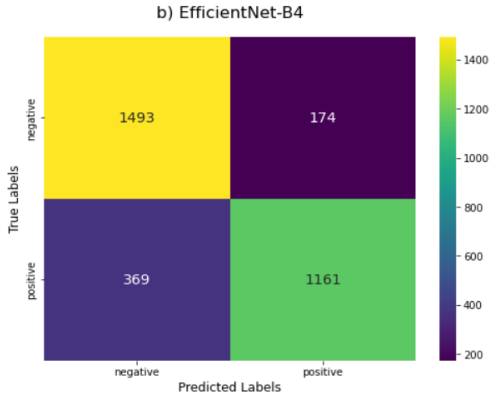
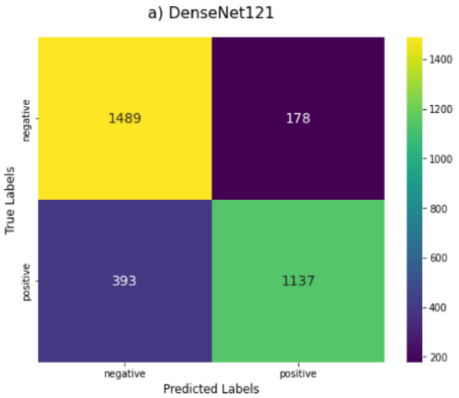


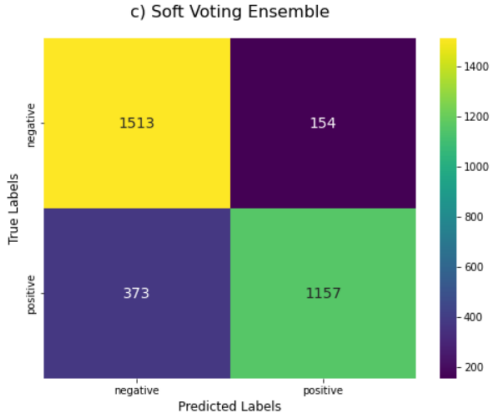
Figure 4. Comparison of precision, recall, and F1-score for the ensemble model across different upper-limb regions



a) Performance of the EfficientNet-B4 model



b) Performance of the DenseNet-121 model



c) Performance of the combined Soft Voting Ensemble method

Figure 5. Overall confusion matrices for comparing the performance of methods on the test dataset

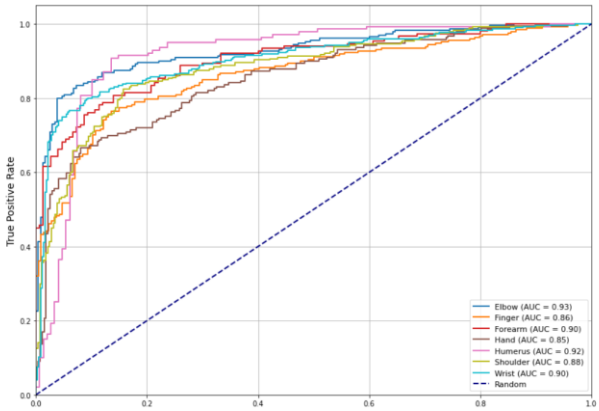


Figure 6. Comparison of ROC curves and AUC values for the Ensemble model across different parts of the upper limb.

Superiority of the ensemble model: By combining predictions from the two models, DenseNet-121 and EfficientNet-B4, the ensemble improved its overall performance, raising the AUC to 0.89. This improvement highlights the strong potential of model combination in building a more robust and stable system. This increase in AUC for the ensemble model demonstrates enhanced discriminative power and greater confidence in the model's diagnostic decision-making. To further assess the effectiveness of the proposed approach, its results were compared with several influential prior studies on fracture detection in radiographic images, particularly those based on the MURA dataset. The DenseNet-121 model was employed for binary classification of abnormalities, and its performance was reported at a radiologist-level standard.

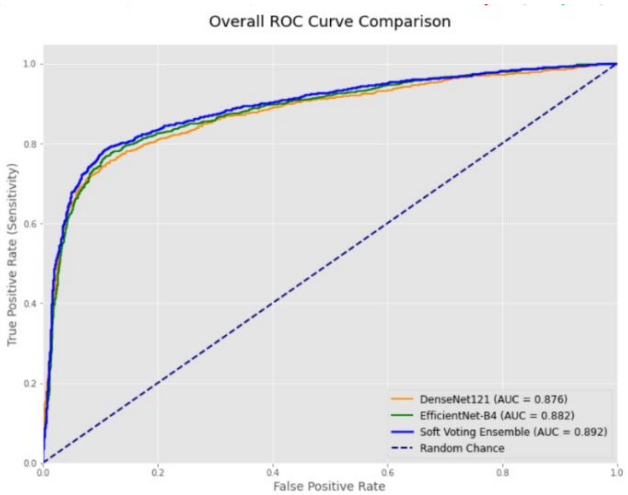


Figure 7. ROC curves for comparing the performance of methods on the test dataset

TABLE 7. Comparison of the proposed method with previous studies on MURA. (*Starred papers indicate that only part or 50% of the dataset was used.*)

Reference	Year	Used Model	Accuracy	AUC	Dataset
(15)*	2019	DenseNet-201, InceptionResNetV2	82	0.88	Humerus finger
(19)	2021	ST-SSAD	-	0.78	All 7 study types
(20)	2024	VGG-16 + ViT	82.88	-	All 7 study types
(4)	2022	Ensemble AlexNet + ResNet18	82.69	0.90	All 7 study types
Proposed Method	-	EfficientNet-B4 + DenseNet-121 (Soft Voting)	**84.1**	**0.90**	Humerus finger
Proposed Method	-	EfficientNet-B4 + DenseNet-121 (Soft Voting)	**83.52**	**0.89**	All 7 study types

However, this approach could not localize the fracture region and was limited only to overall decision-making. Subsequently, studies utilizing more advanced architectures such as MSDNet and DenseNet-169 attempted to enhance diagnostic performance.

While these models demonstrated high accuracy in certain metrics, they also faced challenges such as high computational requirements, complexity in implementation, and limited generalizability (4, 15).

To enable scope-matched comparison, we report results on both the full seven-region setting (Overall Accuracy 83.52%, Sensitivity 90.76%, Specificity 78.66%) and the restricted "finger + humerus" subset used by prior work (15). In the latter, our ensemble attains 84.1% Accuracy and 0.90 AUC, exceeding the 80% Accuracy and 0.84 AUC reported in (15). These results indicate that differences in anatomical coverage and granularity (image- vs study-level) can inflate metrics in narrower settings, and that our method remains competitive even under matched conditions.

Moreover, to enable a fair comparison with Chada (15), who reported results only on the finger and humerus subsets of the MURA dataset, we also evaluated our proposed method under the same restricted setting. In this case, the ensemble model achieved an average accuracy of 84.1% and an AUC of 0.90, outperforming the accuracy (80%) and AUC (0.84) reported by Chada (15). This demonstrates that even when constrained to a limited portion of the dataset, our approach maintains superior discriminative ability and robustness compared to previous methods. In contrast, the proposed model in this study, by combining the two advanced architectures DenseNet-121 and EfficientNet-B4 within an ensemble framework using a soft-voting mechanism, achieved competitive performance and, in some cases, outperformed previous works. To assess the position of the proposed method relative to prior advanced approaches, Table 6 compares its overall performance with several selected studies conducted on the MURA dataset. These results show that, in terms of accuracy,

balanced performance across different regions, and ease of implementation, the proposed method offers a significant advantage over many earlier techniques.

Ultimately, the presented model, with its balanced approach, relative simplicity, and generalizability in real clinical settings, holds strong potential for integration into computer-aided diagnostic systems. Our proposed framework has demonstrated stable and reliable performance by maintaining a balance among different evaluation metrics. In particular, by achieving an overall accuracy of 83.52% and a meaningful AUC of 0.89, the proposed model shows considerable promise as an auxiliary diagnostic tool in real clinical environments.

Table 7 presents a comparative analysis of the proposed ensemble method against several state-of-the-art approaches previously applied to the MURA dataset. As shown, earlier studies such as Chada (15) and others (4, 19, 20) employed models including DenseNet-201, InceptionResNetV2, ST-SSAD, VGG-16 combined with ViT, and ensemble variants of AlexNet and ResNet18, reporting accuracies in the range of 77–83%. In contrast, our proposed method, based on the integration of EfficientNet-B4 and DenseNet-121 through a soft voting strategy, achieved an accuracy of 83.52% and an AUC of 0.89 across all seven study types, thereby outperforming most of the prior works. Moreover, to enable a fair comparison with Chada (15), who reported results only on the finger and humerus subsets of the MURA dataset, we also evaluated our method under the same restricted setting. In this case, the ensemble model achieved an average accuracy of 84.1% and an AUC of 0.90, surpassing the accuracy (80%) and AUC (0.84) reported by Chada (15).

These findings demonstrate that even under limited data conditions, the proposed approach maintains superior discriminative ability and robustness compared to previous methods. This table summarizes the models used, evaluation metrics, and reported results from prior work. Starred papers indicate that only a portion of the dataset, or 50% of it, was utilized. As shown, previous approaches have employed various architectures, including DenseNet, InceptionResNetV2, ST-SSAD, VGG-16 combined with ViT, and ensemble models such as AlexNet and ResNet18. The proposed method integrates EfficientNet-B4 and DenseNet-121 through a soft-voting mechanism and demonstrates competitive performance, achieving higher accuracy compared to most previous studies while maintaining a strong AUC score.

To qualitatively assess the performance of the proposed model, several random predictions on the MURA dataset are presented. Figure 8 shows the ensemble model outputs across different skeletal regions, including the elbow, forearm, wrist, finger, hand, and shoulder. As can be observed, the model successfully identified many cases correctly (e.g., detecting fractures in the shoulder and finger regions), while some misclassifications still occurred (e.g., false negatives in the forearm region). These examples highlight the inherent challenges in automated fracture detection, where subtle abnormalities and structural similarities to normal anatomy may lead to errors. Overall, the qualitative results in Figure 8 complement the quantitative evaluation and confirm that the proposed model demonstrates robust predictive ability, making it a valuable clinical decision support tool in radiological workflows.

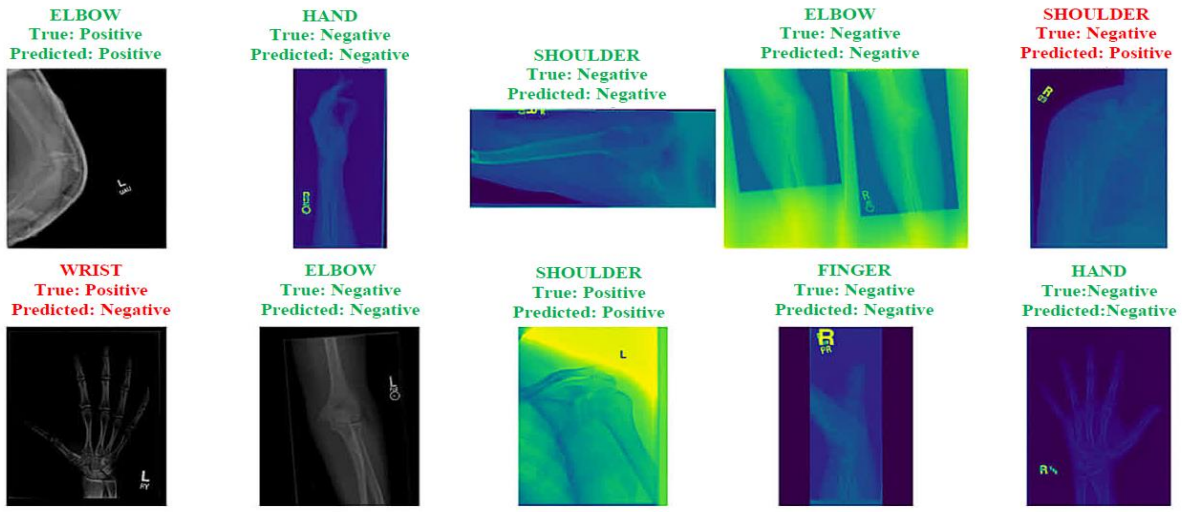


Figure 8. Random sample predictions across different skeletal regions using the proposed ensemble model. The figure presents true positive, true negative, false positive, and false negative predictions on radiographic images of various anatomical regions (elbow, forearm, wrist, finger, hand, and shoulder). The results highlight both the model's reliable performance and the challenges of detecting subtle bone abnormalities in medical imaging

4. 4. 5. Computational Considerations For clinical integration, computational efficiency is as important as accuracy. Our ensemble comprises two pretrained backbones plus customized dense heads. The approximate parameter counts are: EfficientNet-B4 ≈ 19M, DenseNet-121 ≈ 8M, with the added dense layers contributing ≈ 7.7M (2560→1200→2 on the EfficientNet branch) and ≈ 1.6M (1024→512→2 on the DenseNet branch). In total, the ensemble has ≈ 36–37M trainable parameters. At 32-bit precision, the parameter memory footprint is on the order of ≈ 140 MB, which roughly halves under FP16.

We measured end-to-end inference in a Kaggle GPU environment equipped with an NVIDIA Tesla T4 (batch size = 8, input size 380×380, FP32). Including preprocessing and both forward passes, the average latency was ≈ 4.0 s per image. This value serves as our deployment baseline. In settings where lower latency is required (e.g., interactive triage), several standard accelerations are available without retraining the ensemble: mixed-precision (FP16) inference, export to ONNX/TensorRT, and layer-level pruning of the added dense heads. For resource-constrained deployments, a knowledge-distilled single-backbone student network is a practical alternative that preserves most of the ensemble's accuracy while reducing latency and memory. Finally, a simple confidence-gated protocol can route easy cases through a single backbone and reserve the full ensemble for ambiguous cases, trading a small accuracy drop for significant average-latency reductions.

We report latency as time per image at batch size 8 because this setting best matches point-of-care use. Throughput under mini-batching scales predictably on GPUs and can be reported for site-specific hardware if required.

5. CONCLUSION

In this study, an innovative deep learning–based framework was proposed for the detection and classification of bone abnormalities in radiographic images. To this end, two advanced models, EfficientNet-B4 and DenseNet-121, were redesigned and optimized by adding customized dense layers at the output, enabling the extraction of precise and specialized features. To further improve classification accuracy and stability, the predictions of these two models were integrated into an ensemble structure through a soft-voting mechanism. The proposed model was evaluated on the complex and widely recognized MURA dataset, which includes seven distinct anatomical regions. Experimental results demonstrated that combining these two architectures produced reliable, stable, and accurate performance, with the ensemble model achieving an overall accuracy of 83.52%, thus offering competitive results compared to existing methods.

The proposed ensemble model demonstrated robust and consistent performance across all seven anatomical regions in the MURA dataset, highlighting its ability to generalize across diverse skeletal structures with minimal bias toward specific bone types.

The proposed framework contributes to the field by demonstrating that an ensemble of heterogeneous CNN architectures, enhanced with customized dense layers, can effectively capture subtle fracture patterns and improve both accuracy and sensitivity. This approach offers a robust and practical foundation for clinical decision-support systems in medical imaging.

Taken together, the all-regions, image-level evaluation explains why κ may be lower than reports limited to specific regions, while our Sensitivity (90.76%), AUC (0.89), and scope-matched improvements over support the clinical utility and robustness of the proposed ensemble.

The proposed models were implemented in Python 3.7.12 using TensorFlow 2.x with the built-in tf.keras API, and were trained and evaluated in a Kaggle GPU environment equipped with an NVIDIA Tesla T4 accelerator, ensuring reproducible results.

6. FUTURE WORK

Despite the promising results of the proposed method, there remain several opportunities for further improvement and extension in future research. While the current study focuses on achieving high classification accuracy, it does not explicitly address model interpretability. Incorporating methods such as CAM or Grad-CAM in future work can help highlight the most influential regions in the images, providing clinicians with more trustworthy and interpretable decisions. One potential direction is the use of CAM or Grad-CAM to generate heatmaps that highlight the region's most influential in the model's decisions, thereby improving the clinical interpretability and trustworthiness of the system. Another avenue is the incorporation of Vision Transformers (ViT) or hybrid CNN-Transformer architectures, which could enhance the model's ability to capture global contextual relationships within images and improve accuracy in detecting more complex fractures. Additionally, the integration of attention mechanisms would enable the model to selectively focus on the most relevant areas of an image while ignoring non-essential information, ultimately leading to greater precision, robustness, and reliability in diagnostic performance. Furthermore, employing k-fold cross-validation in future studies could provide a more robust and reliable estimate of model performance and generalization across different subsets of the dataset.

Acknowledgements

The authors would like to thank the Faculty of Engineering and Technology at the University of Mazandaran for providing the computational resources necessary for this study.

Funding

This research received no external funding or financial support.

Ethics Approval and Consent to Participate

This study used the publicly available MURA dataset and did not involve experiments on human participants or animals conducted by the authors; therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The authors declare that they have no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT solely for minor language editing and clarity improvement. All scientific content, analysis, and conclusions were produced by the authors, who take full responsibility for the integrity, accuracy, and originality of the work.

Data Availability

The data used in this study are publicly available within the MURA dataset. Any processed results or model outputs can be shared by the corresponding author upon reasonable academic request.

Authors Biosketch

Mohadeseh Mahtabi is a Master's student in Computer Engineering – Artificial Intelligence at the University of Mazandaran, Iran. Her research interests include deep learning, ensemble learning, and medical image analysis, with a focus on automated bone fracture detection.

Sekineh Asadi Amiri is an Associate Professor in the Department of Computer Engineering at the University of Mazandaran, Iran. She received her Ph.D. in Computer Engineering from Shahrood University of Technology in 2016. Her research interests include image processing, deep learning, and intelligent computer-aided diagnostic systems, with a particular focus on medical image analysis and deep neural network–based models.

Samira Mavaddati is an Associate Professor in the Department of Electrical Engineering at the University of Mazandaran, Iran. She received her Ph.D. from Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2016. Her research interests include image processing, signal processing, deep learning, pattern recognition, and intelligent data analysis, with applications in engineering and biomedical systems.

REFERENCES

1. Lu S, Wang S, Wang G. A novel deep learning approach for image segmentation. Multimedia Tools and Applications. 2022;81:44487-503. https://doi.org/10.1007/s11042-022-13287-z
2. Saggi SS, Kuah LZD, Toh LCA, Shah MTM, Wong MK, Abd Razak HRB. A novel approach to structural health monitoring using deep learning. Journal of Civil Engineering. 2023;50(2):115-32.
3. Ioppolo F, Rompe JD, Furia JP, Cacchio A. Clinical application of shock wave therapy (SWT) in musculoskeletal disorders (MSDs). European Journal of Physical and Rehabilitation Medicine. 2014;50(2):217-30.
4. Karthik K, Kamath SS. MSDNet: A deep neural ensemble model for abnormality detection and classification of plain radiographs. Journal of Ambient Intelligence and Humanized Computing. 2023;14(12):16099-113. https://doi.org/10.1007/s12652-022-03835-8
5. Karthik K, Kamath SS. A deep neural network model for content-based medical image retrieval with multi-view classification. The Visual Computer. 2021;37(7):1837-50. https://doi.org/10.1007/s00371-020-01941-2
6. Ying J, Dutta J, Guo N, Hu C, Zhou D, Sitek A, Li Q. Classification of exacerbation frequency in the COPDGene cohort using deep learning with deep belief networks. IEEE Journal of Biomedical and Health Informatics. 2016;24(6):1805-13. https://doi.org/10.1109/JBHI.2016.2642944
7. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. Journal of the American College of Radiology. 2010;7(9):698-704. https://doi.org/10.1016/j.jacr.2010.03.004
8. Mukesh BR, Harish T, Mayya V, Kamath S. Deep learning based detection of diabetic retinopathy from inexpensive fundus imaging techniques. In: 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT); 2021. 1-6. https://doi.org/10.1109/CONECCT52877.2021.9622703
9. García-Floriano A, Ferreira-Santiago Á, Camacho-Nieto O, Yáñez-Márquez C. A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images. Computers and Electrical Engineering. 2019;75:218-29. https://doi.org/10.1016/j.compeleceng.2017.11.008
10. Nedumkunnel IM, George LE, Sowmya KS, Rosh NA, Mayya V. Explainable deep neural models for COVID-19 prediction from chest X-rays with region of interest visualization. In: 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC); 2021. 96-101. https://doi.org/10.1109/ICSCCC51823.2021.9478152

11. Mayya V, Karthik K, Sowmya KS, Karadka K, Jeganathan J. Coviddx: AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data. In: International Conference on Health Informatics (HEALTHINF); 2021. p. 659-66.
12. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, Yang B, Zhu K, Laird D, Ball RL, Langlotz C. MURA: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv. 2017. https://doi.org/10.48550/arXiv.1712.06957
13. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012;25:1097-105.
14. Ebsim R, Naqvi J, Cootes TF. Automatic detection of wrist fractures from posteroanterior and lateral radiographs: A deep learning-based approach. In: International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging; 2018. p. 114-25. https://doi.org/10.1007/978-3-030-11166-3_10
15. Chada G. Machine learning models for abnormality detection in musculoskeletal radiographs. Reports. 2019;2(4):26. https://doi.org/10.3390/reports2040026
16. Yahalomi E, Chernofsky M, Werman M. Detection of distal radius fractures trained by a small set of X-ray images and Faster R-CNN. In: Intelligent Computing – Proceedings of the Computing Conference; 2019. p. 971-81. https://doi.org/10.1007/978-3-030-22871-2_69
17. Yadav DP, Rathor S. Bone fracture detection and classification using deep learning approach. In: International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC); 2020. 282-85. https://doi.org/10.1109/PARC49193.2020.236611
18. Meena T, Roy S. Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. Diagnostics. 2022;12(10):2420. https://doi.org/10.3390/diagnostics12102420
19. Spahr A, Bozorgtabar B, Thiran JP. Self-taught semi-supervised anomaly detection on upper limb X-rays. In: IEEE International Symposium on Biomedical Imaging; 2021. 1632-36. https://doi.org/10.1109/ISBI48211.2021.9433771
20. Dahal S, Thapa R, Panth S. A hybrid deep learning model for musculoskeletal abnormality detection. In: Proceedings of the 15th IOE Graduate Conference; 2024. 1-6.
21. Hardalaç F, Uysal F, Peker O, Çiçeklidağ M, Tolunay T, Tokgöz N, Kutbay U, Demirciler B, Mert F. Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors. 2022;22(3):1285. https://doi.org/10.3390/s22031285
22. Beyaz S, Yayli SB, Kılıc E, Doktur U. The ensemble artificial intelligence method: Detection of hip fractures in AP pelvis plain radiographs by majority voting using a multi-center dataset. Digital Health. 2023;9:1-14. https://doi.org/10.1177/20552076231216549
23. Ghosh M, Hassan S, Debnath P. Ensemble based neural network for the classification of MURA dataset. Journal of Nature, Science and Technology. 2021;4:1-5.
24. Lu X, Chang EY, Du J, Yan A, McAuley J, Gentili A, Hsu CN. Robust multi-view fracture detection in the presence of other abnormalities using HAMIL-Net. Military Medicine. 2023;188(Supplement 6):590-7. https://doi.org/10.1093/milmed/usad252
25. Stanford ML Group. MURA dataset. Available from: https://stanfordmlgroup.github.io/competitions/mura/
26. Zhang P, Yang L, Li D. EfficientNet-B4-Ranger: A novel method for greenhouse cucumber disease recognition under natural complex environment. Computers and Electronics in Agriculture. 2020;176:105652. https://doi.org/10.1016/j.compag.2020.105652
27. Albelwi SA. Deep architecture based on DenseNet-121 model for weather image recognition. International Journal of Advanced Computer Science and Applications. 2022;13(10):559.
28. Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in Neural Information Processing Systems. 2018;31.
29. Agnihotri D, Verma K, Tripathi P, Singh BK. Soft voting technique to improve the performance of global filter based feature selection in text corpus. Applied Intelligence. 2019;49(4):1597-619. https://doi.org/10.1007/s10489-018-1349-1
30. Mim MA, Majadi N, Mazumder P. A soft voting ensemble learning approach for credit card fraud detection. Heliyon. 2024;10:e25466. https://doi.org/10.1016/j.heliyon.2024.e25466
31. Asadi Amiri S, Mohammadpoory Z, Nasrolahzadeh M. A novel content-based image retrieval system using fusing color and texture features. Journal of AI and Data Mining. 2022;10(4):559-68. https://doi.org/10.22044/jadm.2022.12042.2353
32. McHugh ML. Interrater reliability: The kappa statistic. Biochemia Medica. 2012;22(3):276-82.
33. Mavaddati S, Razavi M. An optimized YOLO-ViT hybrid model for enhanced precision in rice classification and quality assessment. International Journal of Engineering, Transactions A: Basics. 2025;38(10):2435-50. https://doi.org/10.5829/ije.2025.38.10a.19
34. Mohammadi M, Talebpour A, Hosseinsabet A. Presenting effective methods in classification of echocardiographic views using deep learning. International Journal of Engineering, Transactions B: Applications. 2024;37(11):2150-61. https://doi.org/10.5829/ije.2024.37.11b.02
35. Farsi H, Noursoleimani S, Mohamadzadeh S, Barati A. Multimodal biomedical image segmentation by using multi-path U-Net. International Journal of Engineering. 2025;38(1):179-93. https://doi.org/10.5829/ije.2025.38.01a.17

Persian Abstract

چکیده

پیشرفت‌های اخیر در شبکه‌های عصبی عمیق موجب بهبود چشمگیری در تحلیل تصاویر پزشکی شده است، با این حال **تشخیص شکستگی استخوان در تصاویر رادیوگرافی** همچنان یک چالش اساسی محسوب می‌شود؛ زیرا ساختار پیچیده استخوان، الگوهای ظریف شکستگی و محدودیت داده‌های برچسب‌خورده، عملکرد مدل‌ها را با دشواری روبه‌رو می‌کند. در این پژوهش، یک چارچوب پیشرفته یادگیری عمیق تلفیقی ارائه می‌شود که دو معماری بهینه‌شده EfficientNet-B4 و DenseNet-121 را از طریق راهبرد ترکیب نرم (Soft-Voting Fusion) برای تشخیص خودکار و دقیق شکستگی‌ها ادغام می‌کند. در این ساختار هیبریدی، از وزن‌دهی تطبیقی و لایه‌های متراکم بهینه‌شده استفاده شده است تا تمایز ویژگی‌ها تقویت شود و توان شبکه در تشخیص جزئیات ظریف شکستگی‌ها افزایش یابد. علاوه بر این، با به‌کارگیری یادگیری انتقالی و تنظیم دقیق (Fine-tuning)، عدم‌توازن داده‌ها کاهش یافته و تعمیم‌پذیری مدل در نواحی مختلف آناتومیکی بهبود یافته است. نتایج حاصل از آزمایش‌های جامع بر روی مجموعه‌داده MURA که شامل تصاویر رادیوگرافی از هفت ناحیه مختلف بدن است، نشان می‌دهد که مدل پیشنهادی با دستیابی به ۸۳/۵۲٪ دقت و ۹۰/۷۶٪ حساسیت عملکردی برتر نسبت به مدل‌های پایه ارائه می‌دهد. پایداری نتایج در تنظیمات مختلف آموزشی، قابلیت اطمینان مدل را برای کاربردهای بالینی تأیید می‌کند. در مجموع، این پژوهش یک **سامانه نوآورانه مبتنی بر یادگیری عمیق تلفیقی** معرفی می‌کند که با بهره‌گیری از ترکیب تطبیقی ویژگی‌ها و نقاط قوت معماری‌های مکمل، دقت بالایی در تشخیص شکستگی‌ها فراهم می‌آورد و گامی مؤثر در جهت توسعه **سامانه‌های هوشمند پشتیبان تصمیم‌گیری رادیولوژی** به‌شمار می‌رود.