ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

AN INNOVATIVE HYBRID MODEL FOR ELBOW BONE FRACTURE DETECTION: INTEGRATING VIT AND CNN

Muhammad Imran Khan[1]
(anmole97@gmail.com),
Javaria Amin[1]**,**
Muhammad Amir Shehzad[1]
Sajid Iqbal[2,*]
[1]Department of Computer Science, University of Wah, Wah Cantt, Pakistan
[2] Department of Computer Science and IT, The University of Lahore, Pakistan
***Corresponding Author : sajid.iqbal@cs.uol.edu.pk**

Abstract
Elbow bone fractures can be quite difficult to detect correctly and if a fracture goes misdiagnosed, it can be treated improperly, resulting in long term complications. Existing approaches such as manual assessment and Convolutional Neural Networks (CNN) based models also struggle with detecting subtle fracture patterns, leading to the demand for more dependable diagnostic aids. Precise classification and fast detection of bone fractures are crucial for efficient clinical diagnosis. Then, traditional techniques, using Convolutional Neural Networks (CNNs), have achieved a great progress, but they still struggle in classifying the subtle fracture subtypes accurately. To alleviate these flaws, the approach in this paper introduces a Hybrid Vision Transformers Convolutional Neural Network (ViT-CNN) model that combines the feature extraction capabilities of CNN with the attention mechanisms of ViTs, resulting in high performance improvement. This hybrid model benefits from the advantages of both architectures, improving the accuracy and reliability of diagnostics. The performance of hybrid model is found better than traditional CNN based approaches with respect to accuracy, sensitivity and specificity. Focusing on subtle patterns of fracture, this model is a powerful resource for increasing fracture detection accuracy and assisting clinicians with accurate diagnosis. The results show that the hybrid ViT-CNN model has the potential to make a substantial positive impact on the future of bone fracture detection and subsequently patient outcomes. The overall aims were to assess the performance successes of this hybrid approach in identifying elbow bone fractures, and whether this may have scope to further improve clinician diagnostic accuracy.

1. Introduction
A bone injury like fractures and deformities is an important part of medical imaging study. The traditional way of having radiologists review CT scans done for lung cancer is a standard practice but has its limitations because of the time it can take and the possibility of errors. This fields has been revolutionized by Artificial Intelligence, or more specifically Deep learning models such as Convolutional Neural Networks (CNNs) that have been yield astonishing accuracy automating the assessment of medical images (X-rays, CT, MRIs) [1].
Vision Transformers (ViTs) were recently introduced for computer vision tasks and showed remarkable image classification performance by utilizing attention mechanisms, thereby enabling global dependencies in images. While CNNs use convolutional filters for local features, ViTs take benefit of self-attention to digest larger contextual information. A hybrid framework retains CNNs feature extraction ability and at the same time, includes ViTs global context processing. Such a synergy overcomes the challenges from each model: fine-tuning ViTs requires significant computation cost and CNNs compete to learn global dependencies that have high time complexity, which is very unsuited for the task of bone detection [2].
Recently, due to their hierarchical structure which learns textures at lower levels and feature abstractions at higher ones, CNNs have become the reference for medical image classification, segmentation and detection tasks. However, CNNs suffer from the weak ability in modeling long-distance dependencies which is important for detecting subtle deformities or fractures. Researchers are investigating complementary models that analyze images in a more integrated way to help remedy this. [3]

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

Vision Transformers (ViTs) depend on on self-attention mechanisms which can capture the relationships between every pixel in an image which makes it suitable to overcome the limitation of CNNs. In contrast to CNNs, ViTs provide a global context, allowing them to identify more complex patterns in medical images and detect bones required both local and global features therefore in that A hybrid CNN+ViT model performed better as it combine significant features from both approaches [4].

Hybrid architecture which fuses CNNs and Vision Transformers take advantage of localized feature extraction and global context modeling. Whereas, the CNNs localize high-impact regions, such as bone edges, fractures and deformations, ViTs preserve the global alignment and context of the image. Using both Anglo-centric and bone-centric feature extraction we see a boost in detection performance and increase robustness against variations in bone structures between patients [5]. Technologically, hybrid models are increasingly being adopted in industries such as image classification, object detection, and even medical imaging, as they can lead to accurate medical applications, such as automated bone detection systems. Whereas the initial computational costs of Vision Transformers hindered their real-world use, these difficulties have been overcome with the development of hardware acceleration as well as efficient algorithms. ViT with CNNs are promising towards attaining high-accuracy imaging diagnostics towards bone [6].

Convolutional Neural Networks (CNNs) have demonstrated great strides in the initial diagnosis of bone fractures, and in combination with Vision Transformers (ViTs), they have reached even greater heights. CNNs work great for extracting local features and minor bone structures but lack in learning long-range dependencies (it becomes unaware of the entire anatomy of the bone). Self-attention allows ViTs to capture the global context across the entire image with spatial dependencies. This setup improves the CNN's local feature extraction and ViTs' context awareness, leading to higher accuracy, fewer false positives, and potential clinical value for detecting bone fractures [7].

The hybrid CNN + ViT model shows superior performance in detecting subtle, multisegmented, and complex bone fractures, especially in challenging anatomical regions. This approach unites the local feature extraction of CNNs with the global context processing of ViT, allowing it to effectively grip both global and local image details. However, this is particularly beneficial for fractures that are obscured by other anatomy or in situations with less obvious anatomical geometry, enabling the model to differentiate between normal and abnormal bone structure. Research has determined that this mixed approach significantly exceeds that of traditional CNNs on their own providing a useful accurate diagnostic aid to applications in clinical settings and thus may well enhance patient care [8].

Here are some key findings of this paper:

1. Improved Diagnostic Accuracy: The hybrid model, which integrates the feature extraction capabilities of CNNs with the attention mechanism of ViTs, is shown to outperform conventional CNNs with respect to diagnostic accuracy, sensitivity, and specificity.
2. Enhanced Fracture Recognition: Through the accurate recognition of slight fracture patterns that are usually overlooked in previously established techniques.
3. Support for Clinicians: This novel approach improves diagnostic accuracy and acts as an aid for clinicians facilitating accurate assessments for better patient care.
4. Clinical Impact: The hybrid mode, differentiate design and performance has the potential to transform diagnostic processes and improve patient outcomes in the medical attention field.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

The rest of the article is structured as follows: Related work is presented in Section II, detailed steps of the proposed method are described in Section III, and Results and discussion are included in Section IV. Section V: Conclusion Finally.

2. Related Work

The detection of bone fractures depends on the manual analysis of X-ray or MRI images, requiring a significant amount of time and expertise, leading to errors (especially in the case of subtle fractures). To mitigate this, researchers have focused on deep learning, and specifically Convolutional Neural Networks (CNN), which learn a hierarchical representation of medical images for fracture detection and classification. However CNNs tend to overfit on small datasets so to prevent that data augmentation methods i.e. rotation, flipping, and shearing are applied to increase the dataset size from few hundreds to around 4000 images. The proposed CNN model delivered 92.44% classification accuracy during experimentations, which was improved using optimizers such as Adam over traditional techniques such as Gray-Level Co-occurrence Matrix (GLCM) and contour extraction based methods with training and validation accuracy [9].

Fractures in hand bones are common and their accurate and timely diagnosis is important to ensure that they heal properly, but the assessment of small fractures based on manual evaluation of X-ray images by radiologists can lead to errors. In response to these challenges, this work proposes a hybrid deep learning model based on highly effective object detection models including YOLO NAS, EfficientDet, and DETR3 to directly localize hand bone fractures in X-ray images. In order to solve the problems associated with the quality of the dataset, the model is trained on a total of 4,736 hand X-rays which are divided into six classes. To enhance detection accuracy and reduce missed diagnoses, the hybrid architecture combines low-level feature extraction and high-level object detection. In ablation study, the composite of depth wise separable multi-convolutional & ReLU showed the highest score amongst the existing algorithms, amounting to 95.84% in testing, making it a commendable alternative to conventional CNNs like ResNet50 and EfficientDet, validating the capabilities of hybrid models in domains where guided deep learning is used to automate medical image analysis tasks like fractures [10].

As demand for radiological services has increased, backlogs of unreported studies have led to increased risk of missed or delayed diagnoses. In order to do so, AI-based solutions are being developed to support radiology workflows. In this study, transfer learning through deep convolutional neural networks (CNNs) using pre-trained networks (e.g. Inception v3) for fracture detection was employed. It achieved an unprecedented area under the curve (AUC) score of 0.954 for fracture classification, from retraining the top layers of the network to classify wrist radiographs. The small training dataset challenge was addressed by using data augmentation techniques including flipping, zooming and rotation which increased the dataset size and improved generalization across changes in wrist position and patient anatomy. The model was fine-tuned with several hyperparameters for improved training efficiency, achieving sensitivity and specificity of 0.9 and 0.88 respectively. The implications of our results are that transfer learning improves computational efficiency while preserving diagnostic ability, providing a path towards the automation of fracture detection and a solution to backlogs in radiology departments. The model's scalability to other medical imaging applications also highlights its promise in contemporary healthcare [11].

Review on edge based bone fracture detection from X-ray images. Edge detection is a critical aspect in Image processing to find areas where the brightness intensity changes significantly and the canny edge detection is particularly highlighted due to its ability to find strong and weak edges as well as its resilience to noise. This distinguishes fracture features from this early stage and sets a solid foundation for high-precision classification tasks in the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

later stage. Methods such as Histogram of Oriented Gradients (HOG) convert the edges into structured data, processed by machine learning algorithms to improve detection accuracy. It also compares the classifiers, and it concludes that Support Vector Machines (SVM) provide the more robust solution as it demonstrates 90% accuracy in the classification of bone fractures. One way to do this is to build an automated pipeline that utilizes these techniques to make the identification of subtle fractures from medical imaging much more efficient [12].

CNNs can distinguish fracture edges even in noisy images by applying strong image segmentation and feature extraction algorithms. Median filter and other preprocessing stages are used to remove noise from input images even though certain edge information is lost because we want to increase image quality. The experimental findings confirm that CNNs surpass the capabilities of traditional techniques like Sobel, Prewitt, and Canny edge detection, primarily in identifying subtle fractures. The above paper also proposes a new Spatial Fuzzy C-Means (SFCM) clustering approach which substantially improves fracture localization and segmentation using the spatial correlation of the pixels themselves. Overall, the combination of SFCM with CNNs promotes the accuracy and reliability of fracture detection, offering a global diagnostic framework to categorize fractures based on their stages. This system automated takes less time to diagnosis but still provide a reliable and accurate result for bone fracture detection [13].

Image preprocessing is focused on in this study to boost bone fracture detection systems. Gaussian filtering reduces noise while preserving sharp edges and contrasting features, making sure that fracture areas are well visualized. Sobel, Prewitt, Roberts, and Canny edge detection methods achieve segmentation of bone regions and separation of features related to fracture, all the while Canny edge detector is found be the most robust among all owing to its noise resistance and retention of fine details, the latter of which is important in the detection of subtle fractures. The study further investigates segmentation methods, reconciling the bone shaft to a common axis to facilitate systematic recognition of fractures. Addition of Hough Transform also enables the identification of a straight line and precise localization of fractures. The automated process aids detection whilst reduce workload on medical experts providing an excellent process for clinical applications [14].

The predefined combination of edge detection and segmentation techniques provides good accuracy for the segmentation of the fracture detection. Edge Detection: Preprocessing steps (such as noise reduction and image enhancement) are applied to enhance the quality of X-ray and CT images. The modified Canny edge detector with histogram equalization for enhanced contrast was the best among the methods evaluated and could detect thin fracture lines, which were not possible with traditional Sobel and Prewitt methods. Image processing strategies were employed to segment the affected bone region, using region-based and edge-based techniques to suppress the surrounding landscape for more focussed analysis. After segmentation, the classified images were classified as either fractured or non-fractured using classifiers such as SVM and KNN, and they achieved the accuracy rate of 85%. The system shows promise for clinical deployment, but the authors point out that cross-modal fracture detection, particularly from low-quality CT images, need to be further improved for better accuracy [15].

The conventional convolutional models used to derive the fracture detection have limitations, which is why this study utilizes the Hybrid-Attention (HA) mechanism integrated within the YOLOv8 architecture to tackle those restrictions. Attention mechanisms, including channel and spatial attention, are combined in this proposed architecture with the idea of focusing on the important features in X-ray images. So channel attention increases the important signal channels and spatial attention marks more critical areas, leading to robust extraction of features that are strongly useful for the process of finding minor cracks or fractures in

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

complex domain images. What is more, the nomenclature distinguished the Cross-Stage Partial (CSP) architecture included in the YOLOv8 backbone, improving the gradient flow and minimizing computation overhead, indicates the model efficiency for high-dimensional medical datasets such as FracAtlas. This design not only allows for localization and detection at the fracture tips, but also at the decoupled head. Including experimental outcomes yielding a 20% improvement on mAP 50, optimized computing cost density, and a relevance to real-time clinical scenarios [16].

CNNs based architectures detect fractures for limited bone areas including calcaneus, wrist, and thigh. This demonstrates that CNNs can be used across multiple datasets by simply resizing and normalizing your inputs in a preprocessing step. For example, calcaneus fracture detection using the SURF method for feature extraction, alongside ResNet-based architectures resulted in an accuracy of 98.0%, demonstrating the high potential for transfer learning to reduce both the overall training and processing time of models while maintaining diagnostic accuracy. It is also worth noting that the study investigates the application of object detection CNNs on wrist fractures and even matches the orthopedic specialty with this with the state-of-the-art AUC that reaches ~96%. We also used more advanced preprocessing techniques like horizontal flip and aspect ratio preserving rescaling to make model robust. The results highlight the increasing role of deep learning approaches for automating diagnostic work in clinical radiology [17].

In this Study, an image processing framework for bone fracture detection with X-ray and computed tomography (CT) images is proposed. The process starts with its preprocessing algorithms, including converting colored images to grayscale, followed by median filtering techniques to suppress the salt-and-pepper noise while also keeping important features intact. First, it ensures that edge detection is performed correctly in this case, using the Sobel method that designates points with a high variation in intensity as being a fracture line/edge. A K-means clustering was used on the regions in both color and intensity space to separate the fractured regions for segmentation, where a high precision in extracting fracture zones was achieved. Moreover, using a Gray Level Co-occurrence Matrix (GLCM), textural features (entropy, contrast, and homogeneity) were extracted from the images and classified with decision tree and neural network models, and reached an accuracy of 85% for classifying the images. This combined algorithm of preprocessing, segmentation and extraction also indicates a move towards automation of bone break detection [18].

In this study a novel dual-phase approaches for the detection and classification of diabetic retinopathy (DR) lesions is proposed. Segmentation was performed using the JSeg model built upon the ResNet-50 backbone of DeepLabv3+ yielding DSC and weighted IoU index results of 0.9820 and 0.9991, respectively, for micro aneurysms, indicating precision for even small lesions. With this innovative approach it outperforms current models, while dealing with issues of lesions with varied size and shape during segmentation. In the classification step, ResNet-101-based features (extracted) were optimized with the Equilibrium Optimizer (EO) and used in classification with Support Vector Machines (SVM) and neural networks (NNs). The model produced a classification accuracy of 99.97% for Grade-4 lesions using 10-fold cross-validation. The JSeg model achieved significantly better segmentation results compared to MSRNet, DeepLabv3, and RefineNet, and the detection of micro aneurysms was more reliable. Compared to the RefineNet (Optic Disc: DSC = 0.9183), the JSeg model's DSC (Optic Disc DSC = 0.9978) showed that JSeg can segment more complex structures without over segmenting. Neural Network classifiers performed remarkably on the classification phase as well, attaining mean ROC of 0.98 for all 10 folds. The model improved classification accuracy and ensured model stability across diverse datasets by helping to overcome common issues such as data imbalance and inadequate

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

feature selection through the incorporation of data augmentation and EO. These results establish a new standard for automated lesion classification in medical imaging [19].

This article explores the Performance of Vision Transformers (ViTs) and transfer learning in detecting kidney conditions such as cysts, stones, and tumors from CT radiographs. The dataset contains 12,446annotated images and enables strong representation for all diagnostic categories. Different pre-processing steps like image scaling, Z-normalizing and random rotation were used in this research model to improve its performance. They compared three of the best ViT variants (EANet, CCT, and Swin Transformer) against of deep learning models such as VGG16, ResNet50, and Inception v3. Hyperparameter tuning was to get classification accuracy for each model. The Swin Transformer recorded the best precision of 0.996, recall of 1.000 with 0.996 F1 score for kidney tumors among the models, resulting in an accuracy of overall 99.30%. The CCT and VGG16 also achieved impressive accuracies of 96.54% and 98.20%, respectively. Unlike these models, ResNet50 and Inception v3 exhibited weaknesses and only achieved on their best performance of 61.60%, as it was unable to capture fine-grained features. In addition, Swin Transformer and VGG16 provided better localization of abnormalities, allowing for more interpretable results as illustrated in the GradCAM visualizations. This study demonstrates the potential of using network architectures like Vision Transformers, specifically Swin Transformer to aid in the imaging diagnostic process [20].

3. Proposed Methodology

The proposed model Hybrid ViT-CNN that utilizes the advantages of ViTs and CNNs in a different manner for performing accurate detection of elbow bone fractures. The ViT observes the entire image and learns relationships holistically and the CNN portion of the model aims to recognize granular, rich features like edges and texture. Thanks to this combination, the model can recognize very small fractures, since it learns to see not only a general structure of the elbow, but all its details. The model creates highly accurate predictions by combining the local and global features, which is why it is a promising diagnostic tool for elbow fracture. It covers a gap that traditional practices couldn't fill, which could improve medical imaging. Figure.1 Shows the Integrated CNN and ViT for accurate elbow fracture classification.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

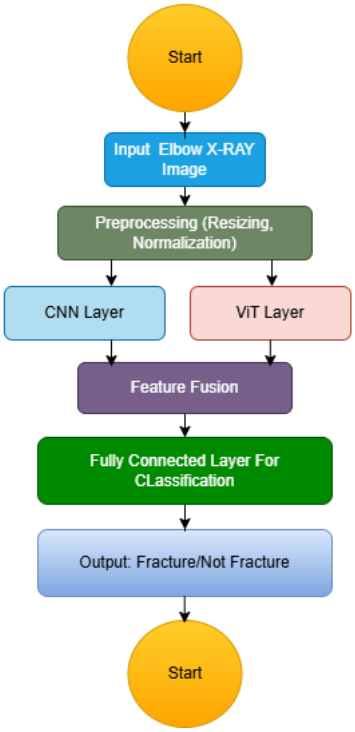


Figure.1 Flow Chart of Proposed Model

It starts with the input of elbow bone X-ray images and goes on to preprocess the data as shown in Figure.1. It then uses parallel CNN and ViT networks for feature extraction to utilize both local and global information. The feature fusion layer integrates these features and then passes it through a fully connected layer to obtain a classification. At last, the model returns fracture diagnosis that is whether or not there exists any fracture of elbow bone.

3.1 Hybrid ViT-CNN Model Architecture

Hybrid Feature extraction using ResNet-101 and tuning by the Equilibrium Optimizer (EO) was performed, followed by classification with Support Vector Machines (SVM) and neural networks (NNs). The model achieved a classification accuracy of 99.97% for Grade-4 lesions using 10-fold cross-validation.

Hybrid ViT-CNN is intended to take advantage of both CNN to acquire the local information and Vision transformers to acquire the global information. It is setup structure with:

- CNN Layers: Get low-level details such as textures and edges.
- ViT Layers: Investigate relationships throughout the entire image using self-attention techniques.
- Feature Fusion: Combines CNN and ViT output to generate a single feature vector for classification.

In the proposed methodology as shown in Figure.2 we represented the whole architecture of the Hybrid ViT-CNN model in detail.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

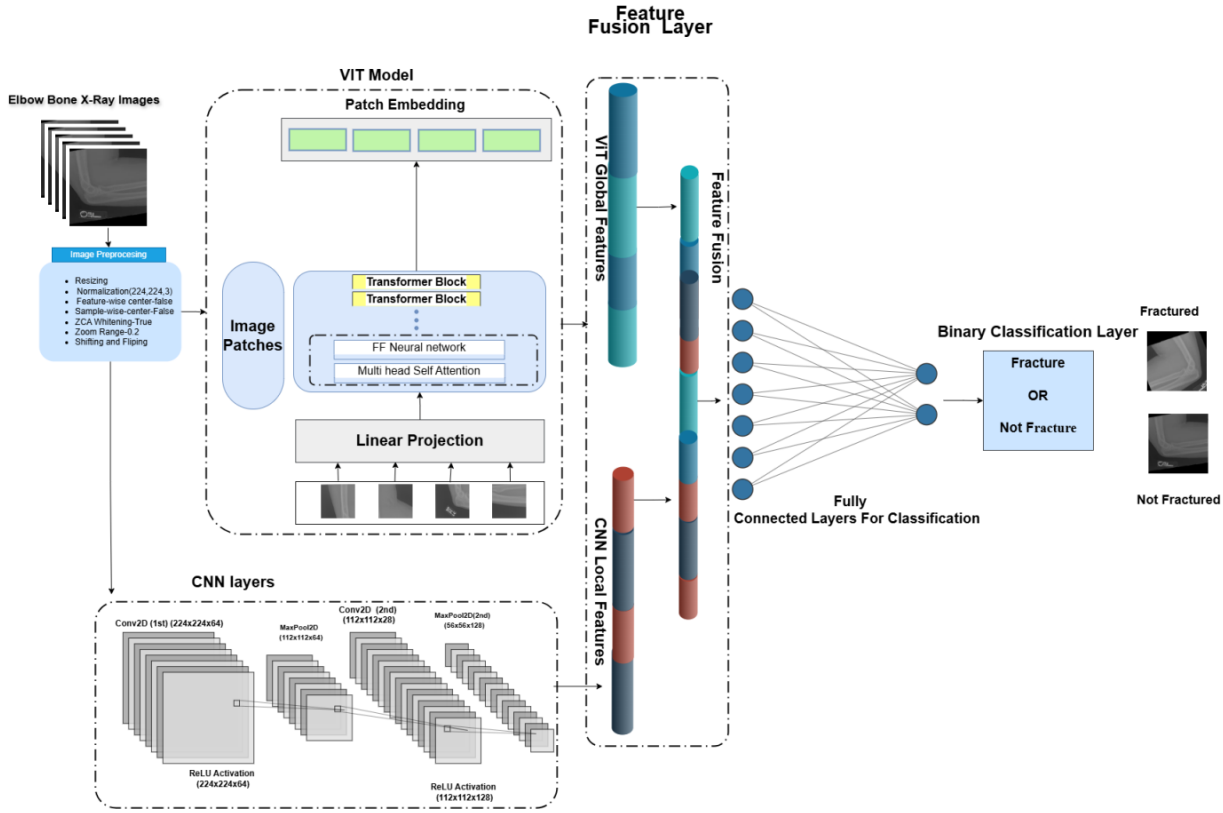


Fig.2 Hybrid ViT CNN Architecture of the Proposed model

The proposed method of identifying elbow bone fractures takes an elbow X-ray image as the first step. The first thing we need to do is proper preparation preprocess the images which represents some standardization of the data, such as scaling and normalization. Resizing ensures that every input image follows the same proportions, which facilitates the model few-shot representation. Normalization improves the training convergence of the neural network after scaling all the pixel values to a common range. The main goal of this stage is to prepare the X-ray image well for the next step (feature extraction) and maintain the accuracy and validity of the diagnostic information of the image.

After extracting features, the outputs of the CNN and ViT layers are aggregated into a single augmented feature representation. This stage of feature fusion fuses both localized and global features which in result makes the model aware of the overall input X-ray. Lastly, the fused features are fed to Fully Connected layers for classification. These layers use non-linear activation functions to capture complex patterns and relationships in the data and can be used to classify or verify if a fracture is present in the input X-ray. The final conclusion is the results of this categorization, which allows us to provide a correct and reliable diagnostic for the presence of a fracture or absence of it.

To understand the hybrid model, break the architecture explanation into three parts. The first part introduces the CNN, highlighting its capacity to produce spatial features in a sequential manner. The second part explains on the Vision Transformer (ViT) encapsulates the global dependencies and contextual relationships. The last part of them works on Feature Fusion, describing how to blend the results of CNN and ViT outputs to unify the advantages of both modalities and improve the performance of the model as a whole.

3.2 Convolutional Neural Network Architecture

The model as shown in Figure.3 CNN Layers consists of an input layer that first receives an image of 224x224x3 in height, width, and RGB channels, respectively. The input to our first convolutional layer consists of 64 filters, with a kernel size of 3x3, stride of 1, and padding of

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

1, which helps extract low-level features such as edges and textures, keeping the input dimensions intact (224x224x64). We then apply an activation function (ReLU in our case) to introduce non-linearity. The spatial dimensions are halved with a max-pooling layer that has a 2x2 kernel size and stride of 2, producing an output of 112x112x64, and this makes the model quite efficient. Another layer of convolution followed by a new activation activation function, so the second convolutional layer uses 128 filters of the identical kernel size, stride, and padding, resulting in deeper and more complex features to be extracted, with dimensions 112x112x128. The next layer is another ReLU activation followed by the second max-pooling layer which reduces the dimensions to 56x56x128, squeezing the features while keeping important spatial information. These multiple layers create hierarchical features from the input image that can be used in a classification (or decision) layer, with many individual neurons aggregating grouped output.

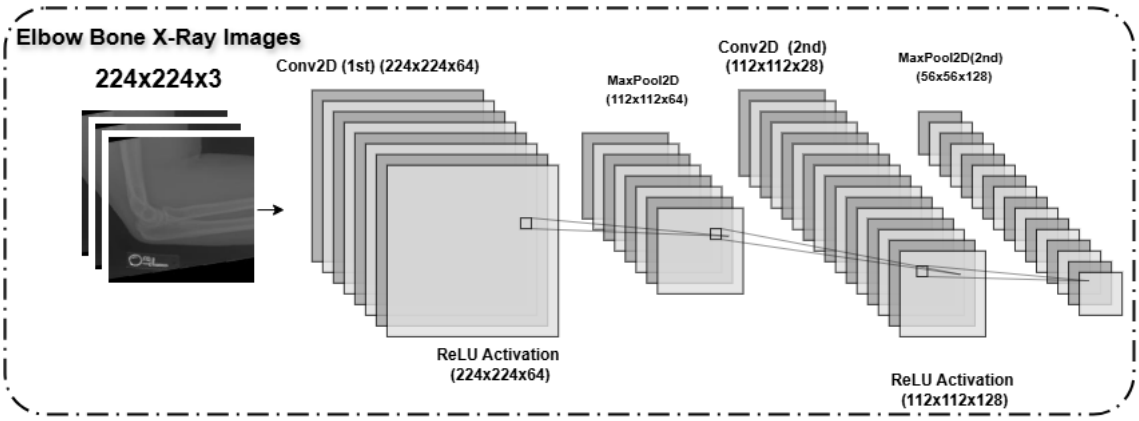


Figure.3 CNN Layers of the Hybrid ViT-CNN Model

$$(f * x)(o, v) = \sum_{m} \sum_{n} x(o + m, v + n) \cdot f(m, n) \quad (1)$$

- x(o,v): Input image (pixel value at (o,v) position)
- f(m,n): Convolution Filter(kernel)
- ∑m∑n: The summation calculates the weighted total of pixel values under the filter position as it moves across the image.
- This operation allows the model to recognize features like edges and textures, important for recognizing fractures in medical images.
- A max pooling is used after each convolutional layer which reduces the size of feature maps in the half retaining the most highlighted features, reducing computing complexity, and to not over fit.

$$MP(x) = \mathbf{max}(x) \quad (2)$$

3.2.1 ReLU Activation

We use Rectified Linear Unit (ReLU) as the non-linearity activation function so that the model can learn complex features.

$$R(x) = \mathbf{max}(0, x) \quad (3)$$

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

This forces the negative values of pixels to be zero and focuses a network on the meaningful specters [21].

3.2.2 Local Feature Detection

CNN Layer Feature like edges and texture are local feature which are combined to formulate higher level of patterns i.e in case of pattern detection in a radiograph and thus performed quite excellently in case of fracture detection [22].

3.3 Vision Transformer

Vision Transformer [2] in Figure.4 exemplifies how CNNs can be improved to utilize global context as well as long-range dependencies, which are particularly important for detecting fractures with nuanced patterns or extending over large regions.

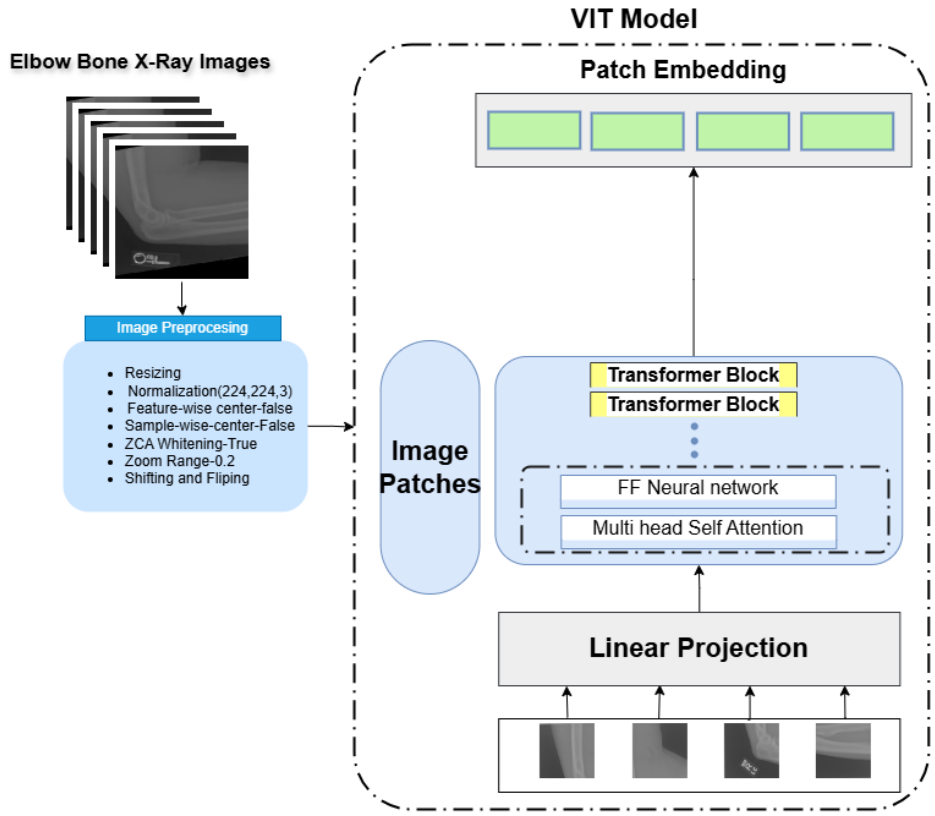


Figure.4 ViT Model of the Hybrid ViT-CNN Model

The module is as shown in Figure.4 initialized using a pre-trained model, typically trained on vast datasets. This model provides a strong foundation of images features that can be optimized for certain applications such as fracture detection.

It allows dividing the input image into fixed-size patches (like 16x16 pixels) and embedding each patch into a vector representation so that the image area can be independently analyzed for tiny fractures while inter-patch correlations can also be learned [23].

Self-Attention Mechanism: Whereas CNNs look at local patches and regions of interest, ViTs model relationships between all patches according to self-attention; Instead of reading the entire image, this helps the model to only pay attention to the relevant parts across the image.

$$\text{Attention}(\text{Y}, \text{E}, \text{A}) = \text{softmax}\left(\frac{YE^t}{\sqrt{d_E}}\right)\text{A} \quad (4)$$

Here, dk is the dimensionality of the keys and Y, E and A are the respective query, key and value matrices derived through patch embeddings. This approach allows the model to gain a

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

more complete understanding of the image, by bringing together proximity (local) and distance (global) relationships [24].

3.3.1 [CLS] Token for Global Features

- ViT uses a [CLS] (classification) token, a learnable embedding that summarizes global context of the image
- The output for the [CLS] token, which is a comprehensive representation of the image post-passing through the transformer layers, accumulates information from every patch [25].

3.4 Feature Fusion

The ViT module then passes local features obtained by the CNN module to a feature fusion technique to merge them with the global features. This enable both coarse and fine level features giving a accurate identification of fractures [26].

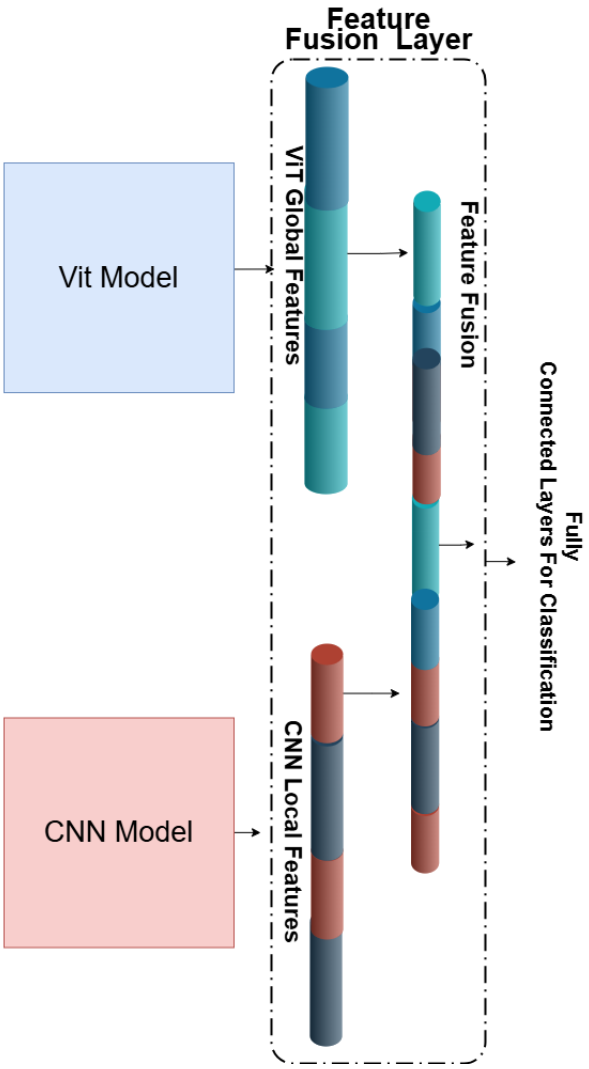


Figure.5 Feature Fusion Layer of Hybrid ViT-CNN Model

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

Feature Fusion as showing in Figure.5 incorporates local patterns from CNN layers and holistic dependencies captured by the global context from the ViT [CLS] token to perform precise identification of fractures.

$$F_{fused} = [F_{CNN}; F_{ViT}] \quad (5)$$

Here in this combination generates a fused feature vector of the CNN's spatial feature maps and the attention-based representations of the ViT, enhancing the fractural recognition performance.

3.4.1 Concatenation of Features

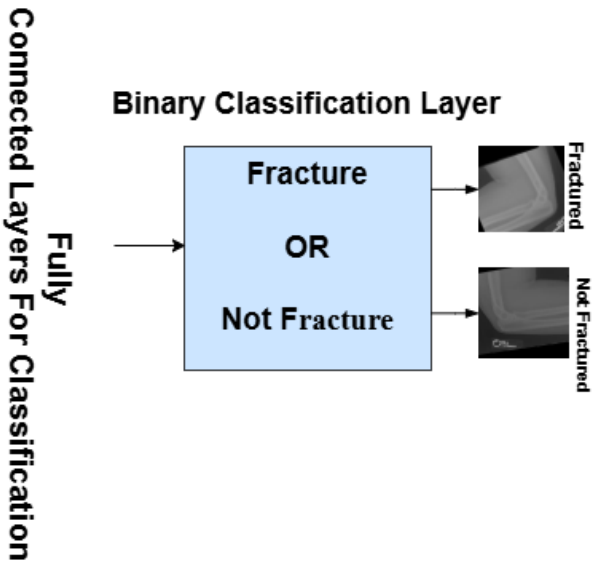
- To get feature vector, CLS token from ViT and CNN layers prior to flattening all outputs are concatenated.
- The ViT [CLS] token remains sensitive to global and contextual relationships across the entire image space, while CNN features only display setting-relational spatial features like edges or textures or patterns.
- As a result, the model can benefit from both large-scale dependencies (very general) and small-scale properties (very specific), which is critical in the domain of medical imaging [27].

3.5 Fully Connected Layers

- Once the concatenation is done on the 2 feature vectors, the resulting ones traverse through a 512-unit fully connected layer. It learns to aggregate per partition through a linear layer that improves the features interactions by minimizing the complexity of the feature space.
- To capture and learn complex patterns, the Rectified Linear Unit (ReLU) activation function introduces non-linarites in this model.

3.6 Binary Classification Layer

Figure.6 shows the binary classification layer of the proposed hybrid ViT-CNN model, which classifies the input image into a fractured elbow bone or a non-fractured elbow bone. The model uses features from the previous layers, and sends them to fully connected layers to be classified. Finally, the model gives a prediction whether the object being labeled is either Fractured or Not Fractured (as seen in the figures labeling the output).



ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

Figure.6 Binary Classification Layer of Hybrid ViT-CNN Model

- This improves feature extraction because it ensures that only positive values go to the next layer.
- The second fully connected layer takes the output from the first fully connected layer and processes the data further. This enables the model to condense the low-dimensional features into one score, thereby allowing the model to predict whether there is a fracture or not (as depicted in the **Figure.6**).
- The output from the last layer indicates class probabilities, with a higher probability indicating a stronger likelihood of a fracture. This is a classification job that is optimized through training with cross-entropy loss [28].

- For controlling the Class score (for instance, fractured or Not), the last classifications layer applies onto the procedure attributes:

$$O = W2H + b2 \quad (6)$$

Where:

- H:Input
- W2: Final classification layer weight matrix.
- B2: Bias vector.
- O:output

Artificial intelligence models derive an equation for binary classification task as shown below $O=W2H+b2$. Here, H, the features that are extracted after the previous layers of the model and these capture the most important information about the input. These features are fed through a fully connected layer with learned weights W2 and bias b2, allowing the model to adjust predictions. The outcome O is a score measuring the probability that the input is from one of the two classes. To give this score a meaning, it is typically converted into a probability through an activation function (for binary classification, we commonly use sigmoid function), this allows the model to correctly label the data as belonging to one of the two classes.

3.7 Cross-Entropy Loss Function

The Cross-Entropy Loss computes the difference between the predicted probabilities and true labels (y). It is defined as:

$$\text{CrossEntropy}(q, p) = -\sum_{i=1}^{g} q_i \log(p)_i \quad (7)$$

Where:
(q) i: True label for class g(0 or 1 for binary classification).
$(p)_i$: Predicted probability for class g .
(g): Number of classes
The loss ensures that the model learns to assign higher probability to the correct class by penalizing inaccurate predictions more severely.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

4. Experiment and Results

We used the MURA (Musculoskeletal Radiographs) dataset, being one of the largest publicly available sets of radiographic images designed for identifying pathologies in musculoskeletal disorders[29]. Therefore, we focused this investigation on the elbow X-ray subset of the dataset as it was created specifically to identify fractures and abnormalities in elbow radiographs. Musculoskeletal disorders are a leading global health challenge, affecting over 1.7 billion people worldwide and contributing to chronic pain and disability. The elbow-specific data from MURA dataset was very helpful for obtaining the CNN-ViT hybrid model I built and evaluated with great results for fracture diagnosis. Using this dataset, my algorithm would improve on the automated diagnostic tools and be especially useful in places where there are not enough qualified radiologists which would further the depth of healthcare and improve efficiency.

Table.1 Details of MURA Dataset [29]

Learning	Training		Testing		Total
ImagesType	Normal Images	Abnormal Images	NormalImages	AbnormalImages	
ElbowBone	1094	660	92	66	1912

Table.1 shows the details of MURA dataset[29]. It shows that there are 1754 images for training in which 1094 are normal and 660 are abnormal, and for validation there are total 158 images were used in which 92 normal and 66 abnormal images. So total images used for elbow bone detection are 1912.

Table.2 Data for Augmentation of the Proposed Model

Augmentation Performance	Value
Rotation is	30%
Width is	20%
Height is	20%
Shear is	20%
Zoom is	20%
Horizontal Flip is	True
Fill Mode is	Nearest

The various data augmentation methods employed to enrich the dataset for training the hybrid ViT-CNN model are described in Table 2. Horizon flipping, 30 degree rotation, 20% width and height shift, 20% shear and zoom ranges was applied to the data in order to increase the images diversity. These augmentations mimic characteristics of real-world environments, aiding the model to generalize better in varied input situations. Fill mode nearest helps to fill empty space in many transformations.

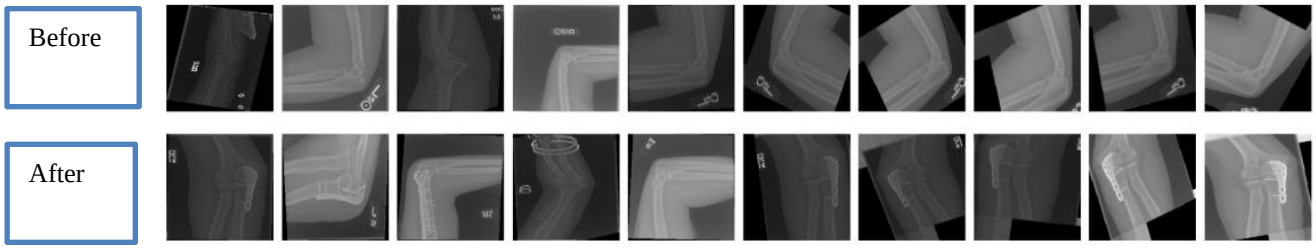
Table.3 Dataset details Proposed Model

Class Name	**Number of Images**		**Dataset Splitting 80-20 Ratio**	
	BeforeAugmentation	**AfterAugmentation**	**Training**	**Testing**
Fracture	2000	4000	3195	805
Not Fracture	2000	4000	3205	795
Total	4000	8000	6400	1600
Training Images	6400			
Testing Images	1600			

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

Dataset used for this study has been divided in Table.3. There were originally 2,000 images for each of the “Fracture” and “Not Fracture” classes. By doing augmentation, the whole dataset was increased to 4,000 images per class (i.e., 8,000 images). Data was divided into 80-20 into train and test. A total of 6,400 images were used for training (3,195 for "Fracture" and 3,205 for "Not Fracture") and 1,600 images were used for testing (805 for "Fracture" and 795 for "Not Fracture"), thus providing balanced representation of images in the training and testing data used to evaluate the model accuracy.

Figure.7 Data Augmentation for 10 sample Images from Proposed Model Dataet



The Figure.7 shows images of elbow bone fractures, before and after the augmentation. The first ten images are of the original dataset that exhibit the inherent and brute visual attributes. The next ten images show the augmented versions created by applying transformations including rotations, zooms, shifts, shears, and horizontal flips. By simulating different conditions for each image in a training epoch, these augmentations increase the diversity of the dataset, which allows the model to generalize better. It demonstrates the power of data augmentation in expanding the dataset to be more diverse and generalized. Here this visualization highlights how preprocessing plays a very critical role in improving the reliability of fracture detection models.

A Kaggle notebook was used to build and test the proposed model along with its GPU100 enablement to speed up testing and training. A Core I7 8th gen powered machine with Windows 10 was used for experiments.

Table.4 Hyper Para Meter of Proposed Model

Hyper parameter	Values
Batch Size	16
LR	1e-4
Number of Epochs	10
Image Resize	(224, 224)
Loss Function	Cross Entropy Loss
Optimizer	Adam

When using a batch size of 16, the model reads 16 images per training iteration, as seen in Table.4 a learning rate of 1e-4 means that the model will adjust its weights no more than 0.0001 at a time during optimization. Note: It is trained for a total of ten epochs (one epoch is a complete pass for the training dataset). All input works are resized to the size of (224, 224). The Adam optimizer is used to minimize the loss function with an adaptive learning rate depending on the momentum of the gradients viewed until that moment. In the case of classification problems, we use the Cross Entropy Loss Function, which is most appropriate to solve these types of problems.

In this section, we present the results of the proposed elbow bone fracture detection model based on hybrid CNN-ViT. It is evaluated on the basis of key measures using accuracy, precision, recall, and, F1-score which are obtained from the experiments done over the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

MURA dataset. A comparative analysis is provided to show that the model can, accurately and identified efficiently, reliably identifying the elbow fractures.

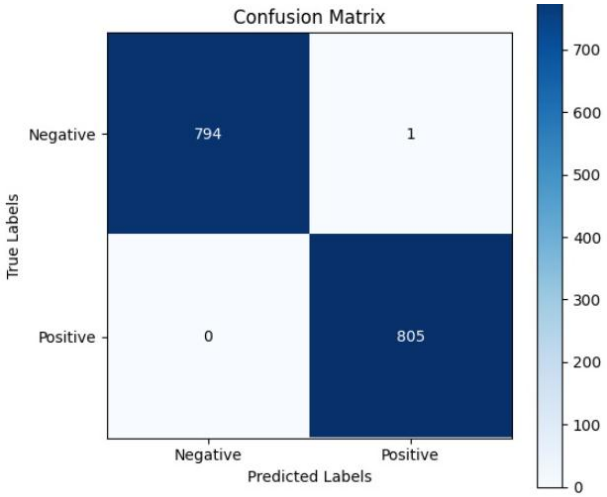


Figure.8 Confusion matrix of Hybrid ViT-CNN Model

Our Proposed model which is based on Elbow Bone Fracture Detection Using Hybrid ViT-CNN Model has confusion matrix in Figure.8 which is showing excellent performance on the test dataset. In the model, out of the 1600 images that were used, it correctly classified 794 healthy cases (true negatives) and 805 fractured cases (true positives), while it did one healthy case classified as fractured (false positive) and did not leave any fractured case out (false negatives). This suggests an almost perfect classification with an extremely low false positive rate and no false negatives, both of which are critical in medical diagnostics. The model appears quite confident about the result, with visualized high accuracy, precision and recall shown in the matrix.

Figure.9 Loss over epochs graph of Hybrid ViT-CNN Model

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

The graph in Figure.9 shows the loss of the proposed Hybrid ViT-CNN model to the epochs in training and validation. This is per the validation loss (orange line) and training loss (blue line) both plot lines show positive downward trends - suggestive that our model is effectively minimizing learning errors.

In Figure.9 showing 10 epochs for the training and validation loss of proposed "Elbow Bone Fracture Detection Using Hybrid ViT-CNN Model ". The rapid decline in both training and validation loss suggests that the model quickly learns the patterns in the data and significantly improves performance during the initial epochs. Both train and validation losses stabilize after 2nd epoch and they do go down as they progress over the epochs but the loss of model remains consistent low for each epoch. In both datasets, the fact that the loss is low and does not evolve significantly against epochs suggests that the model is well-optimized: there is no overfitting or underfitting as the validation loss follows training loss very closely. Thus, the tangential appearance of these curves indicates the model's ability of generalization on the unseen set, which is quite reliable and robust for detecting elbow bone fracture.

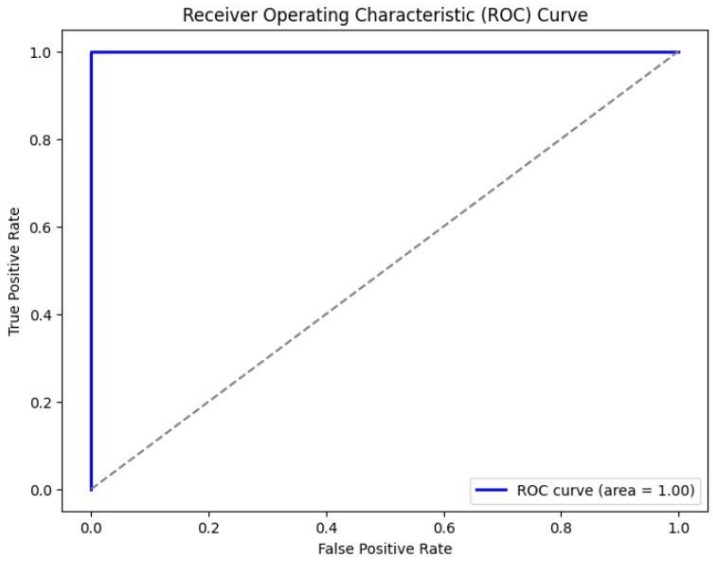


Figure.10 ROC Curve of Proposed Model

ROC curve in the Figure.10 for "Elbow Bone Fracture Detection Using Hybrid ViT-CNN Model" demonstrates excellent model performance. True Positive Rate (sensitivity) against False Positive Rate in the ROC curve, proposed model has AUC = 1.0. Here, such a perfectly accurate classification shows that it must be an ideal classification because it means that it is possible to distinguish perfectly shattered cases from healthy elbow cases.

A steep step curve towards the top left indicates that the model is obtaining a high true positive rate with an almost zero false positive rate. Such a result is desirable in medical applications, where both sensitivity and specificity are essential. The ROC curve illustrates strong performance of proposed model to detect fractures of the bone with no false positives.

Table.5 shows the results that detecting elbow bone fractures, the obtained accuracy of the hybrid model ViT-CNN was 99.93%. It accurately detected 794 out of 800 healthy cases and 805 out of 800 fractured cases out of a total of 1600 images with only one minor error (classifying a healthy case as fractured) and with no missed fractures. Its precision, recall, and F1-scores were almost perfect, with the averages being about 99.9% .This model can be used to accurately detect even the minutest fracture patterns making it a good supportive

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

asset for doctors to diagnose fractures improving the diagnosis quality & therefore improving the patient care round the clock.

Table.5 Performance Evaluation of the Proposed Methodology

PerformanceParameters	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	99.88	1.0	99.93	795
Positive	1.00	99.88	99.94	805
Accuracy	99.93			1600
Macro Average	99.94	99.94	99.94	1600
weighted Average	99.90	99.99	99.90	1600

In Table.6 based on the results, the Proposed Hybrid ViT-CNN model shows an impressive improvement, achieving 99.93% accuracy, 99.88% precision, 100% recall (sensitivity), and 99.94% F1-score, which are statistically significantly winning other methods. For instance, the hybrid deep learning model [30] achieved 84.99% accuracy and 80.44% F1-score, and VGG-16+ViT model [31] achieved 82.88% accuracy along with 88.24% recall. The Inception-ResNet-V2 model [32]followed behind with an 79.6% accuracy. The ability of the Hybrid ViT-CNN model to provide highly accurate and reliable diagnostics outperforms those of other methods and establishes a new reference point for fracture detection elbow classification tasks.The comparison results are shown in Figure 11.

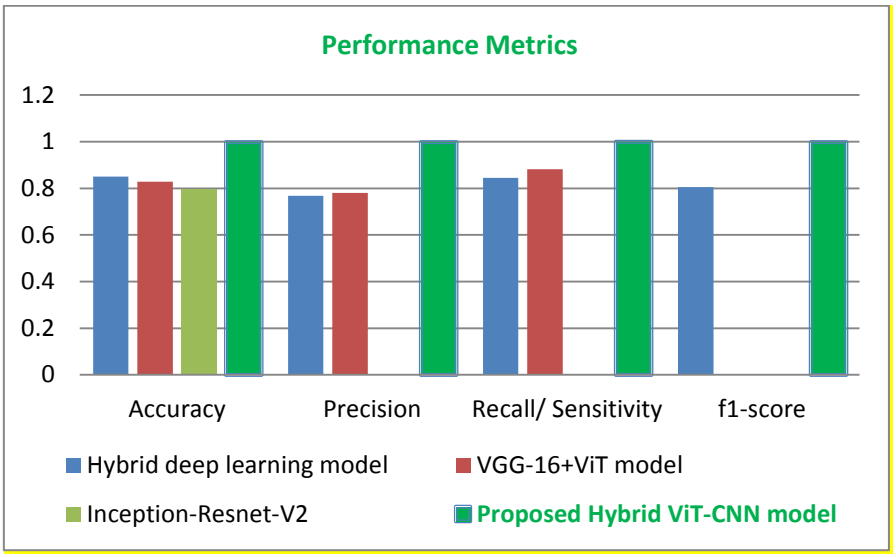
Table.6 Comparison of Specific deep learning methods on the elbow bone classification task in binary classification

Ref#	**Year**	**Dataset**	**Accuracy**	**Precision**	**Recall/Sensitivity**	**F11-score**
Hybrid deep learningmodel [30]	2025	MURA	0.8499	0.7676	0.845	0.8044
VGG-16+ViT model [31]	2024		0.8282	0.78	0.8824	-
Inception-Resnet-V2[32]	2024		0.796	-	-	-
Proposed Hybrid ViT-CNN model			99.93%	99.88%	100%	99.94%

Figure.11 Comparison of Performance Metrics across Various Models for Detecting Elbow Bone Fractures

The performance metrics for different models for elbow bone fracture detection compared, accuracy, precision, recall/sensitivity, and F1-score are shown in Figure.11 using a bar chart. The proposed Hybrid ViT-CNN has the highest values for all the metrics, with an accuracy of 99.93% and recall 100%. More importantly, it also highlights the superiority of the proposed model compared with existing methods.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)



4.1 Advantages and Applications

Proposed hybrid model combines the advantages of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), working together as a powerful tool to detect bone fractures. While CNNs excel in detecting intricate local patterns in images, ViTs introduce a larger context via attention heads. The two combine to form a model capable of detecting even the most subtle patterns of fracture with impressive accuracy. This combined approach reduces the likelihood of misdiagnosis, and provides clinicians with timely and reliable insights, making it a tool that can be pragmatically leveraged to improve patient outcomes.

5. Conclusion

This study proposed a hybrid ViT-CNN model for precise elbow bone fracture detection, leveraging a benchmark in musculoskeletal radiographs the MURA dataset. This approach allows the model to leverage the detailed spatial feature extraction from CNNs along with global attention capabilities from ViTs to achieve a state-of-the-art accuracy of 99.93%. To improve generalization to different data variations, data augmentation was applied. These findings underline the strength of the hybrid approach in the medical imaging domain, especially for automated diagnostic systems. Further research on extending this approach to other bone fracture types, incorporating multi-modal medical imaging data for more accurate diagnosis, and adapting the model for real-time clinical application is possible. Moreover, the application of explainability techniques may increase interpretability of the model's predictions for healthcare professionals.

References

1. Litjens, G., T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, and C.I. Sánchez, A survey on deep learning in medical image analysis. Medical image analysis, 2017. 42: p. 60-88.
2. Alexey, D., An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929, 2020.
3. LeCun, Y., Y. Bengio, and G. Hinton, Deep learning. nature, 2015. 521(7553): p. 436-444.
4. Ashish, V., Attention is all you need. Advances in neural information processing systems, 2017. 30: p. I.
5. Howard, A.G., Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

6. Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. in European conference on computer vision. 2020. Springer.
7. Satyaraj, D., D. Chitra, A. Prassath, L. Vigneash, and N. Murugan, Utilising a Deep Learning Approach for Bone Fracture Detection and Classification. International Journal of Engineering Research and Science & Technology, 2021. 17(1): p. 87-91.
8. Singh, S., Computer-aided diagnosis of thoracic diseases in chest X-rays using hybrid cnn-transformer architecture. arXiv preprint arXiv:2404.11843, 2024.
9. Yadav, D. and S. Rathor. Bone fracture detection and classification using deep learning approach. in 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC). 2020. IEEE.
10. Medaramatla, S.C., C.V. Samhitha, S.D. Pande, and S.R. Vinta, Detection of Hand Bone Fractures in X-ray Images using Hybrid YOLO NAS. IEEE Access, 2024.
11. Kim, D. and T. MacKinnon, Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clinical radiology, 2018. 73(5): p. 439-445.
12. Mohanty, S. and M.R. Senapati. Fracture detection from X-ray images using different Machine Learning Techniques. in 2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS). 2023. IEEE.
13. Sinthura, S.S., Y. Prathyusha, K. Harini, Y. Pranusha, and B. Poojitha. Bone Fracture Detection System using CNN Algorithm. in 2019 International Conference on Intelligent Computing and Control Systems (ICCS). 2019. IEEE.
14. Upadhyay, R.S. and P. Tanwar. A review on bone fracture detection techniques using image processing. in 2019 International Conference on Intelligent Computing and Control Systems (ICCS). 2019. IEEE.
15. Bhakare, D.B., P.A. Jawalekar, and S.D. Korde, A novel approach for bone fracture detection using image processing. International Research Journal of Engineering and Technology, 2018. 5(2): p. 193-195.
16. Meza, G., D. Ganta, and S. Gonzalez Torres, Deep Learning Approach for Arm Fracture Detection Based on an Improved YOLOv8 Algorithm. Algorithms, 2024. 17(11): p. 471.
17. Khatik, I. and S. Kadam, A systematic review of bone fracture detection models using convolutional neural network approach. Journal of Pharmaceutical Negative Results, 2022: p. 153-158.
18. Anu, T. and R. Raman, Detection of bone fracture using image processing methods. Int J Comput Appl, 2015. 975: p. 8887.
19. Amin, J., M.A. Anjum, and M. Malik, Fused information of DeepLabv3+ and transfer learning model for semantic segmentation and rich features selection using equilibrium optimizer (EO) for classification of NPDR lesions. Knowledge-Based Systems, 2022. 249: p. 108881.
20. Islam, M.N., M. Hasan, M.K. Hossain, M.G.R. Alam, M.Z. Uddin, and A. Soylu, Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Scientific Reports, 2022. 12(1): p. 1-14.
21. LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998. 86(11): p. 2278-2324.
22. Krizhevsky, A., I. Sutskever, and G.E. Hinton, Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012. 25.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.01 (2025)

23. Khan, S., M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, and M. Shah, Transformers in vision: A survey. ACM computing surveys (CSUR), 2022. 54(10s): p. 1-41.
24. Vaswani, A., Attention is all you need. Advances in Neural Information Processing Systems, 2017.
25. Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. in International conference on machine learning. 2021. PMLR.
26. Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 2017. 40(4): p. 834-848.
27. Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
28. Goodfellow, I., Deep learning. 2016, MIT press.
29. Rajpurkar, P., J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, and R.L. Ball, Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957, 2017.
30. Singh, G., P. Kumar, and D. Anand, Hybrid Deep Learning Model for Classification and Prediction of Abnormalities in Upper and Lower Extremities of Musculoskeletal Radiographs. SN Computer Science, 2024. 6(1): p. 32.
31. Dahal, S., R. Thapa, and S. Panth, A Hybrid Deep Learning Model for Musculoskeletal Abnormality Detection. 2024.
32. BS, V., K. lakshmi Buchupalli, R.R. Gottimukkula, and S. Kalimuthu, Enhanced Diagnostic Accuracy in Musculoskeletal Radiography: A Comprehensive Ensemble Approach of Deep Learning Models. 2024.