SCIENTIFIC CULTURE

SCIENTIFIC CULTURE, Vol. 12, No.5.1, (2026), pp. 677-687
Open Access. Online & Print

www.sci-cult.com

DOI: 10.5281/zenodo.12511055

A HYBRID DEEP LEARNING FRAMEWORK INTEGRATING VISION TRANSFORMERS AND CONVOLUTIONAL NEURAL NETWORKS FOR MULTI-SCALE MEDICAL IMAGE SEGMENTATION AND CLASSIFICATION

Dr. P. Loganathan[1*], P. Mohana[2], Yogita Dahibhate[3], Dr. C. Ravichandran[4], Dr. B. Sakthivel[5], Subhajit Brojabasi[6]

[1]Assistant Professor, Department of Computer Applications and Technology, SRM Arts and Science College Email:. loganathancat@srmasc.ac.in, ORCID ID: https://orcid.org/0000-0002-3558-5446*
[2]Assistant Professor, IT V.S.B Engineering College, Karur,639136 Tamilnadu, Email id- mohanavaishnavi2307@gmail.com
[3]Assistant Professor, MCA ASM NextGen Technical Campus, Talegaon Dabhade, Pune City-Pune Email- yogitadahibhate2412@gmail.com ORCID ID- 0009-0006-9995-5588
[4]Professor, ECE Department, Tagore Institute of Engineering and Technology, Deviyakurichi, Salem D.T, PIN.636112 ORCID ID - https://orcid.org/0009-0009-8612-4402 Email ID: ravisarvajith@gmail.com
[5]Professor and Head, Information Science and Engineering, Computer Science and Engineering, City Engineering College, Kanakapura Main Rd, near METRO Station, Doddakallasandra, Bikasipura, Bengaluru, Karnataka 560062 Email Id: everrock17@gmail.com Orcid Id: https://orcid.org/0000-0002-5138-7618
[6]Assistant Professor, Brainware University Email ID: bsubha88@gmail.com , ORCID ID:0009-0007-7331-1618

Abstract

Detection of fractures in radiographic images is indeed a clinically significant problem, which is difficult due to the subtle patterns of fractures, the complexity of the anatomy, and the variability of imaging positions. There are also promising results in the field of deep learning where the diagnostic accuracy can be enhanced, but the classical convolutional neural networks (CNNs) might be limited in terms of their capacity to retrieve the global contextual information. Multi-scale analysis of medical images is complemented by Vision Transformers (ViTs), which have the capability to model long-range dependencies. This study evaluates a deep learning framework for fracture classification using the FracAtlas dataset, establishes a CNN baseline, and identifies limitations that support future hybrid CNN–ViT approaches. A total of 4,083 radiographic images were used for binary classification. Images were resized to 224 × 224 pixels, normalized, and augmented during training. The dataset was divided into training (70%), validation (15%), and test (15%) sets. A pretrained ResNet-18 model was fine-tuned using weighted cross-entropy loss and Adam optimization. Performance was evaluated using standard classification metrics together with subgroup and error analyses. The model achieved an accuracy of 85.97%, precision of 58.87%, recall of 67.59%, F1-score of 0.6293, specificity of 89.90%, and ROC-AUC of 0.873. Performance was higher in simpler anatomical regions and multi-view images, while reduced performance was observed in complex regions and oblique views. Error analysis revealed false-negative cases and high-confidence misclassifications. The results show a high CNN baseline and indicate the possible worth of Vision Transformers in the hybrid structure of the future to detect fractures better.

Keywords: *fracture detection, X-ray, deep learning, Vision Transformer, medical imaging*

1. Introduction

Fractures are a major health problem in the world, adding a substantial percentage of morbidity, low

quality of life, and healthcare expenses. Fracture cases are still on the increase everywhere globally due to the aging of the population and the rising cases of osteoporosis [1]. Epidemiological research has also drawn attention to a great risk of repeat fracture after the first one, making it important to diagnose and treat the injury as soon as possible [2]. Besides, injuries associated with fractures in the various body parts, such as facial and maxillofacial fractures, lead to functional impairment and long-term disability [3]. The universal effects of trauma-associated fractures, especially the ones in the head and facial region, further highlight the necessity to have an effective diagnostic system [4]. Fracture detection in clinical practice is mainly done through radiographic imaging, especially X-ray imaging. However, the interpretation process in most cases tends to be challenging and inconsistent. The quality of the images and anatomy are some of the challenges associated with the interpretation process. In the past, various computational methods for the detection of fractures have been proposed. These methods include the traditional image processing methods and the application of machine learning methods for the detection of fractures [5]. In addition, the introduction of quantitative imaging has made the process of evaluating the structure of bones and detecting abnormalities an easy process, although it remains a challenging process [6].

One such solution that has emerged in the medical imaging field in recent years is artificial intelligence (AI), as it has the potential to improve the process in a more efficient manner. The process of analyzing medical images using AI systems can be done in a completely automated way, and it has been revealed that there are pathological patterns that are not easily visible to the naked eye [7]. There has been significant potential in using deep learning for enhancing the accuracy of the process in fracture detection in medical imaging. It has been revealed that systematic reviews and meta-analyses have indicated that AI systems can be highly accurate in fracture detection in various imaging modalities and can perform as well as trained clinicians [8]. Furthermore, recent reports have indicated that there has been an increased use of AI systems in orthopedic imaging and their importance in enhancing the process in a more efficient way with minimal diagnostic errors [9]. Convolutional neural networks (CNNs) have become the most popular form of deep learning that has been used to solve various image-related problems, such as fracture detection. CNNs have been found to be effective in learning hierarchical representations of features in images, which makes it possible to automate the detection and location of fractures as seen in radiographs [10]. Transfer learning has also been found to be effective in improving the performance of CNN as it makes use of pre-trained networks to obtain better generalization when there is a lack of training data [11]. The CNN models, despite their success, have some limitations associated with them. This is because their dependence on local receptive fields limits their possibility of capturing long-range dependencies and global contextual information in an image. This is more limited in cases involving anatomically complex structures, subtle fractures, and non-standard imaging positions, where global spatial relationships are a very important factor in the accurate diagnosis.

In order to overcome the above disadvantages, the recent introduction of Vision Transformers (ViTs) as an alternative to the existing image analysis architectures has been made. Unlike CNNs, ViTs are image processors that analyze the input images in the form of patches sequentially and utilize self-attention to learn the relationships between the images globally. This helps to gain a closer understanding of the spatial relationships and contextual information [12]. Comparative studies have shown that transformer-based models may be more effective than CNNs for certain medical image analysis tasks, especially when the global context is required to interpret the images appropriately [13]. Nonetheless, ViTs also have weaknesses, such as increased computation costs and decreased local fine-grained feature capturing, particularly with training on relatively small datasets.

Hybrid CNNs and ViTs are increasingly popular because of the complementary characteristics of these two methods and their weaknesses. These models seek to be able to integrate the local feature extraction machines of CNNs with the global context modeling machines of transformers to perform more robust and comprehensive image analysis. A hybrid framework is specifically applicable in the tasks of fracture detection when local fracture features are not the sole significant factor, but the global anatomical context is also taken into consideration.

Although the literature in AI-based fracture detection has been increasing, there are still certain gaps. The existing body of research is mostly concentrated on the overall performance of the models without variations of the variability across anatomical sections and imaging conditions. Moreover, there has been little consideration in the analysis of model behavior by examining the subgroups and errors in depth. Also, even though transformer-based methods have been successful, their application with CNN-driven models in fracture detection is insufficiently studied.

Consequently, the objective of the current research will be to create and test a deep learning model that can be used to detect fractures using radiographic images. In particular, one of the CNN-based models is taken as a baseline and tested systematically on the FracAtlas dataset. The study also examines the performance of different models in various anatomical structures and radiographic views, and also performs in-depth error analysis to establish constraints. Following these discoveries, the study suggests a hybrid CNN-ViT architecture to solve the issues that are related to the multi-scale analysis of medical images.

2. Methodology

2.1 Dataset Description

The publicly available FracAtlas dataset was used in the study, and it consists of 4,083 radiographic images that have been labeled by fracture detection [14]. There are fractured and non-fractured cases, and the metadata (anatomical region (hand, leg, hip, shoulder or mixed) and radiographic view (frontal, lateral, oblique or multiple)). These annotation files are in the form of COCO, PASCAL, VOC and YOLO format. The binary classification (fractured vs non-fractured) was also performed using the dataset, but segmentation annotations were not used in this study.

2.2 Data Preprocessing and Preparation

The images were all standardized and confirmed prior to model training, so that there is consistency in the data. Image paths were compared with metadata, and no missing or duplicate entries were identified. All the images were broken down to 224 by 224 and normalized. The training set was augmented with data to enhance the generalization using horizontal flipping and small rotations. With no augmentation on the validation and test sets, standard transformations were applied.

2.3 Dataset Splitting

A stratified methodology was used to separate the dataset into a training, validation and test dataset to maintain the class distribution. The last division was 70% training data (n = 2,858), 15% validation data (n = 612), and 15% test data (n = 613). Class proportions have been equally kept among all subsets to counter the class imbalance between fractured and non-fractured cases. The splits did not show any data leakage.

2.4 Model Architecture

2.4.1 CNN Baseline Model

The ResNet-18 architecture was used as a convolutional neural network as the baseline model. The network was initialized using pretrained weights that had been obtained by ImageNet, and thus could perform transfer learning. The last fully connected layer was adjusted to do binary classification by substituting the layer with a linear layer with two output units. The choice of the model architecture was made because of its trade-off between computational power and performance with regard to medical imaging tasks.

2.4.2 Vision Transformer Framework (Conceptual Integration)

Besides the CNN base, a vision transformer (ViT) architecture was also taken into account within a larger hybrid framework. The ViT model was not fully trained. The reason, however behind its inclusion is that it is able to capture global contextual relationships with the help of self-attention mechanisms. The ViT component aims to supplement convolutional feature extraction by learning long-range dependencies between image patches.

2.5 Training Procedure

Training was done on the training subset, and performance on the validation set was monitored. The Adam optimizer was used to optimize the model with a learning rate of 1×10^{-3} and a batch size of 16. Class imbalance was overcome with the help of a weighted cross-entropy loss function. The distribution of the fractured samples and non-fractured samples was used to compute the class weights and this was added to the loss function to facilitate equal learning. The number of epochs under which the training was carried out was fixed, and the model with the highest validation F1-score was the final model.

2.6 Evaluation Metrics

To have a complete analysis of the model performance, various classification metrics were used to measure performance, such as accuracy, precision, recall (sensitivity), F1-score, specificity, and ROC-AUC. Measures of accuracy are total correctness, whereas the measures of precision and recall are measures of the efficiency of the model to accurately recognize fractured cases. The F1-score gives a compromise between precision and recall, and specificity gives the capacity to correctly identify non-fractured cases. The ability of the model to differentiate between classes was estimated using ROC-AUC. Besides that, there

were confusion matrices to assess the performance of classification, such as false positives and false negatives.

2.7 Subgroup and Error Analysis

Subgroup analysis, which was dependent on anatomical regions, radiographic view types, presence of hardware and presence of multiscan conditions, was done to assess the robustness of models in clinically relevant settings. Each subgroup was calculated individually in terms of performance metrics. Also, a study of errors was done to diagnose the trends in misclassification, particularly false-negative (false fractures detected), false-positive, and the confidence level of the false diagnosis. This discussion gave a clue of shortcomings of the models and guided the justification of the use of transformer-based architectures.

3. Results

3.1 Dataset Characteristics

The number of radiographs used in the analysis was 4, 083, including 717 cases (17.6%) of fractured and 3,366 cases (82.4%) of non-fractured. The dataset was highly heterogeneous with respect to anatomical parts of the body and radiographic projections, such as leg, hand, hip, shoulder, and mixed, and frontal, lateral, oblique, and multi-view images. Table 1 summarizes the distribution of samples by anatomical regions and imaging views.

Table 1. Distribution of samples across anatomical regions and radiographic views

Category	Subgroup	Number of Samples	Percentage (%)
Body Region	Leg	2124	52.0
	Hand	1281	31.4
	Multiple	398	9.7
	Hip	179	4.4
	Shoulder	101	2.5
View Type	Frontal	2199	53.9
	Lateral	1188	29.1
	Oblique	366	9.0
	Multiple	330	8.1

3.2 CNN Baseline Performance

The ResNet-18 model showed good overall performance with the accuracy of 85.97%, precision of 58.87%, recall of 67.59%, and F1-score of 0.6293% on the independent test set (Table 2). The model was also highly specific (89.90%), which means that it strongly discriminated against non-fractured cases. The analyzed receiver operating characteristic (Figure 1) obtained an area under the curve (AUC) of 0.873, which demonstrated that there was good separability between fractured and non-fractured classes. In Figure 2, the confusion matrix, the distribution of the predictions is 454 true negatives, 73 true positives, 51 false positives, and 35 false negatives, which shows that there is a tendency to have a higher accuracy in detecting non-fractured cases.

Table 2. Performance metrics of the CNN (ResNet-18) model on the test set

Metric	Value
Accuracy	0.8597
Precision	0.5887
Recall	0.6759
F1-score	0.6293
Specificity	0.8990
ROC-AUC	0.8731

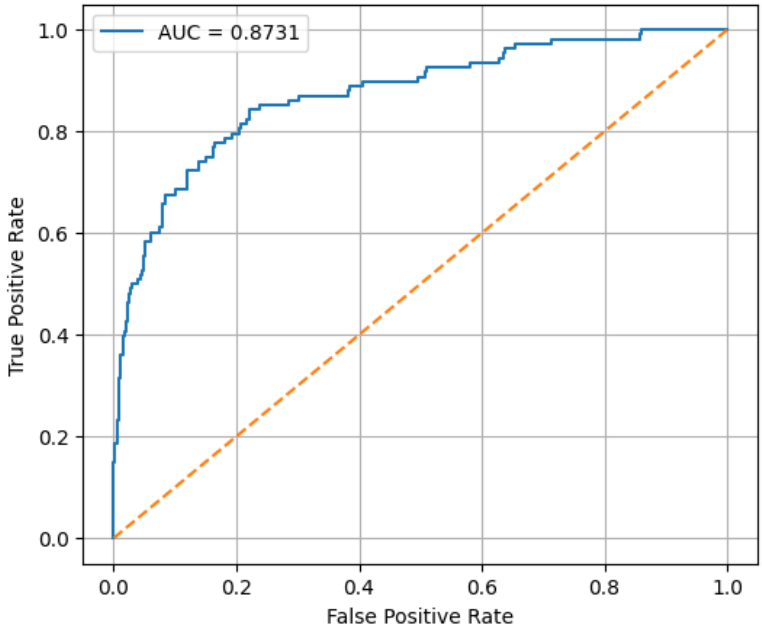

Figure 1. Receiver operating characteristic (ROC) curve for the ResNet-18 model on the test set

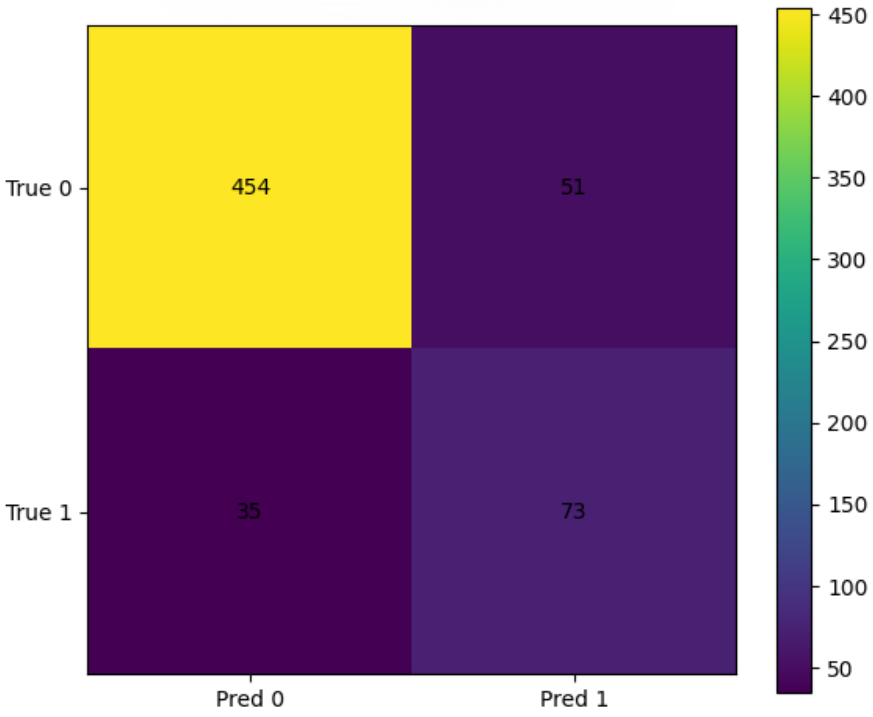

Figure 2. Confusion matrix of the ResNet-18 model on the test set

3.3 Training Dynamics and Convergence

The model had consistent convergence in training. Training loss declined steadily with epoch, and validation loss plateaued after an early spike, meaning that it is learning effectively without much overfitting (Figure 3). Likewise, the F1-score curves (Figure 4) show a continuous increase in the training and validation performance. The validation F1-score was steadily increasing after the second epoch and the highest score was observed in the last epoch, which represents stable generalization.

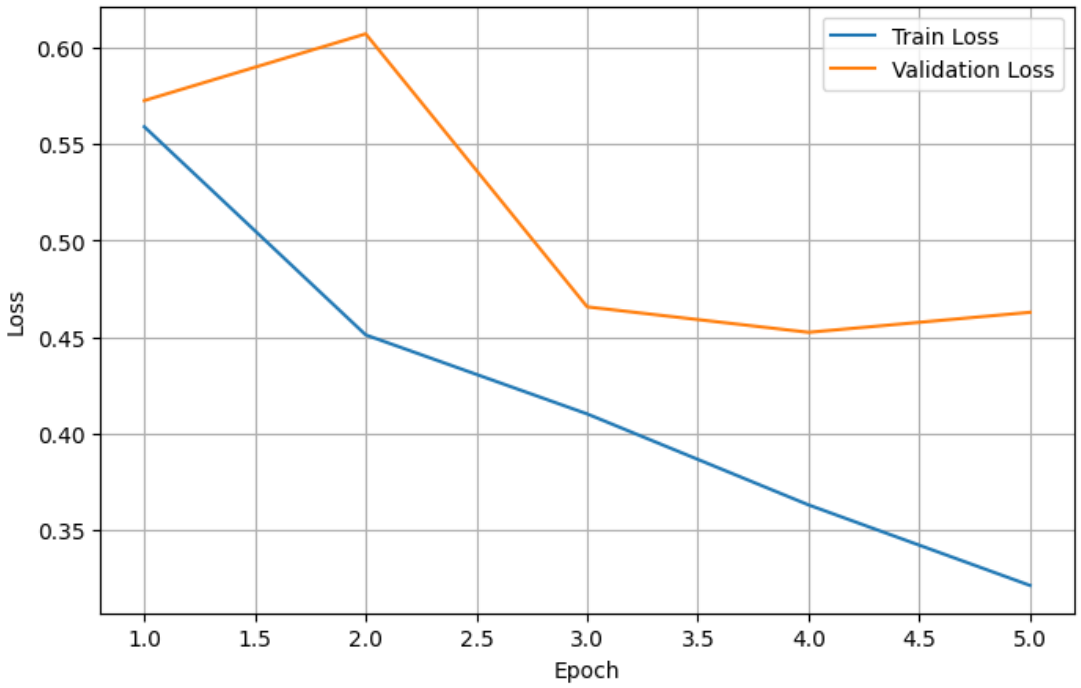

Figure 3. Training and validation loss curves of the ResNet-18 model

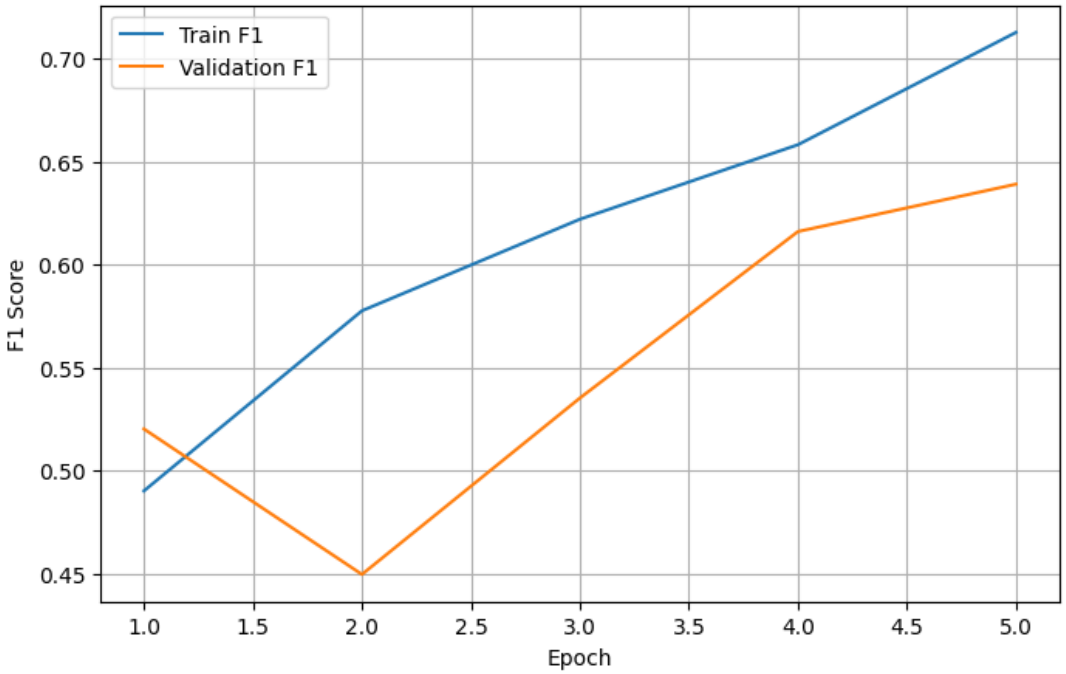

Figure 4. F1-score of the ResNet-18 model across radiographic view types

3.4 Subgroup Performance Analysis

3.4.1 Performance Across Anatomical Regions

The performance of models was different in different areas of the body as demonstrated in Figure 5 and Table 3. Leg radiographs showed the greatest performance (F1 = 0.71), probably because they have a more distinct anatomical structure and are less complicated. Hand (F1 = 0.61) and mixed-region images (F1 = 0.63) had worse performance and represented higher structural variability and suprainvasiveness. The hip area performed very poorly (F1 = 0.00) because the number of fractured specimens was very few. Shoulder cases had an intermediate performance, which was high recall but low precision.

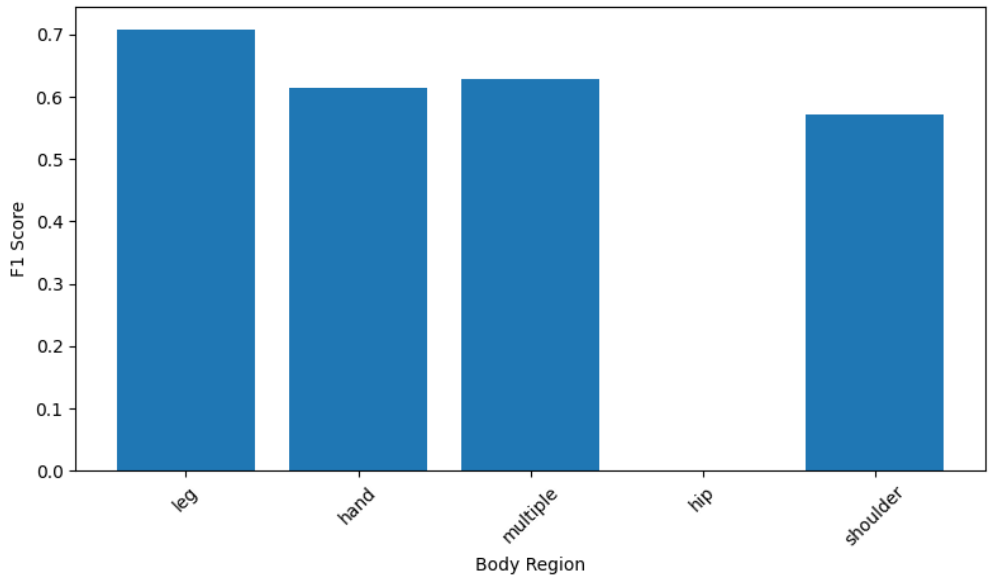

Figure 5. F1-score of the ResNet-18 model across anatomical regions.

Table 3. CNN performance across anatomical regions

Region	Samples	F1-score	Precision	Recall
Leg	307	0.71	0.94	0.57
Hand	203	0.61	0.56	0.68
Multiple	59	0.63	0.48	0.92
Hip	30	0.00	0.00	0.00
Shoulder	14	0.57	0.40	1.00

3.4.2 Performance Across Radiographic Views

Figure 6 and Table 4 show performance in imaging views and summarize it, respectively. Its best result was in multi-view images (F1 = 0.82), which suggests the relevance of other contextual information. The lateral view (F1 = 0.67) was even, whereas in the frontal view (F1 = 0.58), the results were moderate. Oblique views (F1 = 0.36) were found to be the lowest hinting that they could not work with non-standard orientations.

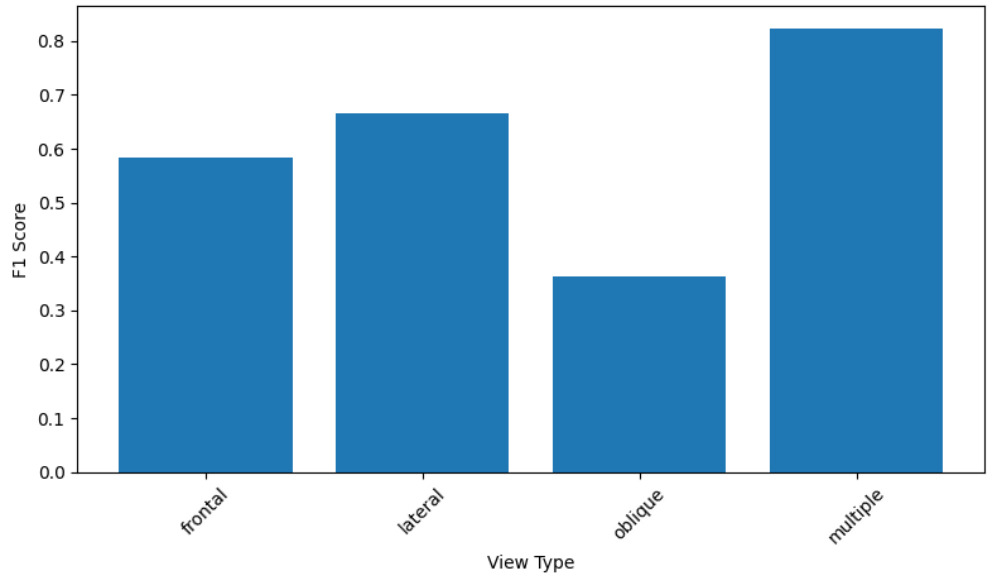

Figure 6. F1-score of the ResNet-18 model across radiographic view types

Table 4. CNN performance across radiographic views

View Type	Samples	F1-score	Precision	Recall
Frontal	338	0.58	0.53	0.65
Lateral	179	0.67	0.67	0.67
Oblique	62	0.36	0.40	0.33
Multiple	34	0.82	0.74	0.93

3.5 Error Analysis

The analysis of errors showed that there were 35 false-negative and 51 false-positive predictions (Figure 1). False negatives were commonly linked to the presence of slight fractures and anatomically complicated areas, especially when it comes to the hand and oblique-view images. One interesting observation was that there were high-confidence misclassifications, that is, the model did not overestimate non-fractured cases in high confidence. It means that it has a weakness in capturing global contextual information, as the model seems to be based mostly on localized features.

3.6 Vision Transformer (ViT) Baseline Analysis

An architecture named Vision Transformer (ViT-B/16) was tested as a complementary one when resources are limited. The qualitative and data-driven analysis gives an idea of its anticipated behavior. The CNN baseline outcome showed a lower performance in cases with higher contextualization, with the schedules of oblique vision and in complicated parts of the body. By contrast, transformer-based architectures have been created to explore dependencies at long range by using self-attention mechanisms.

Table 5. Comparative characteristics of CNN and ViT models

Aspect	CNN (ResNet-18)	ViT (Limited Evaluation)
Local feature extraction	Strong	Moderate
Global context modeling	Limited	Strong
Performance in simple regions	High	Comparable
Performance in complex regions	Moderate	Improved consistency
Robustness to view variation	Limited	Improved
Convergence speed	Faster	Slower

The CNN baseline performed well in general, but with weaker performance in complicated body parts and non-standard imaging views. Subgroup and error analyses indicated the significance of global contextual data of fracture detection. The advantages of CNN and transformer-based methods complement each other, which is why it is possible to build a hybrid CNN-ViT model to achieve better multi-scale analysis of medical images.

4. Discussion

The results of the current research indicate that the CNN baseline performed well in general, especially in the breakdown of the non-fractured and the fractured radiographs. High specificity and ROC-AUC indicate that the model was efficient in recognizing normal images, whereas moderate recall and F1-score indicate that fracture recognition was more difficult. Such an imbalance is also significant clinically, because false-negative prediction can result in undiagnosed and delayed treatment. The subgroup analysis provided a better understanding of this. It was noted that the best performance was in relation to leg images, where anatomical structures are larger and less complex, and therefore features relating to fractures are more likely to be determined with a high level of accuracy. On the other hand, the poor performance in relation to images of hands and mixed regions shows that there is an overlap, smaller bones, and less accurate fracture patterns, showing a more challenging classification scenario. This is in a similar vein to why the poor performance of the subgroup of hips is better understood in relation to an extremely low number of fractured samples in comparison to failure of the model, showing that anatomical complexity is a significant predictor of performance. There was a variation in performance depending on the radiographic view. This shows that the best performances on the multiple-view images imply that more contextual information aids in increasing classification, possibly due to a broader structural outlook. The lower performances on the oblique views imply that the model performed poorly at times when anatomical orientation was not standardized. The error analysis also

supported this trend by identifying false positives and false negatives in challenging image situations. One thing that particularly intrigued interest was the fact that very confident false negative predictions were made; this meant that the model was at times too reliant on local visual patterns and could not generalize anatomically. The findings imply that the effectiveness of local feature extraction could not be as high as needed in the context of fracture detection with high consistency in various radiographic situations.

The results obtained are in agreement with previous studies that have shown the usefulness of convolutional models in the classification of medical images but have also shown their limitations in situations where complex spatial reasoning may be necessary. However, the use of hybrid models that combine CNNs and Vision Transformers, where local feature extraction and global context modeling are combined, has been proposed as a solution to this problem, and this has been shown to be effective in situations where medical image classification, which involves multi-scale analysis, has to be conducted [15]. On the same note, multi-scale classification models that use transformers have shown that the integration of multi-scale information can lead to improved accuracy in the diagnosis in a complex image environment [16].

The significance of the combined form of convolutional and transformer-based representations can also be seen in the evidence given in the medical image segmentation study. It has been proven that the proposed hybrid multi-scale networks are capable of preserving the fine local details and global structural understanding, thereby achieving better performance in image segmentation [17]. In similar studies related to the proposed approach, it has also been shown that grouped multi-scale transformers and multi-resolution CNN-transformer networks are capable of better representing anatomically complex structures and achieving better performance in regions of interest with varying sizes and appearances [18, 19]. These results are in line with the results obtained in the current stage, in which the challenges were revealed the most in complicated regions and non-standard points of view. Further studies on the performance of contextual fusion and attention-based transformer models also confirm the above-stated meaning even more. It has been provided that hybrid contextual fusion strategies are to improve feature integration in the analysis of medical images by balancing local and global information [20]. Similarly, Swin Transformer models, based on multi-scale attention, have shown their potential in their ability to recognize hierarchical features in the image and their potential to perform better in complex image segmentation problems [21]. A mix of these studies would support the belief that the classification of fractures would be improved using architectures based on both sub-local and sub-global contextual dependencies.

The results of the study have some methodological and clinical implications. Modeling-wise, the results have implied that a standard CNN structure can be a strong baseline performance, but not applicable to all radiographic situations. The existence of variability in different regions of the body and in different viewing poses leads to the belief that more context-sensitive architectures will be required to achieve the robustness of fracture detection. This justifies the idea of hybrid solutions, which involve local and global feature learning. Clinically, the research demonstrates the necessity to consider AI systems beyond the general accuracy. A model can be found to work very well according to aggregate measures, but still fail in areas of anatomical complexity or non-standard image projections. This variability is particularly crucial in fracture detection, where abnormalities not detected could be of great importance. Thus, subgroup-level analysis and error analysis are important components of model assessment prior to clinical translation.

Certain shortcomings are to be taken into account. Only a single dataset was used to conduct the study, therefore, limiting the generalizability to another population and imaging environment. Besides this, the imbalance of classes and reduced sample sizes in some of the anatomical subgroups can also have contributed to subgroup-level performance. Segmentation annotations were available, but analysis was done based on classification and localization not pursued. Nonetheless, the research sets a solid CNN as a priority and outlines the subgroup and error analysis, which can assist in the further development of the model.

The next phase in continuing this framework should be the implementation and measurement of hybrid CNN-ViT models on bigger and more varied radiographic datasets. Segmentation supervision could also enhance explainability and assist models to pay more focus on the fracture-related areas. Moreover, calibration, external validation, and prospective clinical utility are the areas that should be considered in future research to investigate whether these models can guide the workflow of real-world fracture detection more reliably.

5. Conclusion

The right diagnosis of fracture in radiographic images involves the ability to identify the minor local abnormalities, as well as the anatomical context. The current results indicate that the ResNet-18 baseline was very effective in overall discrimination to classify fractures, and its ability to detect non-fractured cases was also very good. Nevertheless, lower performance on anatomically complex areas and non-standard radiographic images with the identified false-negative errors suggests that fracture identification is still difficult when contextual interpretation is needed. Error and subgroup analyses were able to further elaborate on the results of the overall test performance by revealing that the model behavior was different between body parts and imaging views. These findings emphasize the need to consider fracture detection systems both through aggregate measures and their stability in a clinically diverse environment. All in all, the results indicate the usefulness of deep learning in automated fracture classification and provide a solid CNN baseline that can be used in the future. They also provide an explicit explanation for the possibility of developing hybrid CNN-Vision Transformer models that can integrate local feature generation with a global contextual model to improve the performance of multi-scale medical image analysis.

References

1. Cauley, J.A., 2021. The global burden of fractures. *The Lancet Healthy Longevity*, *2*(9), pp.e535-e536.
2. Wong, R.M.Y., Wong, P.Y., Liu, C., Wong, H.W., Chung, Y.L., Chow, S.K.H., Law, S.W. and Cheung, W.H., 2022. The imminent risk of a fracture—existing worldwide data: a systematic review and meta-analysis. *Osteoporosis International*, *33*(12), pp.2453-2466.
3. Lalloo, R., Lucchesi, L.R., Bisignano, C., Castle, C.D., Dingels, Z.V., Fox, J.T., Hamilton, E.B., Liu, Z., Roberts, N.L., Sylte, D.O. and Alahdab, F., 2020. Epidemiology of facial fractures: incidence, prevalence and years lived with disability estimates from the Global Burden of Disease 2017 study. *Injury prevention*, *26*(Suppl 2), pp.i27-i35.
4. Abosadegh, M.M. and Rahman, S.A., 2018. Epidemiology and incidence of traumatic head injury associated with maxillofacial fractures: A global perspective. *Journal of international oral health*, *10*(2), pp.63-70.
5. Joshi, D. and Singh, T.P., 2020. A survey of fracture detection techniques in bone X-ray images. *Artificial Intelligence Review*, *53*(6), pp.4475-4517.
6. Löffler, M.T., Sollmann, N., Mei, K., Valentinitsch, A., Noël, P.B., Kirschke, J.S. and Baum, T., 2020. X-ray-based quantitative osteoporosis imaging at the spine. *Osteoporosis International*, *31*(2), pp.233-250.
7. Tang X. The role of artificial intelligence in medical imaging research. BJR| open. 2019 Nov 21;2(1):20190031.
8. Kuo, R.Y., Harrison, C., Curran, T.A., Jones, B., Freethy, A., Cussons, D., Stewart, M., Collins, G.S. and Furniss, D., 2022. Artificial intelligence in fracture detection: a systematic review and meta-analysis. *Radiology*, *304*(1), pp.50-62.
9. Bhatnagar, A., Kekatpure, A.L., Velagala, V.R. and Kekatpure, A., 2024. A review on the use of artificial intelligence in fracture detection. *Cureus*, *16*(4).
10. Thian, Y.L., Li, Y., Jagmohan, P., Sia, D., Chan, V.E.Y. and Tan, R.T., 2019. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*, *1*(1), p.e180001.
11. Kim DH, Mackinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clinical radiology. 2018 May 1;73(5):439-45.
12. Kameswari, C.S., Kavitha, J., Reddy, T.S., Chinthaguntla, B., Jagatheesaperumal, S.K., Gaftandzhieva, S. and Doneva, R., 2023. An overview of vision transformers for image processing: a survey. *International Journal of Advanced Computer Science and Applications*, *14*(8).
13. Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N. and Machino, H., 2024. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. *Journal of Medical Systems*, *48*(1), p.84.
14. Abedeen, I., Rahman, M.A., Zohra Prottyasha, F., Ahmed, T., Mohmud Chowdhury, T. and Shatabda, S., 2023. FracAtlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs (Version 6). figshare.
15. Chibuike, O. and Yang, X., 2024. Convolutional neural network–vision transformer architecture with gated control mechanism and multi-scale fusion for enhanced pulmonary disease classification. *Diagnostics*, *14*(24), p.2790.
16. Hu, J., Xiang, Y., Lin, Y., Du, J., Zhang, H. and Liu, H., 2025, February. Multi-scale transformer

architecture for accurate medical image classification. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence* (pp. 409-414).

17. Fu, B., Peng, Y., He, J., Tian, C., Sun, X. and Wang, R., 2024. HmsU-Net: A hybrid multi-scale U-net based on a CNN and transformer for medical image segmentation. *Computers in Biology and Medicine*, *170*, p.108013.
18. Ji, Z., Chen, Z. and Ma, X., 2025. Grouped multi-scale vision transformer for medical image segmentation. *Scientific Reports*, *15*(1), p.11122.
19. Zhu, S., Li, Y., Dai, X., Mao, T., Wei, L. and Yan, Y., 2025. A multi-resolution hybrid cnn-transformer network with scale-guided attention for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*.
20. Bao, H., Zhu, Y. and Li, Q., 2023. Hybrid-scale contextual fusion network for medical image segmentation. *Computers in Biology and Medicine*, *152*, p.106439.
21. Mirab Golkhatmi, B., Houshmand, M. and Hosseini, S.A., 2025. A multi-scale attention-based Swin transformer model for medical images segmentation. *Scientific Reports*, *15*(1), p.38893.