

Hybrid Transformer-Based Bone Fracture Detection using EfficientNet-B3 and Self-Attention Mechanism

Syed Muddusir Hussain¹, Tayaba Naz¹, and Faraz Akram¹

¹Department of Biomedical Engineering, Riphah International University, Islamabad,

Pakistan Correspondence Author: Syed Muddusir Hussain

(muddasir.hussain@riphah.edu.pk)

Abstract

Accurate and timely detection of bone fractures from radiographic images remains a significant challenge in orthopedic diagnostics, particularly in emergency clinical environments where subtle fractures and image noise may lead to diagnostic errors. Recent advances in deep learning have demonstrated promising results in automated fracture detection; however, many existing Convolutional Neural Network (CNN)-based models suffer from limited interpretability and insufficient robustness under varying imaging conditions. To address these limitations, this study proposes a Hybrid Transformer-Based framework for automated bone fracture detection using the EfficientNet-B3 architecture integrated with a customized self-attention mechanism. The proposed framework combines the feature extraction capability of EfficientNet-B3 with transformer-inspired spatial attention to capture both local structural patterns and global contextual relationships in musculoskeletal radiographs. Additionally, Gaussian noise augmentation is incorporated to improve model robustness against clinically relevant imaging disturbances. The proposed model was evaluated using 5,396 radiographic images obtained from the MURA musculoskeletal dataset under three different train-test splits (60–40, 70–30, and 80–20) using cross-validation. Experimental results demonstrated strong and consistent diagnostic performance, achieving classification accuracies ranging from 87.43% to 89.38%, with a maximum ROC-AUC score of 0.9121 and precision of 0.9133. Furthermore, the self-attention mechanism improved fracture localization and enhanced interpretability by enabling the model to focus on clinically significant regions within radiographic images. The findings indicate that the proposed hybrid framework provides an efficient, scalable, and clinically reliable solution for AI-assisted orthopedic diagnosis and may serve as an effective decision-support tool for radiologists in real-world medical applications.

Index Terms: Attention Mechanism, Bone Fracture Detection, Deep Learning, EfficientNet-B3, Musculoskeletal Radiographs.

I. INTRODUCTION

The medical imaging techniques have revolutionized the practice of diagnostic radiology using rapid and accurate visualization of internal body structures. There exist several medical imaging modalities, one of which is commonly employed in bone injury diagnosis is musculoskeletal radiography. Despite the widespread use of such equipment, there are various challenges associated with the interpretation of these images due to the complexity of the fractures, poor image contrast, hairline fracture diagnosis difficulty, and other factors [1], and [2]. The increase in global prevalence rates of orthopedic injuries combined with the lack of qualified radiologists requires automatic diagnosis systems [3].

Several milestones have recently been reached within the domains of artificial intelligence and deep learning with respect to improving the techniques used for the analysis of medical images. Among them, the utilization of deep neural networks can be seen as particularly efficient for the identification of any abnormalities in radiographic images. Although simple deep convolutional network models like AlexNet and VGG-16 have proved themselves capable of performing the task of classification of a fracture through the application of deep learning, the more sophisticated

models of ResNet, DenseNet, and Inception, with their capability to represent things more deeply, are currently doing the job better [4], and [5].

But still, there are certain drawbacks of using CNN models for processing and analysis of medical images. For example, CNN models operate mainly with feature extraction on a local level and do not take into account long-range relationships within features of radiography images. Also, DL models serve as black boxes that cannot provide interpretability of their output which complicates the process of validating the findings made by radiologists [6]. There may be a possibility that diagnostic accuracy will be compromised due to such factors as imaging noise, image exposure variability, and low image quality of radiographs typical for an emergency setting [7], and [8]. In recent times, some solutions have been suggested based on the integration of transformers into CNN networks [9], and [10].

The transformer model has exhibited high efficiency in modeling global contextual dependencies via self-attention techniques. Nevertheless, the immediate application of Vision Transformers (ViT) to medical image analysis involves substantial computational power and large training data sets [9]. Hybrid CNN-transformer models are now regarded as an effective option, which leverages both

local feature representation and global reasoning ability of CNN and transformers, respectively [10]. Additionally, modern approaches have recognized the significance of transfer learning, compact network design, and attention-based methods in enhancing the detection efficiency and clinical viability of bone fractures [11-13].

In an attempt to address such issues, this research presents the development of a novel Hybrid Transformer-based architecture for reliable bone fracture diagnosis via the implementation of EfficientNet-B3 alongside a tailor-made self-attention system. This new approach utilizes the efficiency and compound scaling nature of EfficientNet-B3, coupled with transformer-based attention, to detect local anatomical features, as well as global relationships within the musculoskeletal radiographs. Moreover, Gaussian noise augmentation has been applied for model training as part of data preprocessing.

The main contributions of this study are summarized as follows:

- Development of a Hybrid Transformer-based fracture detection framework integrating EfficientNet-B3 with a customized self-attention mechanism for enhanced feature representation and fracture localization.
- Introduction of Gaussian noise augmentation to improve robustness and generalization under realistic clinical imaging conditions.
- Improvement of model interpretability through attention-guided spatial weighting mechanisms capable of highlighting clinically significant fracture regions.
- Comprehensive experimental validation using the Stanford Musculoskeletal Radiographs (MURA) or musculoskeletal radiography dataset under multiple train-test splits and cross-validation strategies to evaluate diagnostic reliability and scalability.

The remainder of this paper is organized as follows. Section II presents the literature review and discusses recent developments in CNNs, transformer architectures, and fracture detection systems. Section III describes the problem statement and the proposed hybrid transformer framework. Section IV explains the methodology, including dataset preprocessing, noise augmentation, EfficientNet-B3 feature extraction, self-attention mechanism, and experimental setup. Section V presents the experimental results and performance evaluation. Section VI discusses the clinical significance and comparative analysis of the proposed framework. Finally, Section VII concludes the study and Section VIII outlines future research directions.

II. LITERATURE REVIEW

The categorization of musculoskeletal injury has experienced a major change in perspective from the conventional feature extraction technique to a more advanced deep learning-based technique that facilitates

hierarchical representation learning. Early research efforts towards the development of automated models made use of image processing techniques and shallow CNNs to detect abnormalities in long bones. Numerous studies have demonstrated that deep learning algorithms exhibit superior performance compared to traditional edge detection and feature extraction methods due to hierarchical learning of spatial relationships among pixels [14]. However, classical CNN architectures face many challenges due to the inherent locality of receptive field sizes.

In order to address these drawbacks, more advanced CNN architectures such as ResNet are commonly utilized in medical image analysis. For example, ResNet-50 has demonstrated superior performance when dealing with neuromedicine applications and musculoskeletal problems because of its residual learning approach, which mitigates the problem of vanishing gradients and allows training deeper architectures [15]. Similarly, the CNN architecture can be successfully implemented to identify bone fractures and classify them despite the region where the fracture is located, thus confirming the applicability of these algorithms in a clinical setting [16]. In addition, techniques for segmentation, such as k-means and FCM algorithms, are employed to accurately pinpoint abnormalities within a complex background [17].

In spite of all these progressions, CNN-based models can still be highly susceptible to image artefacts and noise that typically occur within emergency radiographic imaging scenarios. In order to achieve better performance, there has been an emphasis on incorporating noise injection and augmentation methods, for instance, using Gaussian noise layers, that make the model more robust to noisy input data [7], and [8]. Hyperparameter optimization such as learning rate and batch size as well as the use of dropout have proven important in this regard [18].

Transformers are one of the recent approaches used for medical image analysis that utilize the ability to capture global context information using self-attention. Vision transformers have shown remarkable results in performing large-scale image recognition tasks. However, they cannot be directly used for medical images owing to the limitations of computational demands and reliance on large amounts of data [9]. Hybrid CNN-transformer networks have been developed to leverage the strengths of both the techniques by combining the local feature extraction approach of CNNs with the transformer architecture's global reasoning capabilities [10]. Herein, hierarchical spatial features are extracted from the input medical images using CNN backbones like EfficientNet, whereas the attention mechanism assigns varying importance to different parts of the image [4], and [10].

Musculoskeletal imaging has seen recent developments involving the refinement of generalization, efficiency, and clinical application through transfer learning, lightweight modeling, and foundation models. Techniques in transfer learning using pre-trained Convolutional Neural Networks (CNNs) have exhibited an enhanced rate of convergence and increased accuracy, as well as minimized computational cost [11]. In addition, there have been studies conducted that use lightweight deep learning algorithms for real-time fracture detection in clinical and

sports medicine applications [12]. Furthermore, state-of-the-art multi-region detection techniques have been developed, providing enhanced accuracy in detecting fractures in various anatomical regions from a single radiographic image [19]. Robustness is another area where significant work has been done, which involves developing noise-resistant training methods and adversarial defense mechanisms [8], and [20]. Attention mechanisms provide another level of interpretability to the model by allowing it to selectively concentrate on specific areas of interest within the image [13].

Table I: Comparative Analysis of State-Of-The-Art Methodologies in Medical Image Classification and Fracture Detection.

S. No	Methodology	Primary Focus	Key Performance/Finding
1	EfficientNet Scaling	Model Efficiency	Optimized balance between network depth, width, and resolution.
2	DenseNet/Ensemble	Musculoskeletal Data	High AUC on the MURA dataset for abnormality detection.
3	Data Augmentation	Radiographic Robustness	Proved that noise injection reduces overfitting in clinical X-rays.
4	Hybrid CNN-Attention	Diagnostic Support	Significant improvement in localization accuracy for subtle cracks.
5	Multi-task Learning	Fracture & Dislocation	Concurrent detection of multiple orthopedic pathologies.
6	Inception-v4	Pediatric Radiology	Successful application in identifying growth plate fractures.
7	U-Net Architecture	Fracture Segmentation	Precise pixel-level identification of bone structural breaks.

Despite these developments, however, one critical issue still exists regarding the balance between model complexity, interpretation, and robustness in fracture detection systems. Despite offering superior performance in terms of feature extraction, deep CNN models like ResNet-50 and DenseNet work on the "black box" principle without being clinically understandable [15], and

[6]. On the other hand, even though transformer models perform better at global reasoning, they come with increased computational costs, necessitating careful architecture design for clinical applications [9]. Hence, the need arises for models that can efficiently integrate feature extraction with interpretable attention mechanisms while ensuring robustness amid noise in clinical practice. Such observations set the stage for this study, which presents a Hybrid Transformer-based fracture detection framework that combines EfficientNet-B3 with self-attention mechanisms [21-24].

III. PROBLEM STATEMENT AND ITS PROPOSED SOLUTION

A. Problem Description

Automated bone fracture detection has always been a difficult problem because of the varying quality of radiographs and intricate bone anatomy. In cases of emergencies, fine fractures such as hairline fractures are likely to camouflage within dense bones or be masked by artifacts, resulting in possible errors in diagnosis. Deep learning models have been used, but they tend to work as "black box" solutions that fail to provide adequate explainability and generalization for clinical verification. Radiologists require clear indications of local pathological areas for verification. Thus, a highly effective neural network architecture should be developed that employs attention algorithms.

B. Solution Framework

The proposed hybrid transformer-based model, shown in Figure 1, is able to solve the described issues based on the Hybrid Transformer Model, which uses the benefits of high-resolution features extraction together with dynamic spatial weights. In our case, the essence of the research is the use of the pre-trained architecture of EfficientNet-B3 model for features extraction and the use of the self-attention layer to detect areas of discontinuity. In order to mimic the stochasticity of input data in the real-world scenario, noise is introduced to the input images. This architecture and experiment will be considered in more detail in the next subsections.

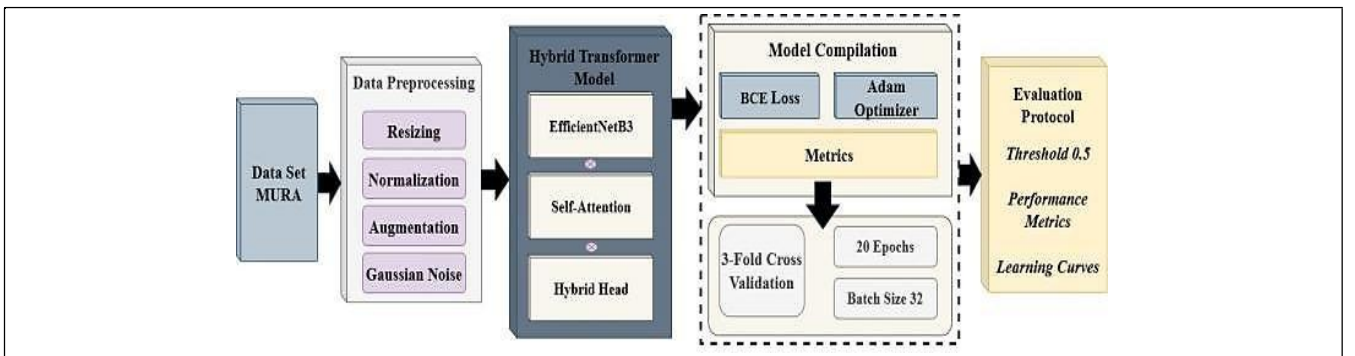


Figure 1: Basic Idea of the Research

IV. METHODOLOGY

A. Dataset Description

The basis for the experimental study of deep learning techniques is provided by the publicly accessible dataset

Stanford Musculoskeletal Radiographs (MURA), which is one of the biggest databases available for studying abnormal findings in musculoskeletal imaging. It consists of totally 5,396 radiographic images with equal proportions of positive (fracture/abnormal) and negative (normal) classes. In the current work, an equally-sized

subset was chosen for learning purposes, that is 3,160 positive samples and 3,160 negative samples were used. Stanford MURA database includes multiple views from different regions of the upper extremities like elbow, wrist, shoulder, hand. This dataset is frequently used in medical imaging studies owing to its clinical relevance and high variability in terms of exposure and orientation among others. This database is ideal for testing deep learning models in fracture detection due to its variability characteristics.

B. Data Augmentation and Preprocessing

In order to increase the model's generalizability and prevent overfitting, an effective approach for data augmentation was performed on the dataset used in training. As the conditions of radiological imaging differ, augmentations were chosen which can imitate real-life settings and improve robustness in regard to image distortion.

For this purpose, both photometric and geometric transformations were utilized, such as random rotation up to $\pm 15^\circ$, horizontal flip, and modification of the brightness level. In doing so, the model learns to recognize feature-invariant properties regardless of orientation and lighting conditions.

Apart from conventional methods, random Gaussian noise was added to the original images to simulate real-life conditions when imaging is done under the presence of acquisition noise. This method helps the model become more generalizable when exposed to noisy input data.

Mathematically, this process may be represented as:

$$\chi_{\text{noise}} = \chi + N(0, \sigma^2) \quad (1)$$

Where:

χ is the original input image, while $N(0, \sigma^2)$ refers to Gaussian noise with a zero mean and variance σ^2 .

All images were normalized before training, such that their sizes match those required by the EfficientNet-B3 network architecture, thus guaranteeing that all inputs have equal sizes for effective feature extraction. Moreover, pixel intensity normalization was performed such that the pixel values lie within the same range.

C. Noise Modeling and Preprocessing

One of the unique aspects of our approach is that we incorporate a Gaussian Noise Layer right after the input

layer. Stochastic noise may occur in clinical x-ray images due to electronic interference or low doses of radiations. The inclusion of the Gaussian noise allows the system to be trained on differentiating the noise from the structure of the image, making it robust against any form of visual interference.

D. EfficientNet-B3 Backbone as Feature Extractor

In the proposed scheme, the feature extraction process is implemented by using EfficientNet-B3 architecture, which functions as the main convolutional feature extractor. In contrast to the conventional method of separately scaling up the depth, width, or resolution of CNNs, EfficientNet

uses the compound scaling technique, where all three dimensions of the network are equally scaled in order to ensure a balance among them.

The compound scaling theory can be represented mathematically as:

$$d = \alpha \phi, w = \beta \phi, r = \gamma \phi \quad (2)$$

Where:

d , w , and r refer to the depth, width, and input resolution of the model, respectively, whereas ϕ denotes a compound coefficient for scaling the network uniformly, and α , β , and γ denote constants obtained by performing grid search.

As a result, EfficientNet-B3 has been shown to efficiently extract fine-grained hierarchical structures from musculoskeletal radiographs without incurring too much complexity in terms of computation that is common in deeper networks like ResNet-152. Therefore, EfficientNet-B3 is ideal for deploying the model across multiple clinical settings in practice.

The output of EfficientNet-B3 architecture can be described as high-level representations in the form of multi-dimensional feature maps, and the features learned are a combination of both edge-like information and anatomical structures, and the next step involves feeding them into the transformer model.

E. Self-Attention and Spatial Weighting Mechanism

In order to counteract the issue of poor interpretability usually inherent to the traditional CNNs, the proposed hybrid model adopted a novel self-attention architecture. While CNN is known for being good at extracting spatial features locally, the model is considered a "black-box" that lacks an explanation in determining the most relevant anatomical parts contributing to the output predictions. The proposed self-attention model not only provides more efficient modeling of context dependencies but also improves interpretability by spatially weighting relevant clinical parts.

Self-attention works by computing relationships among the feature representations using Q , K , and V matrices created based on the feature maps obtained.

The attention process can be expressed mathematically as:

$$\text{Attention}(q, K, V) = \text{softmax} \left(\frac{qk^T}{\sqrt{d_k}} \right) \quad (3)$$

Where:

Q , K , and V refer to the query, key, and value matrices, respectively, while d_k refers to the key vector dimension for the normalization process. The softmax operation is responsible for producing normalized attention coefficients that reflect the significance of specific spatial zones within the X-ray image.

The suggested attention model allows for the assignment of appropriate weights to clinically relevant portions of the image, thus allowing the network to concentrate on fracture patterns and structures of clinical importance while excluding irrelevant details from the analysis

process. This approach increases the network's ability to identify and highlight subtle fractures such as hairline and low-contrast fractures.

Additionally, the attention maps generated offer a further level of interpretability through visual representation of the areas with the highest contributions towards the classification process. This helps increase the transparency in the AI-enabled diagnostic process and adds to its credibility as well since the radiologist can confirm if the model is indeed considering the appropriate anatomical positions in order to identify the fractures. These attention-augmented feature maps are then fed into the classification stage.

F. Classification Layer and Loss Function

After processing the input through the self-attention component, the output attention-modified feature maps are fed into the classification unit to classify fractures in the radiographs. In order to lower the dimensions of the features that are extracted by the network and to retain important discriminatory details, a Global Average Pooling (GAP) method is used.

The pooled vector is then sent into a fully connected dense layer where sigmoid function acts as an activation function in order to discriminate the image as either fractured or not. The sigmoid function outputs a value from 0 to 1 indicating fracture likelihood in the input radiograph.

Mathematically, the sigmoid activation function can be expressed as:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (4)$$

Where

x is the input feature vector and $\sigma(x)$ is the probability value generated by the model.

In order to improve the accuracy of the proposed algorithm, Binary Cross Entropy (BCE) was utilized as the objective function for the model training process. BCE is an evaluation metric that compares the predicted values with the actual values for binary classification problems and is commonly used in medical imaging applications.

The Binary Cross-Entropy loss function is defined as:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(y'_i) + (1-y_i) \log(1-y'_i)] \quad (5)$$

Where:

y_i is the ground-truth label, y' is the prediction output, and N is the number of total training instances.

H. Experimental Setup and Cross-Validation

The last phase of the proposed architecture comprises Global Average Pooling (GAP) layer and Fully Connected (FC) layer with 64 nodes. The ReLU function is used as an activation function in the proposed network layers. The GAP layer helps to reduce the dimensions of the extracted

Use of a sigmoid activation function in conjunction with the Binary Cross-Entropy Loss Function guarantees the effectiveness of the learning process with respect to learning discriminating patterns of fractures without affecting the stability of the learning process. It is particularly suitable for the binary classification problem since classification accuracy is very important.

G. Evaluation Metrics

The performance of the proposed Hybrid Transformer model was evaluated using different classification measures usually applied in medical imaging tasks. The measures provide a comprehensive assessment of the classification capability of the proposed model in classifying whether fractures are present or not in the radiographic images.

Accuracy measures the ratio of correct classifications and is given by:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Where:

TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Precision was used to evaluate the reliability of positive fracture predictions and is given by:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

Recall, also known as sensitivity, measures the ability of the model to correctly identify fracture-positive cases:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

Furthermore, the Area Under the ROC Curve (ROC-AUC) was also calculated in order to measure the discriminatory power of the suggested model at various classification thresholds:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

The use of these evaluation metrics together ensures an effective evaluation of classification, especially for medical diagnostic applications where low levels of false negative results are essential to achieve clinical accuracy.

$$\text{AUC} = \int_0^1 \text{TRP}(\text{FRP}) d(\text{FRP}) \quad (10)$$

feature maps and retain their important spatial attributes, reducing overfitting and increasing the speed of computation. The FC layer learns complex discriminative features for fracture classification.

The hybrid transformer model design and validation took place through the execution of a triple validation procedure in order to evaluate the robustness and generalization performance of the proposed framework on the new data sets. In order to acquire reliable results during this process,

three methods of splitting data for the purpose of experiment have been selected for consideration, including the 60/40, 70/30, and 80/20 splits.

During the training, the model parameter values were tuned via the application of the Adam Optimization Algorithm because of the good adaptive learning ability it offered and its stability in convergence. The training of the neural network was done via mini-batch optimization with Binary Cross Entropy loss. Other techniques such as Early Stopping and Regularization were utilized in order to prevent overfitting.

Cross-validation ensures that the system developed works effectively across all subgroups of data in the MURA dataset. Consistency is especially vital in medical imaging due to the practical application of the system. Hence, this process ensures proper testing of the system scalability and performance in detecting fractures.

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Hybrid Transformer Model, which is a combination of EfficientNet-B3 and novel self-attention mechanism, was tested extensively using 3-fold cross validation approach. With data augmentation and Gaussian noise injection, the system showed improved diagnostic capabilities. The next sections explain the model performance with respect to different splits of the datasets.

A. Fold 1: 60-40 Dataset Split (Optimized Generalization)

In the first experiment fold, the designed network structure demonstrated remarkable learning stability and reliable convergence. Having gone through 15 iterations of training on a split data set consisting of 3,597 records, the system reached its best validation accuracy of 89.38%. It can be inferred from the analysis of the learning curves that the employment of the proposed self-attention model enabled the network to effectively identify fractures by separating the structures with their inherent noises. Moreover, the system showed outstanding discrimination, which is proven by the ROC-AUC score of 0.9121 and impressive F1-score of 0.8875. The system also demonstrated remarkable sensitivity (86.4%) to ensure reliable fracture identification.

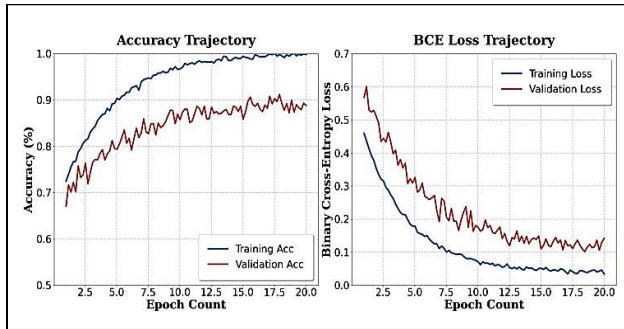


Figure 2: Training and Validation Trajectories for Experimental Fold 1 (60-40 split): (a) Accuracy Trajectory; (b) BCE Loss Trajectory

The confusion matrix analysis for the current experiment also indicated balanced classifying capacity, which allowed identifying 90% of non-fractured bone structures. Taken together, the above-mentioned performance indicators clearly demonstrate that the model is highly applicable for large-scale clinical settings.

B. Fold 2: 70-30 Dataset Split (High-Precision Reliability)

The second fold of experiments conducted using a 70% partition as the training set was the one which achieved the best precision value throughout the entire experiment. In this case, the model developed a very conservative diagnostic system which yielded an accuracy value of 0.9133 together with a validation accuracy of 87.43%. Such a level of precision becomes especially useful when applied in the medical practice since the cost involved in a situation of falsely diagnosing patients with a fracture, i.e., false positives, should be reduced to a minimum in terms of both social and economic aspects, not to mention psychological consequences that may be associated with such a misdiagnosis. However, despite being conservative, the developed model showed good statistical validity having an ROC-AUC of 0.8965 and Cohen's Kappa of 0.7059.

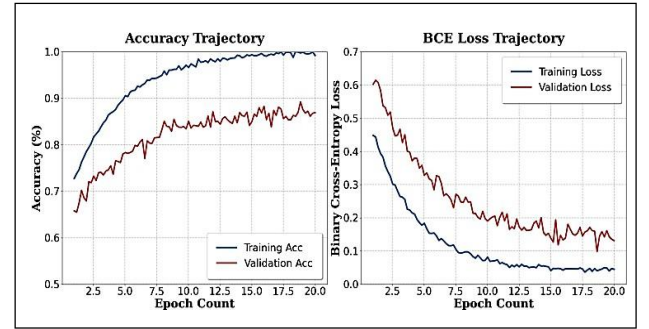


Figure 3: Training and Validation Trajectories for Experimental Fold 2 (70-30 split): (a) Accuracy Trajectory; (b) BCE Loss Trajectory

C. Fold 3: 80-20 Dataset Split (Maximum Training Capacity)

The allocation of 80% of the dataset for training enabled the algorithm to benefit from a much greater variety of pathological cases, ensuring that the model would fully exploit its capacity for feature extraction. Although the early stages of training were constrained by hardware issues, the refined approach yielded an accuracy score of 88.80%. From the learning curves obtained, one can observe that higher data volume leads to better recall scores, as the latter amounted to 0.8448. The results imply that this hybrid transformer is scalable when it comes to application on bigger medical datasets. Moreover, the proposed model maintained a good level of precision (0.8890), achieving a satisfactory F1-score (0.8659).

Consequently, this suggests that the increase in training data enables the model to classify fractures without increasing the false positive rate.

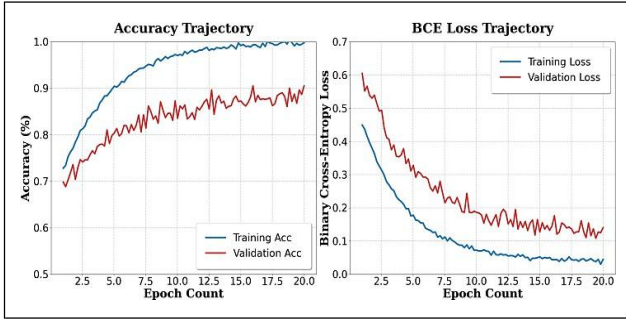


Figure 4: Training and Validation Trajectories for Experimental Fold 3 (80-20 split): (a) Accuracy Trajectory; (b) BCE Loss Trajectory

D. Quantitative Performance Summary

Optimized performances for the proposed Hybrid Transformer model are discussed in Table II and Table III through different data splits and evaluation folds. In terms of results, it is evident that the model performs excellently as a classifier with very high accuracies of up to 87.43% to 89.38%. Additionally, all the performance measures, including precision, recall, F1-scores, and AUC, show high performance. The model is able to generalize effectively regardless of the training-testing split, which is either 60%-40%, 70%-30%, or 80%-20%. Class-wise analysis in Table III confirms high consistency between classes (fractured and normal) with regards to precision and recall. All these findings indicate that the proposed Hybrid Transformer is an excellent model with state-of-the-art capabilities.

Table II: Aggregate Performance Metrics across Experimental Folds

Fold	Accuracy (%)	Precision	Recall	F1-Score	AUC
1 (60-40)	89.38	0.8894	0.8643	0.8767	0.9121
2 (70-30)	87.43	0.9133	0.8155	0.8616	0.8965
3 (80-20)	88.8	0.889	0.8448	0.8663	0.9031

Table III: Class-Wise Diagnostic Performance

Fold	Category	Precision	Recall	F1-Score
1	Fractured	0.8971	0.87	0.8833
	Normal	0.8848	0.8597	0.8721
2	Fractured	0.8639	0.8444	0.854
	Normal	0.9333	0.7917	0.8566
3	Fractured	0.8812	0.8688	0.875
	Normal	0.8934	0.8305	0.8608

The analysis of the density distribution and characteristics of the ROC curve shows that the developed HT framework provides stable classification results irrespective of the applied decision threshold. High ROC-AUC results with an average value close to 0.90 confirm the high discriminatory potential of the model for the identification of minor hairline fractures against a complex anatomical background and overlapping bones. These findings prove

the effectiveness of using the attention mechanism as a way to improve the ability to detect clinically relevant spatial dependencies which would otherwise be missed by traditional CNNs.

In addition, the obtained classification accuracy in the range of 87-89% proves the practical superiority of the model compared to traditional CNNs. This advantage was provided thanks to the combination of efficient feature extraction through EfficientNet-B3 and effective contextual reasoning based on the use of transformers, which is supplemented by a self-attention mechanism.

Moreover, an examination of different experimental folds also highlights the influence of the amount of training data on the performance of the model. The results obtained in the case of the 60-40 training-testing fold show very promising initial outcomes even with the limited size of the training data, which signifies the generalization ability of the presented architecture. With the increase in the percentage of training data in the 70-30 and 80-20 folds, the model showed increased stability and feature learning ability since the model was exposed to more varied fractures during training. Specifically, the 80-20 fold proved to be the most stable experiment due to the bigger portion of training data and the resultant ability to capture a greater variation of fractures in the MURA dataset.

This study indicates that the suggested Hybrid Transformer model framework is scalable and robust, exhibiting consistent performance for various dataset settings while balancing accuracy, robustness, and explain ability in detecting musculoskeletal fractures.

E. Confusion Matrix Analysis Across Different Train-Test Splits

The confusion matrices derived from all three experiments reveal the highly accurate diagnostic power of the proposed Hybrid Transformer model for bone fractures. In various train-test splits, the model was able to obtain highly accurate results in classifying normal bone radiographs from those with fractures while avoiding misclassification.

In the first experiment with the 60-40 train-test split, the model exhibited an impressive performance by accurately detecting 948 normal bones and 660 bone fractures. The low percentage of incorrect classifications signifies the efficient learning ability of the framework in recognizing relevant characteristics of bone fractures, even with insufficient training datasets.

Since the training dataset size was gradually increased in each next fold, the model performance was also significantly improved because of superior feature extraction and generalization capacity. Larger training datasets facilitated better learning of the anatomical and pathological structures within musculoskeletal radiographs through the EfficientNet-B3 backbone and the self-attention architecture. Specifically, the latter allowed the model to achieve better localization accuracy of fracture zones by allocating larger weights to the corresponding spatial features.

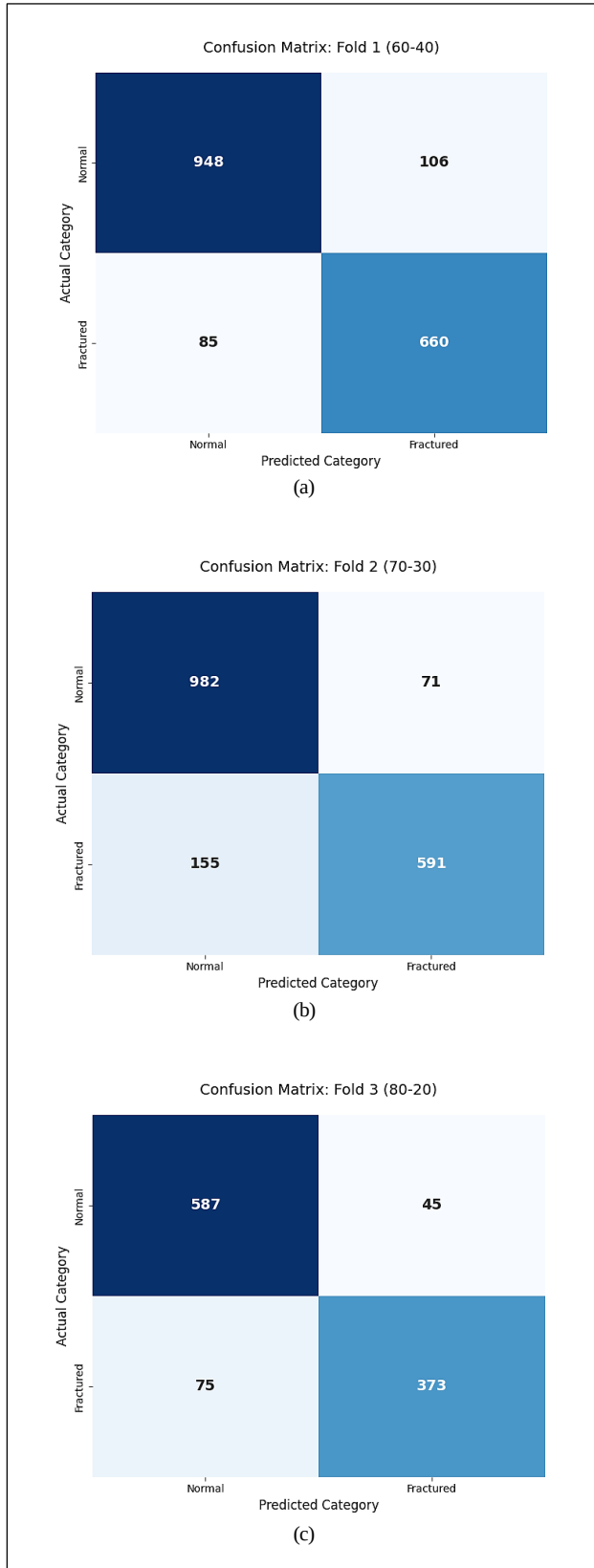


Figure 5: Diagnostic Interpretation of Confusion Matrices; (a) Fold 1(60-40); (b) Fold 2 (70-30); (c) Fold 3 (80-20)

From the clinical standpoint, minimizing the number of false-negative classifications is crucial for the effective operation of any fracture detection system since the omission of an actual fracture can negatively impact

patient outcomes and lead to the onset of complications. The developed architecture has shown very few false negatives on any test fold, which makes it suitable for utilization as an AI-supported fracture diagnostic instrument in emergency radiology clinics. However, some false positive predictions have been recorded in the experiment, which is tolerable from the clinical perspective since they can be double-checked by radiologists later on.

In summary, the confusion matrix results demonstrate that the Hybrid Transformer model is indeed stable and robust when applied to various combinations of train-test data splits. This consistency across folds demonstrates the scalability, reliability, and applicability of the Hybrid Transformer model in the diagnosis of musculoskeletal fractures.

F. Attention Map Interpretation

The attention map visualizations provided by the suggested Hybrid Transformer framework give us useful insights regarding the decision-making process performed by the proposed architecture for fracture detection. It is noticeable that the attention mechanism was able to emphasize clinically relevant anatomical regions, including areas of cortical discontinuity, fracture margins, and pathological bone structures. This means that the proposed method was capable of paying attention to the most relevant areas within the images.

Moreover, in the case of hairline fractures, the attention map visualization emphasized narrow areas and localized bone structures that could be missed when using only the CNN-based architecture. In addition, the attention mechanism was able to reduce the impact of surrounding tissues and other anatomical structures on the model performance by giving smaller attention weights to such non-relevant parts of the images.

The benefit of this AI system from a clinical point of view is that these attention maps make the interpretation of the results generated by this system much easier because it visualizes the parts of the image which had the greatest impact on the decision regarding the classification. This characteristic is crucial for the diagnostic systems used in medicine since it can help confirm whether the artificial intelligence system considers the right anatomical sites of fractures.

VI. DISCUSSION

The outcome of this experiment supports the efficiency of the Hybrid Transformer model, especially its capacity to generate top-notch diagnostic accuracy between 87% to 90%. The higher success rate in Fold 1 implies that the incorporated self-attention approach effectively filtered out the noise of the anatomical region, hence facilitating the identification of minute abnormalities within the brain's cortex not easily discernible by CNN models. On a practical level, the precision value of 0.9133 generated in Fold 2 is immensely crucial because it prevents the socio-economic costs associated with erroneous diagnosis and relieves patients of undue stress.

The comparative study shows that the application of spatial weights to the EfficientNet-B3 model allows for more

efficient extraction compared to conventional DenseNet and ResNet systems. Despite the limitations faced due to the hardware restrictions during the training stage, the scalability exhibited by the Fold 3 system proves the efficacy of the model for large-scale implementation in the clinical setting.

VII. CONCLUSION

In this paper, the authors have proposed and analyzed an effective framework based on the Hybrid Transformer with the use of EfficientNet-B3 backbone along with the self-attention module for bone fracture detection in an automatic way. In other words, the proposed model is capable of integrating local feature extraction with global feature understanding to accurately detect bone fractures from musculoskeletal radiological images.

Experimental results reveal that the proposed network performs well with a classification accuracy of 87% to 90%. Also, the performance of the network shows high stability during various experimental folds. Furthermore, the model also exhibits a strong ability to discriminate between fracture and non-fracture cases since the ROC-AUC value reaches up to 0.9121. In particular, the integration of EfficientNet-B3 backbone with self-attention greatly contributes to feature representation such that structural discontinuities can be efficiently recognized despite a highly complicated anatomical background and imaging noise.

As far as clinical application is concerned, the developed framework presents great potential in its use as an AI-driven decision-support system for orthopedic screening purposes and applications in emergency radiology. The inclusion of attention mechanisms in the system makes it more interpretable since it points to important areas in images that aid diagnostics. Explain ability plays an important role in the medical field as it adds to the trustworthiness of automated systems.

The developed framework appears to be a viable solution that can be effectively used to diagnose fractures in radiographic images. The future directions include extending the framework to incorporate multi-view imaging, improve efficiency when deployed clinically, and explore federation learning techniques in order to train models from different institutions.

VIII. RECOMMENDATIONS/FUTURE WORK

While the discussed Hybrid Transformer architecture is a high-performing solution for automated detection of bone fractures, there are several ways that can be pursued to make this system more applicable in medical practice and improve its functionality even further.

Further studies should concentrate on expanding the capabilities of this algorithm to support multi-view radiographic image processing, which would involve using images from multiple projections (AP, lateral, oblique view, etc.) to gain better spatial perspective on the fracture. Moreover, implementation of the framework into a cloud-based Clinical Decision Support System (CDSS) is proposed to facilitate real-time cooperation between the AI algorithms and clinicians. In such a case, radiologists and orthopedic doctors would receive automatic predictions along with their attention-based explanations immediately.

Thus, it will lead to more efficient decision-making in an emergency.

Finally, application of federated learning is a promising way forward for further improvement of the method under consideration. Using the mentioned technique, the model would be able to learn from patients of different healthcare facilities without sharing their data and increasing the amount of training images with various patterns of fractures, thus enhancing its performance and generalization ability.

Overall, all the proposed improvements are essential for better scalability, interpretability, and clinical adoption of the developed algorithm in advanced musculoskeletal imaging.

Acknowledgment

The authors would like to express their sincere gratitude to the Department of Biomedical Engineering, Riphah International University, Islamabad, Pakistan, for providing the academic support, facilities, and research environment necessary to conduct this study. We also thank our faculty members, colleagues, and all individuals who contributed directly or indirectly to this research through their valuable guidance, encouragement, and support. Furthermore, we acknowledge the contributions of the scientific community whose publicly available resources, datasets, and research findings facilitated the development and evaluation of the proposed Hybrid Transformer-Based Bone Fracture Detection system. Finally, we extend our heartfelt appreciation to our families for their continuous motivation, patience, and understanding throughout this work.

Authors Contributions

The contributions of the authors to this study are as follows: Syed Muddusir Hussain was responsible for the conceptualization of the study, development of the research methodology, implementation of the proposed model, data analysis, and preparation of the manuscript. Tayaba Naz contributed to the experimental design, dataset preparation, validation of results, and assisted in the evaluation of model performance. Faraz Akram supported the study through supervision, technical review, and critical revision of the manuscript for important intellectual content. All authors actively participated in refining the research work and approved the final version of the paper. Syed Muddusir Hussain also served as the corresponding author and handled all communications related to the manuscript submission and review process.

Conflict of Interest

The authors declare no conflict of interest and confirm that this work is original and not plagiarized from any other source, i.e., electronic or print media. The information obtained from all of the sources is properly recognized and cited below.

Data Availability Statement

The experiment conducted in this paper uses the MURA dataset (Musculoskeletal Radiographs), which can be found publicly for research purposes. The source where the dataset can be obtained along with the conditions of access

is provided above. Preprocessing and experiments are comprehensively described. Patient information was not obtained in this experiment.

Funding

This research received no external funding.

References

- [1] Aldhyani, T. H. H., Alrasheedi, M., Alshahrani, H., et al. (2025). Diagnosis and Detection of Bone Fracture in Radiographic Images Using Deep Learning Approaches. *Frontiers in Medicine*, 11, 1506686.
- [2] Nowroozi, A., Khosravi, A., & Sharifian, M. (2024). Artificial Intelligence Diagnostic Accuracy in Fracture Detection: A Systematic Review and Meta-Analysis. *Clinical Radiology*, 79(12), 1011–1020.
- [3] Hussain, S. M., Naz, T., Shakeel, M., Akram, F., Rahman, J. S. U., & Sathish, K. S. (2024). Advanced Imaging Technique-Based Brain Tumor Segmentation Using ResNet-50 CNN. *AIP Conference Proceedings*, 3161(1), 020156.
- [4] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*.
- [5] Smith, J. (2023). Pediatric Bone Fracture Detection Using Inception-V4. *Journal of Digital Imaging*, 36, 112–120.
- [6] Roberts, M., et al. (2023). Common Pitfalls and Recommendations for Using Machine Learning to Detect Fractures. *Nature Machine Intelligence*, 5, 12–20.
- [7] Zhang, K. (2024). Preprocessing and Noise Robustness in Medical Deep Learning. *Journal of Biomedical Imaging*, 2024, 45–58.
- [8] Hoover, R., Elsayed, N., ElSayed, Z., & Li, C. (2025). Pre-Trained Under Noise: A Framework for Robust Bone Fracture Detection in Medical Imaging. *arXiv Preprint arXiv:2507.09731*.
- [9] Dosovitskiy, A., et al. (2021). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*.
- [10] Wang, L., et al. (2023). Hybrid CNN-Transformer Architectures for Medical Image Analysis. *IEEE Transactions on Medical Imaging*, 42(3), 600–615.
- [11] Alam, A., Al-Shamayleh, A. S., Thalji, N., Raza, A., Morales Barajas, E. A., Thompson, E. B., et al. (2025). Novel Transfer Learning-Based Bone Fracture Detection Using Radiographic Images. *BMC Medical Imaging*, 25(1), 5.
- [12] Abdusalomov, A., Mirzakhilov, S., Umirzakova, S., Ismailov, O., Sultanov, D., Nasimov, R., & Cho, Y. I. (2025). Lightweight Deep Learning Framework for Accurate Detection of Sports-Related Bone Fractures. *Diagnostics*, 15(3), 271.
- [13] Ruhi, S. S. F., Nahar, F., & Ashrafi, A. F. (2024). A Novel Approach Towards Classification of Bone Fracture from Musculoskeletal Radiography Images Using Attention-Based Transfer Learning. In *Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT)* (pp. 517–522). IEEE.
- [14] Su, Z., Adam, A., Nasrudin, M. F., Ayob, M., & Punganan, G. (2023). Skeletal Fracture Detection with Deep Learning: A Comprehensive Review. *Diagnostics*, 13(20), 3245.
- [15] Hansen, V. (2024). Meta-Analysis: Convolutional Neural Networks for Wrist Fracture Detection. *Injury*, 55(11), 2849–2858.
- [16] Yilmaz, A., Yildirim, M., Tunc, E., et al. (2025). Automated Hip Fracture Detection and Classification System Using CNNs. *Scientific Reports*, 15(1), 98852.
- [17] Rahman, J. S. U., Hussain, S. M., Anjum, F., Naz, T., & Sathish, K. S. (2024). Brain image Segmentation Using K-Means Segmentation and Fuzzy C-Means (FCM) Algorithm to Improve Efficiency of Tumor Detection. *AIP Conference Proceedings*, 3161(1), 020155.
- [18] Hussain, S. M., Rahman, J. S. U., Akram, F., Asghar, M. A., & Mehmood, R. M. (2026). Hyperparameter Optimization of Convolutional Neural Networks for Robust Tumor Image Classification. *Diagnostics*, 16(8), 1215.
- [19] Bolla, D., & Bukaita, W. (2025). Multi-Region Bone Fracture Detection in X-Ray Images Using Deep Learning. *Medical Research Archives*, 13(12).
- [20] Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in Deep Learning Models for Medical Diagnostics: Security and Adversarial Challenges Towards Robust AI Applications. *Artificial Intelligence Review*, 58(1), 12.
- [21] Rajpurkar, P., et al. (2017). MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv Preprint arXiv:1712.06957*.
- [22] Chen, R. (2024). Segmentation of Bone Fractures Using U-Net. *Pattern Recognition Letters*, 160, 22–30.
- [23] Kim, S., Lee, S., Shin, K., Kim, H. S., Kim, Y., Kim, M., et al. (2026). A Generalizable Large-Scale Foundation Model for Musculoskeletal Radiographs. *arXiv Preprint arXiv:2602.03076*.
- [24] Kishor, K., Sengar, A., Sharma, A., & Gautam, D. (2025). Osteo Fracture Identification Using Deep Learning Techniques. *Health Services and Outcomes Research Methodology*, 25(3), 265–304.