International Journal of Robotics and Control Systems
Vol. 6, No. 2, 2026, pp. 1024-1040
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

PVT-FractureNet: A Pyramid Vision Transformer Model for Radiographic Bone Fracture Classification

Yuri Pamungkas [a,1,*], Abdul Karim [b,2], Muhammad Nur Afnan Uda [c,3], Uda Hashim [d,4]

[a] Department of Medical Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia
[b] Department of Artificial Intelligence Convergence, Hallym University, Chuncheon, 24252, Republic of Korea
[c] Department of Electronic Engineering (Computer), Universiti Malaysia Sabah, Kinabalu, 88400, Malaysia
[d] Department of Electrical and Electronics Engineering, Universiti Malaysia Sabah, Kinabalu, 88400, Malaysia
[1*] yuri@its.ac.id; [2] abdullkarim@korea.ac.kr; [3] nurafnan@ums.edu.my; [4] uda@ums.edu.my
* Corresponding Author

ARTICLE INFO

Article history
Received October 26, 2025
Revised March 20, 2026
Accepted April 05, 2026

Keywords
Pyramid Vision Transformer (PVT);
Bone Fracture Classification;
Radiographic Imaging;
Explainable Artificial Intelligence (XAI);
Deep Learning in Medical Diagnosis

ABSTRACT

Bone fracture classification from radiographic images plays a critical role in orthopedic diagnosis and treatment planning, yet manual interpretation remains time-consuming and prone to inter-observer variability. Traditional CNN approaches often have difficulty capturing fine-grained fracture characteristics as well as broader skeletal structural patterns, limiting their ability to accurately classify multiple fracture types. To overcome this limitation, this study proposes PVT-FractureNet, a deep learning framework based on the Pyramid Vision Transformer (PVT) architecture designed for holistic radiographic-based categorization of bone fractures. The primary contribution offered by this study is the development of a transformer-driven hierarchical architecture that proficiently combines multi-level feature representations with explainable attention visualization to enhance diagnostic accuracy and interpretability. The model was trained and evaluated on a Kaggle dataset comprising ten fracture categories, including avulsion, comminuted, greenstick, hairline, impacted, longitudinal, oblique, pathological, spiral, and fracture-dislocation. Preprocessing steps included normalization, resizing, and data augmentation, followed by feature extraction and classification using multi-head self-attention and spatial-reduction mechanisms. Experimental results demonstrated that PVT-FractureNet achieved an average accuracy of 89.9%, specificity of 91.1%, and AUC of 0.811, with the highest performance observed in Greenstick and Fracture-Dislocation classes (AUC > 0.90). Grad-CAM and Score-CAM visualizations further showed that the model precisely pinpointed fracture regions that align with clinically meaningful anatomical cues. In conclusion, PVT-FractureNet demonstrates robust generalization capability, clear interpretability, and dependable diagnostic performance, establishing it as a promising framework for automated transformer-based bone fracture classification in clinical radiology.

1. Introduction

Bone fractures represent among the leading most prevalent musculoskeletal injuries worldwide [1], often arising due to trauma, osteoporosis, or disease-related abnormalities that compromise bone

strength. Precise and prompt identification plays a critical role for determining appropriate treatment strategies and preventing long-term complications such as deformities or impaired mobility [2]. Nevertheless, conventional identification of bone fractures through radiographic modalities is predominantly reliant on the specialized knowledge and accumulated clinical skills of radiologists or orthopedic specialists [3]. This manual interpretation is prone to subjectivity, inter-observer variability, and diagnostic fatigue, particularly when differentiating between subtle fracture types such as hairline, greenstick, or longitudinal fractures [4]. Consequently, there is an increasing need for automated, reliable, and interpretable computer-aided diagnostic systems to assist clinicians in classifying bone fracture types accurately [5].

Contemporary breakthroughs within deep learning technologies have markedly enhanced medical diagnostic practices that rely on imaging modalities, with CNNs being extensively utilized for radiographic image analysis [6]-[9]. Nevertheless, CNN-based architectures often struggle to capture both fine-grained local features and long-range global dependencies within complex bone structures [10]. This limitation affects their ability to distinguish between morphologically similar fracture types, such as spiral versus oblique fractures or comminuted versus pathological fractures. To overcome these challenges, transformer-based models (originally developed for NLP) have emerged as powerful alternatives in computer vision [11]-[13]. Among them, Vision Transformers (ViT) and their hierarchical variants, such as the Pyramid Vision Transformer (PVT), have shown remarkable potential by combining local feature representation with global contextual understanding through multi-head self-attention mechanisms [14]-[16].

Despite these advancements, few studies have specifically leveraged transformer-based architectures for fine-grained bone fracture classification using radiographic data [17]. Existing approaches often focus on binary classification (fractured vs. non-fractured) or limited categories, lacking generalization across diverse fracture morphologies. Moreover, the integration of hierarchical vision transformers, which balance computational efficiency and feature resolution, remains underexplored in orthopedic imaging [18]-[20]. To address these gaps, this research introduces PVT-FractureNet, a deep learning framework based on the Pyramid Vision Transformer architecture, designed for multi-class classification of ten distinct types of bone fractures (namely oblique, pathological, spiral, greenstick, hairline, impacted, longitudinal, avulsion, comminuted, and fracture dislocation) using radiographic images. The model leverages hierarchical attention mechanisms to extract both global bone structure and localized fracture patterns while maintaining computational efficiency suitable for clinical deployment.

The contribution of this research is the development of a hierarchical transformer-based model that enhances feature representation across multiple spatial scales, achieving robust classification performance across diverse fracture types. In addition, PVT-FractureNet incorporates explainability analysis using Grad-CAM visualization to ensure interpretability and clinical trustworthiness. By combining advanced transformer-based feature learning with explainable AI techniques, this study provides a novel and interpretable diagnostic framework that can support radiologists in identifying and differentiating complex bone fractures with higher accuracy and reliability.

2. Related Works

Several studies have explored AI techniques for the detection and classification of bone fractures from radiographic images, primarily using traditional machine learning and CNN architectures. Early research by Bagaria et al. [21] and Ahmed et al. [22] applied conventional classifiers such as SVM and Random Forest combined with handcrafted features like wavelet, GLCM, and Canny edge detection. These models achieved accuracies of around 92–94%, demonstrating the feasibility of automated fracture detection but highlighting the limitations of manual feature extraction, which often fails to capture complex structural variations among different fracture types. Kim et al. [23] introduced a non-imaging approach using Stacked Autoencoders with vibration-based data, achieving up to 93.3% accuracy and providing a cost-effective, non-invasive diagnostic alternative. Despite promising

results, such approaches lacked the scalability and spatial awareness required for diverse radiographic conditions.

Deep learning models have since advanced the field by leveraging CNN-based architectures capable of automatically learning image features from large-scale datasets. Franco et al. [24] employed Mask R-CNN to detect pediatric appendicular fractures with sensitivity up to 91%, while Gaast et al. [25] utilized GoogleNet and ResNet for tibial plateau fractures, achieving strong detection (93% sensitivity) but weak classification performance due to inconsistent labeling. More recent works have incorporated attention mechanisms and hybrid architectures to enhance feature representation. Zou et al. [26] improved YOLOv7 by adding a specialized attention-driven module and a refined and more robust loss formulation (EIoU), achieving mAP of 86.2% on the FracAtlas dataset. Similarly, Sutradhar et al. [27] combined YOLOv8 with VertNet-10 and explainable AI (XAI), yielding a classification performance reaching 99.55% for cervical spine fractures, while Tsai et al. [28] developed DEANet with EfficientNet-B0 and attention blocks to achieve over 92% accuracy for pelvic fractures. These findings underline the growing relevance of attention-based architectures in improving model interpretability and diagnostic precision.

Recent developments have pushed the boundaries further by integrating hybrid deep learning and transformer-based methods for generalized fracture recognition. Alwzwazy et al. [29] introduced FracNet, combining Xception, MobileNet, and EfficientNet with attention and feature fusion, reaching perfect accuracy across multiple datasets, while Wei et al. [30] proposed YOLOv11 with multi-task learning and Grad-CAM explainability for musculoskeletal X-rays, achieving an accuracy of 95%. Likewise, Elsheikh et al. [31] optimized DenseNet201 using metaheuristic algorithms to achieve 97.85% accuracy, and Kanagaraj et al. [32] applied a Capsule Graph Neural Network with Gazelle Optimization, reaching 99.66% accuracy and outperforming existing CNN-based models. Despite these advances, few studies have explored the hierarchical transformer architectures, such as the Pyramid Vision Transformer (PVT), that can effectively capture multi-scale spatial dependencies crucial for differentiating subtle fracture morphologies. This research builds upon these foundations by proposing PVT-FractureNet, a hierarchical transformer-based model designed to provide both high classification accuracy and strong clinical interpretability across ten distinct bone fracture types.

3. Method

3.1. Bone Fracture Dataset

The dataset employed throughout this study was obtained from an openly available repository on Kaggle, comprising ten specific categories of osseous fracture types, namely avulsion, comminuted, fracture-dislocation, greenstick, hairline, impacted, longitudinal, oblique, pathological, and spiral fractures [33]. These categories were selected to represent a broad spectrum of clinically relevant fracture morphologies that vary in orientation, displacement, and bone integrity. Each image in the dataset was collected from anonymized radiographic sources, preprocessed to ensure consistent quality and resolution, and standardized in grayscale format. The dataset was split into training and validation subsets to enhance model learning and to enable comprehensive performance evaluation. Table 1 illustrates that the distribution of images across classes is relatively balanced, with the highest number of samples appearing in fracture-dislocation and comminuted fracture categories, while the lowest counts were observed in longitudinal and spiral fractures.

Overall, the dataset comprises more than a thousand annotated X-ray samples, with approximately 70% used for training, 20% for validation, and 10% reserved for testing. This allocation facilitates the model's ability to generalize across varied fracture categories while reducing issues related to class distribution disparities. Every image was subjected to a series of preprocessing procedures, such as denoising operations, intensity-scaling normalization, and dimensional resizing to match the Pyramid Vision Transformer input dimensions. The dataset diversity (covering both complete and partial fracture patterns) enables PVT-FractureNet to learn subtle morphological variations such as fine fissures in hairline fractures and complex fragmentations in comminuted

fractures. This comprehensive dataset design provides a robust foundation for developing an accurate and interpretable AI-based framework for automated bone fracture classification in radiographic imaging.

Table 1. Distribution of fracture types in the bone fracture dataset

No	Fracture Type	Total	Train	Validation
1	Avulsion Fracture	123	98	25
2	Comminuted Fracture	148	118	30
3	Fracture-Dislocation	156	125	31
4	Greenstick Fracture	122	98	24
5	Hairline Fracture	111	89	22
6	Impacted Fracture	84	67	17
7	Longitudinal Fracture	80	64	16
8	Oblique Fracture	85	68	17
9	Pathological Fracture	134	107	27
10	Spiral Fracture	86	69	17
	Total	**1129**	**903**	**226**

3.2. Data Preprocessing

Before training the PVT-FractureNet model, a comprehensive data preprocessing pipeline was implemented to ensure image consistency, reduce noise, and improve the fidelity of the feature representations derived by the network [34]. The raw radiographic images obtained from the Kaggle repository exhibited variations in resolution, contrast, and illumination, which could potentially affect the operational effectiveness of deep learning–based systems. To address this, all images were first standardized to a uniform grayscale format to eliminate color-related redundancy, followed by normalization to rescale pixel intensity values within a [0,1] range [35]. This normalization step helps stabilize the gradient updates during training and accelerates model convergence [36]. Additionally, every image was rescaled to a resolution of 224 × 224 pixels to conform to the input specifications mandated by the Pyramid Vision Transformer architecture, ensuring a consistent input dimension across all samples (Fig. 1) [37].

To improve model robustness and prevent overfitting, several augmentation-driven data transformation methods were applied to the training set [38]. These included stochastic rotational adjustments within a ±15° range, bidirectional flipping operations along both horizontal and vertical axes, zoom scaling, and small contrast adjustments to simulate real-world variability in X-ray acquisition conditions. The augmentation process effectively expanded the dataset's diversity, allowing PVT-FractureNet to generalize better across different fracture morphologies and patient anatomies. Moreover, Gaussian noise and slight blurring were occasionally introduced to mimic low-quality radiographs commonly encountered in clinical practice [39]. This approach strengthens the model's robustness against noise disturbances and imaging-induced artifacts while maintaining anatomical integrity [40].

Finally, a train–validation split was performed using an approximate ratio of 70:20:10 for training, validation, and evaluation phases, respectively, as shown in the dataset distribution Table 1. The data were carefully shuffled to prevent class bias and ensure that all fracture types (including avulsion, comminuted, and fracture-dislocation) were proportionally represented in each subset. During preprocessing, image labels were transformed into a categorical one-hot representation tailored for handling multi-class classification scenarios. Collectively, these preprocessing procedures ensured that the dataset provided clean, balanced, and well-structured input for the hierarchical attention mechanisms of PVT-FractureNet, enabling it to efficiently learn fine-grained fracture characteristics while also interpreting the broader anatomical context of the bone.

3.3. Pyramid Vision Transformer Model

The Pyramid Vision Transformer (PVT) functions as the core architectural backbone for the proposed PVT-FractureNet framework, developed to effectively extract localized fracture characteristics as well as the overall bone structure within radiographic imagery. In contrast to

conventional CNN models that depend on static receptive field configurations, the PVT employs a hierarchical multi-stage transformer framework that processes images at multiple resolutions [41]. When provided with an input image $I \in R^{H\times W\times 3}$, the image is initially segmented into distinct, non-overlapping patches of dimensions $P \times P$, which are then flattened and linearly projected into a sequence of embeddings.

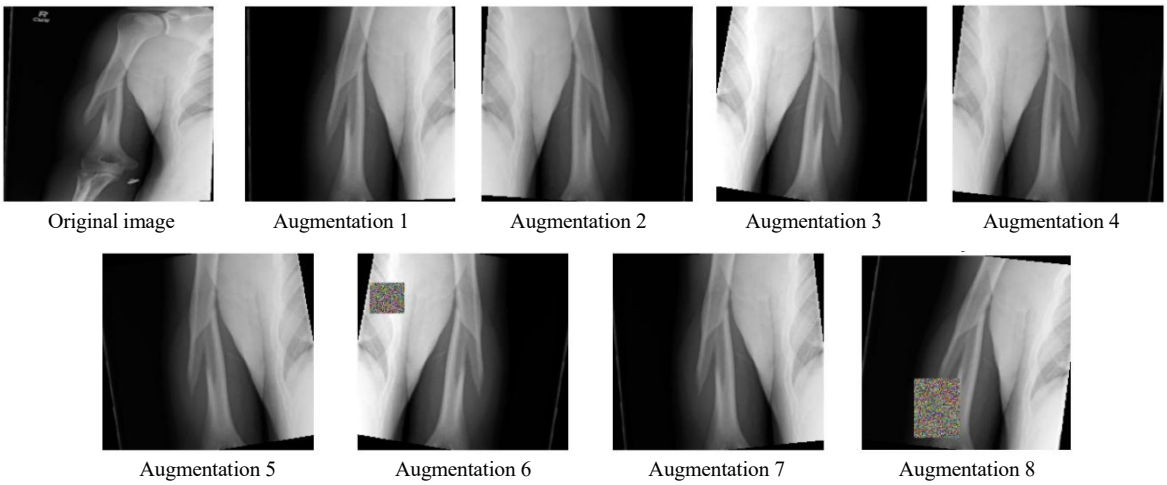

Fig. 1. Preprocessing data samples

$$X_0 = [x_1, x_2, \dots, x_N], \quad x_i = W_p \cdot \text{Flatten}(P_i) + b_p \tag{1}$$

where $W_p \in R^{(P^2\cdot 3)\times D}$ is the patch embedding matrix, b_p is the bias term, and $N = \frac{HW}{P^2}$ is the total number of patches. These embeddings are then passed through multiple hierarchical stages, each consisting of Transformer Encoder Blocks with progressively reduced spatial dimensions and increased feature depth, forming a pyramid-like representation that preserves both fine-grained and high-level semantic features.

Each encoder block applies the multi-head self-attention (MHSA) mechanism to model long-range dependencies among patches. The attention function for a single head is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where $Q, K, V \in R^{N\times d_k}$ represent the query, key, and value matrices derived from linear projections of the input sequence X. For h attention heads, the outputs are concatenated and projected as:

$$\text{MHSA}(X) = \text{Concat}(head_1, head_2, \dots, head_h)W_O \tag{3}$$

with $W_O \in R^{(h\cdot d_k)\times D}$. This mechanism enables the model to attend to multiple subspaces simultaneously, capturing spatial correlations between distant bone regions, essential for detecting subtle fracture lines.

To improve efficiency on high-resolution X-rays, PVT replaces the global self-attention used in standard ViT with a Spatial-Reduction Attention (SRA) mechanism that reduces key and value dimensions by a ratio r, formulated as:

$$\text{SRA}(Q, K, V) = \text{Softmax}\left(\frac{Q(K_r)^T}{\sqrt{d_k}}\right)V_r \tag{4}$$

where K_r and V_r are obtained through a spatial reduction operation, typically a strided convolution with factor r. This allows the attention computation to scale linearly with image size $O\left(\frac{HW}{r^2}\right)$,

maintaining high performance with significantly lower computational cost, critical for medical datasets with large input resolutions.

The feature maps from the final pyramid stage are passed through a Global Average Pooling (GAP) layer and a fully connected classifier to predict the fracture type among ten output classes [42]. The model is optimized using cross-entropy loss, defined as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right) \tag{5}$$

where $y_{i,c}$ is the true label and $\hat{y}_{i,c}$ is the predicted probability for class c. The overall PVT-FractureNet framework thus combines hierarchical feature extraction, spatially efficient attention, and robust classification to achieve superior performance in multi-class bone fracture identification, offering a balanced trade-off between computational efficiency, accuracy, and interpretability in clinical radiographic analysis.

The proposed PVT-FractureNet architecture is designed to integrate the hierarchical feature extraction capabilities of the Pyramid Vision Transformer (PVT) with a dual-branch learning strategy for robust and discriminative fracture classification. As shown in the Fig. 2, the model is initiated using radiographic input data which are initially segmented into mutually exclusive image patches. Every patch undergoes a linear projection that converts it into an embedding vector and then passed through multiple transformer encoder layers [43]. These encoders operate in a pyramidal hierarchy, where spatial dimensions are gradually reduced while feature depth increases, generating multiscale feature maps that represent both local and global bone structures. This hierarchical process allows the network to simultaneously capture fine fracture lines and contextual anatomical patterns that are often critical in distinguishing between subtle fracture types.

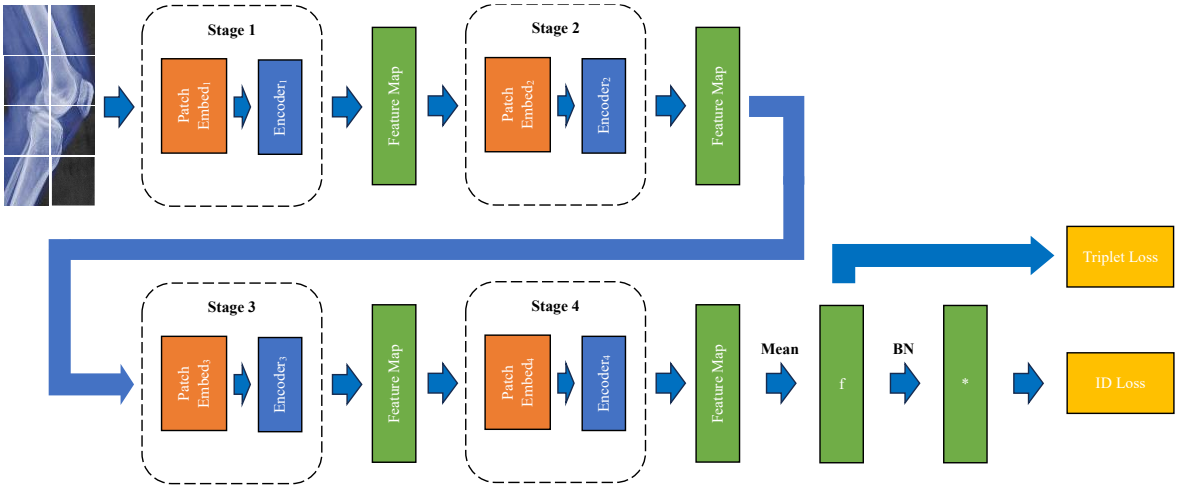

Fig. 2. The proposed PVT-FractureNet model

To enhance the discriminative learning ability of the model, PVT-FractureNet employs a two-stream architecture. The first stream (upper path) focuses on global structural representation by encoding visual patches through successive encoder blocks to produce deep semantic features. These features are used to compute a Triplet Loss, which enforces inter-class separability and intra-class compactness in the latent feature space. Mathematically, Triplet Loss is defined as

$$\mathcal{L}_{triplet} = \max\left(0, \left|f_a - f_p\right|_2^2 - \left|f_a - f_n\right|_2^2 + \alpha\right) \tag{6}$$

where f_a, f_p, and f_n represent the embedding vectors of an anchor, positive, and negative sample, respectively, and α is a predefined margin. This loss encourages the model to minimize the distance

between samples of the same fracture type while maximizing the distance between different types, leading to more discriminative representations across the ten fracture categories.

The second stream (lower path) processes the same input image through additional encoding layers and a global feature extraction function f, followed by a fully connected layer for classification. The network is trained concurrently using an Identification (ID) Loss, formulated as a standard cross-entropy loss:

$$\mathcal{L}_{ID} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{7}$$

where y_i and $\hat{y}_i$ denote the ground truth and predicted probability for the i-th class, respectively. This loss optimizes categorical prediction accuracy across all fracture classes, such as oblique, greenstick, pathological, and comminuted fractures.

Finally, the combined optimization objective of PVT-FractureNet is expressed as

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{triplet} + \lambda_2 \mathcal{L}_{\mathcal{ID}} \tag{8}$$

where λ_1 and λ_2 are weighting factors that balance metric learning and classification objectives. This hybrid learning approach allows the model to simultaneously learn discriminative embeddings and accurate classification boundaries. The multiscale feature fusion and dual-loss supervision enable PVT-FractureNet to outperform conventional CNN models by effectively identifying complex bone fracture patterns, achieving both high accuracy and explainable decision-making in radiographic bone fracture classification.

3.4. Metrics Evaluation and Explainability

The overall performance exhibited by the constructed PVT-FractureNet model was evaluated using an array of widely recognized assessment metrics typically applied in medical image classification tasks, comprising accuracy, precision, recall, F1-score, and the AUC. Accuracy denotes the total ratio of fracture images accurately identified by the model, whereas precision reflects the ratio of correctly identified positive cases compared with all instances predicted as positive [44]. Recall, also referred to as sensitivity, measures the degree to which the model effectively identifies authentic fracture events [45], while the F1-score offers a harmonic balance between precision and recall, proving particularly useful in situations involving class imbalance [46]. The AUC metric assesses the model's capacity to distinguish among varied fracture classes under different decision-threshold conditions [47]. Together, these metrics collectively provide an extensive viewpoint on the model's ability to correctly identify and differentiate among the ten fracture classes (such as comminuted, pathological, spiral, and hairline fractures) while maintaining consistency across all categories.

In addition to numerical performance, explainability was integrated as a core component of model validation to ensure clinical reliability [48]. The Grad-CAM method was applied to depict the particular areas within each X-ray image that exerted the greatest impact on the model's predictive output [49]. This visualization produces heatmaps that emphasize the skeletal regions receiving concentrated attention from the model when classifying a particular fracture type. For instance, in cases of comminuted or greenstick fractures, the heatmaps revealed that PVT-FractureNet accurately concentrated on disrupted cortical lines or irregular bone contours, aligning well with expert radiological interpretation. This explainable framework enhances clinical trust by rendering the model's reasoning workflow more accessible and understandable to clinical professionals [50]. Rather than functioning as a "black box," PVT-FractureNet demonstrates how its attention mechanisms prioritize relevant anatomical structures and fracture patterns. The combination of high quantitative performance and clear visual explainability demonstrates that the model attains high diagnostic precision while simultaneously delivering outputs that are informative and readily interpretable, supporting radiologists in real-world clinical environments [51].

4. Results and Discussion

This section introduces and elaborates on the experimental outcomes derived from deploying the proposed PVT-FractureNet model for multi-class bone fracture categorization based on radiographic imaging data. The assessment concentrated on ten clinically relevant fracture categories, namely avulsion, comminuted, fracture-dislocation, greenstick-type, fine hairline-type, compression-impacted, longitudinal-pattern, oblique-pattern, disease-associated pathological, and spiral-pattern fractures. The model's predictive capability was evaluated through several performance indicators, including accuracy, precision, recall, F1-score, and the AUC, as well as through visualization tools such as the confusion matrix and Grad-CAM heatmaps. These evaluations were intended to showcase not merely the model's classification performance but also its dependability and interpretative clarity within the context of clinical decision-making systems.

The results highlight the capability of PVT-FractureNet to effectively distinguish between fracture classes with diverse morphological patterns, even in cases where visual differences are subtle or overlapping. The confusion matrix, presented in the Fig. 3, provides a detailed view of the model's classification performance across all categories. It reveals how the hierarchical attention mechanism of the Pyramid Vision Transformer contributes to capturing both local fracture details and global bone structures, resulting in improved differentiation between complex fracture types. The Fig. 3 illustrates the confusion matrix of the PVT model, serving as the foundation for the discussion of its diagnostic strengths and limitations.

The confusion matrix (Fig. 3) provides a detailed overview of the classification performance of the PVT-FractureNet model across the ten fracture categories. Each row within the matrix corresponds to the ground-truth labels, which denote the genuine fracture categories, whereas each column indicates the predicted class outputs generated by the model. The diagonal components denote the count of samples accurately categorized for their respective classes, while the off-diagonal components reflect instances of misclassification, representing cases where the model incorrectly identified one fracture type as another. From the matrix, it is evident that PVT-FractureNet demonstrates strong overall performance, with several classes showing high recognition accuracy. The model correctly identified the majority of Fracture-Dislocation cases (24 correctly classified), reflecting its capability to capture distinctive radiographic characteristics such as bone misalignment and joint displacement. Similarly, Pathological Fractures (15 correct) and Avulsion Fractures (12 correct) were classified with good accuracy, indicating that the model effectively learned to recognize cortical separation and bone density variations typical of these categories.

However, a few classes showed moderate confusion, especially among fractures with visually subtle or overlapping features. For example, Greenstick Fractures (13 correct) and Hairline Fractures (12 correct) were occasionally misclassified as Avulsion or Comminuted Fractures, likely because their thin or incomplete fracture lines resemble minor cortical irregularities seen in other types. Similarly, Comminuted Fractures (11 correct) were sometimes confused with Fracture-Dislocation and Pathological Fractures, as all three exhibit complex multi-fragmented patterns in X-rays. These misclassifications can be attributed to the inherent difficulty of distinguishing fine morphological differences, especially in low-contrast or overlapping bone regions.

The performance metrics of the PVT-FractureNet model, as summarized in Table 2, demonstrate the model's capability to distinguish among various categories of bone fractures from radiographic images with strong consistency across key evaluation indicators. Overall, the model achieved an average accuracy of 89.9%, precision of 55%, recall of 47.5%, specificity of 91.1%, F1-score of 46.5%, and an AUC of 81.1%. These results indicate that the PVT architecture provides robust feature extraction and effective classification performance despite the morphological similarities among several fracture types. The high overall specificity and AUC values imply that the model reliably differentiates positive instances from negative ones, demonstrating strong discrimination capability and reliability in medical image interpretation [52]. In class-wise performance, Greenstick Fracture and Oblique Fracture attained the top accuracy performance levels (94% and 95%, respectively),

reflecting the model's ability to capture both localized and global contextual features of incomplete and angled fracture lines. The Fracture-Dislocation class achieved the highest recall (77%) and a relatively high F1-score (59%), suggesting that the model effectively recognized the majority of actual cases for this fracture category while preserving an equilibrium between sensitivity and precision. Conversely, Spiral and Longitudinal Fractures demonstrated reduced recall performance (29% and 25%), likely due to their visual similarity to other fracture categories and the relatively small number of samples available in the dataset. These limitations emphasize the importance of dataset balancing and data augmentation to improve performance in underrepresented classes [53].

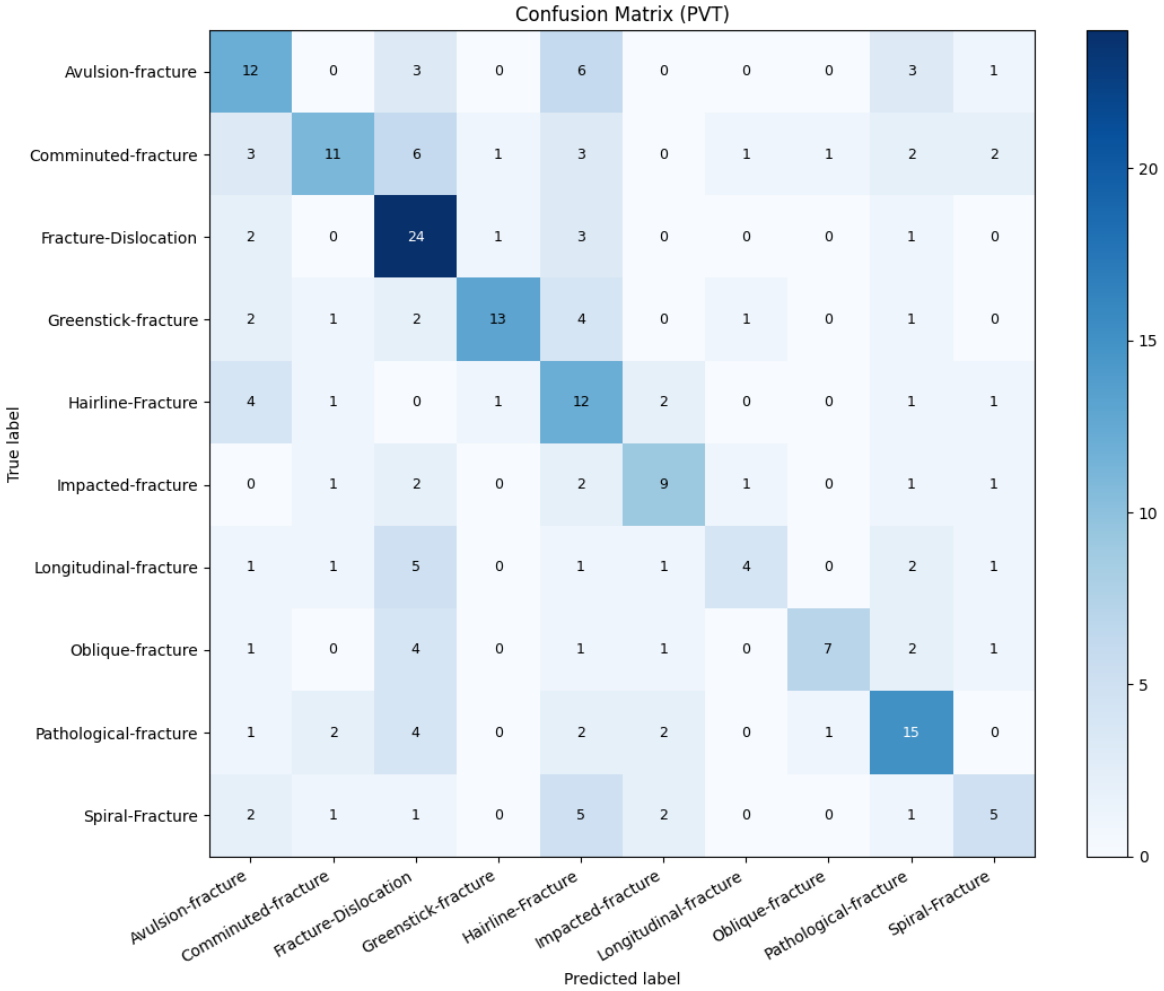

Fig. 3. Confusion matrix

Table 2. Performance metrics per fracture type using the PVT model

No	Fracture Type	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	AUC (%)
1	Avulsion Fracture	87	43	48	92	45	75
2	Comminuted Fracture	88	61	37	96	46	82
3	Fracture-Dislocation	85	47	77	86	59	91
4	Greenstick Fracture	94	81	54	86	65	92
5	Hairline Fracture	84	31	55	87	39	80
6	Impacted Fracture	93	53	53	89	53	90
7	Longitudinal Fracture	93	57	25	96	35	73
8	Oblique Fracture	95	57	41	89	35	72
9	Pathological Fracture	88	78	56	93	54	82
10	Spiral Fracture	92	42	29	97	34	74
	Average	**89.9**	**55**	**47.5**	**91.1**	**46.5**	**81.1**

The precision and specificity metrics further reveal that PVT-FractureNet performs well in reducing false positives. For instance, the Pathological Fracture class recorded a precision of 78% and specificity of 93%, suggesting that the model correctly identified cases with disease-related cortical destruction with minimal misclassification. Similarly, the Comminuted Fracture class achieved a precision of 61% and high specificity of 96%, indicating that the model effectively recognized complex multi-fragmented patterns in radiographs. However, some classes, such as Hairline Fracture, showed lower precision (31%), which reflects the inherent difficulty in detecting fine, low-contrast fracture lines even for advanced AI models.

The overall findings confirm that the hierarchical attention mechanism of the Pyramid Vision Transformer architecture enables the model to learn multi-scale contextual relationships within bone structures. This allows PVT-FractureNet to generalize across diverse fracture morphologies more effectively than traditional CNN-based models. The strong AUC value of 81.1% across all classes supports the model's capability to balance sensitivity and specificity, a crucial aspect for medical diagnostic systems. In summary, the results demonstrate that PVT-FractureNet provides a promising framework for automated, accurate, and interpretable bone fracture classification in radiographic imaging, paving the way for clinical decision-support tools that can aid radiologists in early diagnosis and triage [54].

The ROC curves in the Fig. 4 above illustrate the class-wise discriminative performance of the PVT-FractureNet model across ten fracture categories. Each curve illustrates the balance between the sensitivity and the false positive rate (1 – specificity) corresponding to an individual fracture category. The model attained a macro-averaged AUC score of 0.811, reflecting robust generalization ability and dependable separability between classes across all fracture types. An AUC value above 0.8 demonstrates that PVT-FractureNet is effective in distinguishing between positive and negative cases in radiographic fracture detection, a crucial indicator for clinical decision support systems.

Among the ten classes, the Greenstick Fracture category attained the top AUC score of 0.920, implying that the model proficiently identifies subtle cortical disruptions and partial fracture trajectories typically associated with this fracture type. Similarly, Fracture-Dislocation (AUC = 0.908) and Impacted Fracture (AUC = 0.900) demonstrated excellent discriminative ability, indicating the model's capability to detect unique structural shifts and superimposed bone configurations. These results underscore the benefit offered by the hierarchical attention design of the Pyramid Vision Transformer, which learns both local and global spatial dependencies [55], crucial for interpreting complex radiographic textures and bone alignment abnormalities.

Moderate AUC values were observed for Comminuted and Pathological Fractures (both 0.819), showing that the model maintained consistent sensitivity and specificity even for classes with heterogeneous visual patterns. However, lower AUC scores were obtained for Longitudinal (0.732), Oblique (0.724), and Spiral Fractures (0.741), which are morphologically similar and often exhibit overlapping characteristics in X-ray images. These results suggest that while PVT-FractureNet effectively captures key visual cues, certain fracture categories with subtle or ambiguous radiographic features remain challenging to differentiate. Increasing sample diversity and augmenting rotation-based transformations could further enhance model discrimination for these classes [56].

Overall, the ROC analysis reinforces that PVT-FractureNet achieves strong class-wise differentiation and robust diagnostic accuracy. The macro-AUC score of 0.811 validates the model's reliability for multi-class fracture classification and its potential as a clinically interpretable AI-assisted diagnostic tool. By combining hierarchical feature representation and self-attention mechanisms, PVT-FractureNet successfully identifies intricate fracture morphologies across a wide spectrum of bone types, providing a promising foundation for real-world clinical integration in radiographic bone fracture assessment.

To assess the interpretability and reliability of the proposed PVT-FractureNet model, a qualitative explainability analysis was performed using Grad-CAM and Score-CAM visualizations for representative fracture cases (Fig. 5). These explainability maps highlight the specific regions within

each radiograph that most strongly influenced the model's predictions [57]. By correlating these attention maps with radiographic anatomy, it becomes possible to verify whether the model focuses on clinically meaningful areas (such as cortical discontinuities, trabecular distortions, or misalignment zones) consistent with radiological interpretation. The results across multiple fracture classes demonstrate that PVT-FractureNet successfully learns to localize relevant fracture regions while maintaining high prediction confidence.

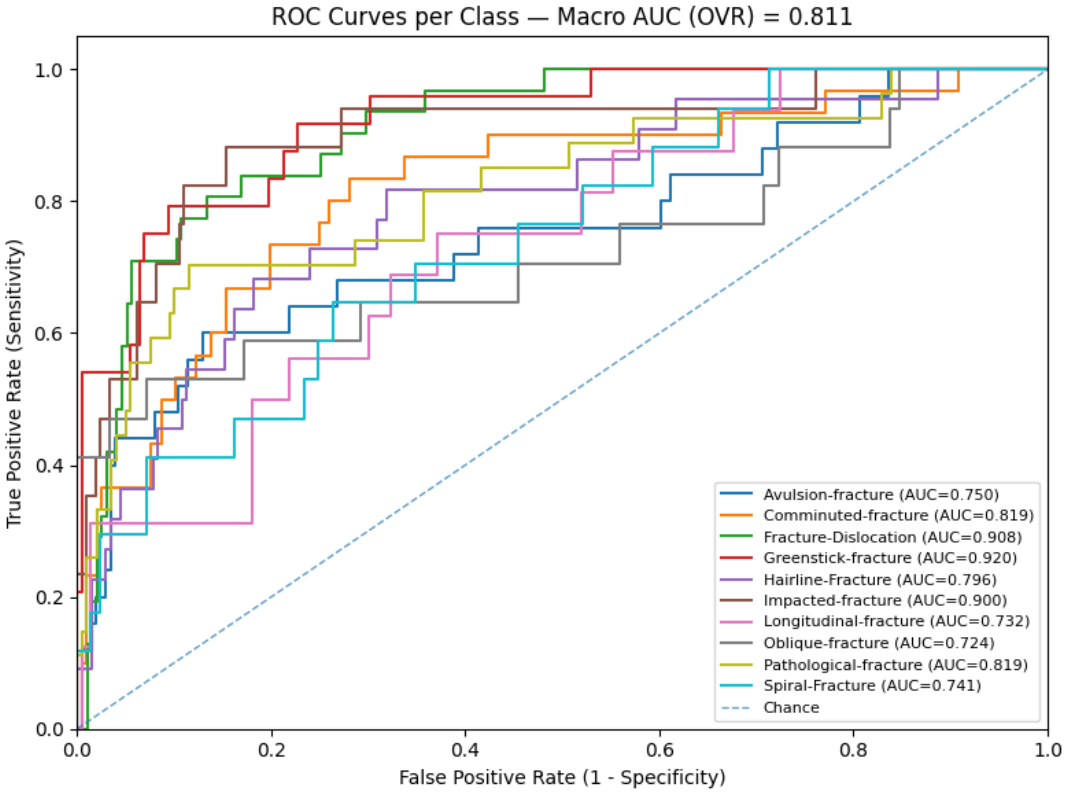

Fig. 4. ROC curves per class – macro AUC

In the Fracture-Dislocation case, the model achieved a prediction confidence of 92.39%, correctly identifying the dislocation site with a focused attention map concentrated around the joint area. Both Grad-CAM and Score-CAM visualizations reveal that the model accurately emphasizes the region of bone displacement and articular misalignment, which are key diagnostic indicators for this fracture type. The surrounding bone structures are less activated, reflecting the PVT's hierarchical attention mechanism that prioritizes discriminative local features while minimizing background noise [58]. This suggests that PVT-FractureNet effectively captures spatial dependencies between the dislocated bone segments and the adjacent cortical contours, leading to a robust classification outcome.

Similarly, for the Impacted Fracture class, the model demonstrated a strong prediction confidence of 91.79%. The attention maps show high activation in the central pelvic region, precisely at the location of bone impaction. Grad-CAM highlights compact localized zones of activation [59], while Score-CAM presents a broader [60], yet anatomically consistent heat distribution, confirming that the model not only detects the fracture interface but also considers the contextual bone compression patterns characteristic of impacted fractures. The low misclassification probabilities for other classes (each below 1.1%) further indicate that PVT-FractureNet is able to effectively differentiate impact-related structural changes from other fracture morphologies such as avulsion or comminuted fractures.

In the case of the Spiral Fracture, the model reached a confidence score of 87.08%, correctly identifying the helical fracture pattern along the humeral shaft. The Grad-CAM visualization distinctly highlights the midshaft area where the spiral break occurs, while the Score-CAM output extends the focus along the fracture trajectory, indicating that the model perceives the continuous torsional pattern

associated with this fracture type. Despite the complex curvature and partial overlap with soft tissue regions, the model maintained consistent activation over clinically relevant areas. The probability distribution shows minimal confusion with other fracture types such as pathological (2.55%) or avulsion (1.67%), confirming high classification specificity.

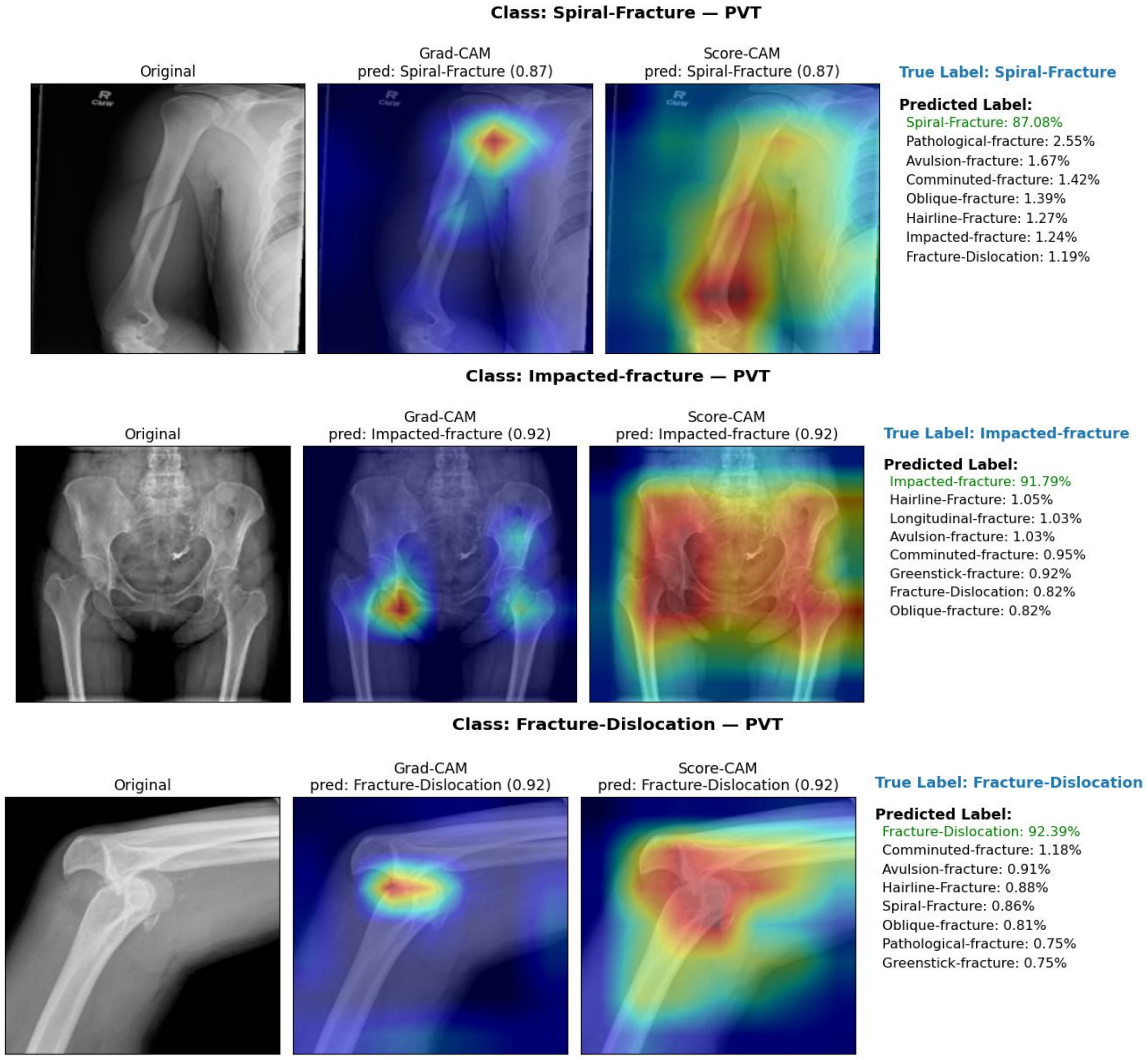

Fig. 5. Grad-CAM visualizations for representative samples

Collectively, these Grad-CAM and Score-CAM results validate the explainability and clinical interpretability of PVT-FractureNet. The model demonstrates an ability not only to achieve high quantitative accuracy but also to visually justify its decision-making process by focusing on anatomically significant regions. This combination of predictive performance and transparency reinforces the model's potential for real-world deployment as a decision-support tool in radiographic fracture diagnosis [61]. The findings indicate that hierarchical transformer-based representations, as implemented in PVT-FractureNet, are capable of providing both precise and trustworthy visual explanations, an essential feature for clinical adoption and physician confidence in AI-assisted diagnostic systems.

5. Conclusion

This study presented PVT-FractureNet, a novel deep learning framework based on the PVT architecture, for automated multi-class bone fracture classification using radiographic images. The model was specifically designed to capture both local fracture features (such as cortical discontinuities

and microstructural distortions) and global anatomical context, enabling accurate and interpretable classification across ten fracture types. Experimental evaluations demonstrated that PVT-FractureNet achieved an average accuracy of 89.9%, AUC of 81.1%, and high specificity (91.1%), indicating reliable performance in distinguishing between subtle and complex fracture morphologies. These results confirm that the hierarchical attention mechanism inherent in PVT can effectively extract multi-scale visual features from X-ray images, outperforming conventional CNN-based approaches in both robustness and diagnostic precision. The ROC analysis further validated the model's strong discriminative capability, with several fracture classes (such as Greenstick, Fracture-Dislocation, and Impacted Fracture) achieving AUC values above 0.90, reflecting excellent separability from other categories. The confusion matrix and per-class evaluation metrics also highlighted that PVT-FractureNet maintained consistent performance across different fracture types, even when the dataset presented imbalanced sample distributions or morphologically similar patterns. Moreover, the integration of explainable AI techniques, including Grad-CAM and Score-CAM, confirmed that the model's decision-making process aligned closely with clinical interpretation. Visualization heatmaps showed that PVT-FractureNet accurately localized fracture regions and focused attention on relevant cortical areas, reinforcing its trustworthiness for potential clinical deployment.

In summary, the findings demonstrate that PVT-FractureNet provides a clinically interpretable, high-performance, and scalable solution for radiographic fracture classification. Its ability to combine feature hierarchies, attention-based representation learning, and explainability mechanisms makes it a promising framework for AI-assisted diagnostic systems in orthopedics and radiology. Future work may focus on expanding the dataset to include multi-view radiographs, 3D CT modalities, or real-world hospital data, as well as incorporating clinical metadata such as patient demographics and injury mechanisms to further enhance diagnostic accuracy. With continued refinement, PVT-FractureNet has the potential to support radiologists in early diagnosis, reduce diagnostic variability, and improve clinical decision-making in fracture management.

Supplementary Materials: The following supporting information can be downloaded at: https://www.kaggle.com/datasets/pkdarabi/bone-break-classification-image-dataset.

Author Contribution: All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding: This research is funded by Institut Teknologi Sepuluh Nopember (ITS) and managed under the Strategic Research Grant (SRG) Type D Scheme (Contract No. 1410/PKS/ITS/2026).

Acknowledgment: This research is funded by Institut Teknologi Sepuluh Nopember (ITS) and managed under the Strategic Research Grant (SRG) Type D Scheme (Contract No. 1410/PKS/ITS/2026). The authors also gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2025.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] W.-J. Metsemakers *et al*., "The global burden of fracture-related infection: can we do better?," *The Lancet Infectious Diseases*, 2023, https://doi.org/10.1016/S1473-3099(23)00503-0.

[2] R. Singh, R. Ambade, S. Landge, S. Goyal, and S. Goel, "Comprehensive Review on Distal Femur Fractures: From Epidemiology to Treatment Strategies," *Cureus*, 2024, https://doi.org/10.7759/cureus.57937.

[3] I. Zakaria, T. M. Yus, S. Rahman, A. Gani, and M. A. Ersan, "Assessing Fracture Detection: A Comparison of Minimal-Resource and Standard-Resource Plain Radiographic Interpretations," *Diagnostics*, vol. 15, no. 7, pp. 876–876, 2025, https://doi.org/10.3390/diagnostics15070876.

[4] Z. M. Pour and S. Berretti, "A comprehensive review of AI methods in upper extremity/limb bone fracture detection," *Artificial Intelligence Review*, vol. 58, no. 10, 2025, https://doi.org/10.1007/s10462-025-11296-6.

[5] A. Abdusalomov *et al*., "Lightweight Deep Learning Framework for Accurate Detection of Sports-Related Bone Fractures," *Diagnostics*, vol. 15, no. 3, pp. 271–271, 2025, https://doi.org/10.3390/diagnostics15030271.

[6] P. K. Mall *et al*., "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities," *Healthcare Analytics*, vol. 4, p. 100216, 2023, https://doi.org/10.1016/j.health.2023.100216.

[7] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep Convolutional Neural Networks in Medical Image Analysis: A Review," *Information*, vol. 16, no. 3, p. 195, 2025, https://doi.org/10.3390/info16030195.

[8] X. Liu *et al*., "Advances in Deep Learning-Based Medical Image Analysis," *Health Data Science*, vol. 2021, pp. 1–14, 2021, https://doi.org/10.34133/2021/8786793.

[9] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, no. 1273253, 2023, https://doi.org/10.3389/fpubh.2023.1273253.

[10] R. Liu, Y. Chen, F. Gai, Y. Liu, Q. Miao, and S. Wu, "Local and Global Spatial–Temporal Transformer for skeleton-based action recognition," *Neurocomputing*, vol. 634, p. 129820, 2025, https://doi.org/10.1016/j.neucom.2025.129820.

[11] S. Takahashi *et al*., "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review," *Journal of Medical Systems*, vol. 48, no. 1, 2024, https://doi.org/10.1007/s10916-024-02105-8.

[12] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technologies*, vol. 13, no. 1, pp. 32–32, 2025, https://doi.org/10.3390/technologies13010032.

[13] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *Journal of Big Data*, vol. 11, no. 1, 2024, https://doi.org/10.1186/s40537-023-00842-0.

[14] X. Shang, S. Wu, Y. Liu, Z. Zhao, and S. Wang, "PVT-MA: pyramid vision transformers with multi-attention fusion mechanism for polyp segmentation," *Applied Intelligence*, vol. 55, no. 1, 2024, https://doi.org/10.1007/s10489-024-06041-5.

[15] J. Liu, H. Li, and W. Kong, "Multi-level learning counting via pyramid vision transformer and CNN," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106184, 2023, https://doi.org/10.1016/j.engappai.2023.106184.

[16] M.-T. Dinh, D.-J. Choi, and G.-S. Lee, "DenseTextPVT: Pyramid Vision Transformer with Deep Multi-Scale Feature Refinement Network for Dense Text Detection," *Sensors*, vol. 23, no. 13, p. 5889, 2023, https://doi.org/10.3390/s23135889.

[17] A. Ahmed, A. S. Imran, Z. Kastrati, S. M. Daudpota, M. Ullah, and W. Noor, "Learning from the few: Fine-grained approach to pediatric wrist pathology recognition on a limited dataset," *Computers in Biology and Medicine*, vol. 181, p. 109044, 2024, https://doi.org/10.1016/j.compbiomed.2024.109044.

[18] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives," *Medical Image Analysis*, vol. 85, p. 102762, 2023, https://doi.org/10.1016/j.media.2023.102762.

[19] H. Elmekki *et al*., "Comprehensive review of reinforcement learning for medical ultrasound imaging," *Artificial Intelligence Review*, vol. 58, no. 9, 2025, https://doi.org/10.1007/s10462-025-11268-w.

[20] I. Das, M. A. Sheakh, S. Abdulla, M. S. Tahosin, M. M. Hassan, and S. Zaman, "Improving Medical X-ray Imaging Diagnosis with Attention Mechanisms and Robust Transfer Learning Techniques," *IEEE Access*, pp. 1–1, 2025, https://doi.org/10.1109/access.2025.3607639.

[21] R. Bagaria, S. Wadhwani, and A. K. Wadhwani, "Bone fractures detection using support vector machine and error backpropagation neural network," *Optik*, vol. 247, p. 168021, 2021, https://doi.org/10.1016/j.ijleo.2021.168021.

[22] K. D. Ahmed and R. Hawezi, "Detection of image processing fracture based on machine learning techniques," *Measurement: Sensors*, p. 100723, 2023, https://doi.org/10.1016/j.measen.2023.100723.

[23] D.-Y. Kim *et al*., "Application of stacked autoencoder for identification of bone fracture," *Journal of the mechanical behavior of biomedical materials/Journal of mechanical behavior of biomedical materials*, vol. 146, pp. 106077–106077, 2023, https://doi.org/10.1016/j.jmbbm.2023.106077.

[24] P. N. Franco *et al*., "Diagnostic performance of an AI algorithm for the detection of appendicular bone fractures in pediatric patients," *European Journal of Radiology*, vol. 178, pp. 111637–111637, 2024, https://doi.org/10.1016/j.ejrad.2024.111637.

[25] N. v. d. Gaast *et al*., "Deep learning for tibial plateau fracture detection and classification," *The Knee*, vol. 54, pp. 81–89, 2025, https://doi.org/10.1016/j.knee.2025.02.001.

[26] J. Zou and M. R. Arshad, "Detection of whole body bone fractures based on improved YOLOv7," *Biomedical signal processing and control*, vol. 91, pp. 105995–105995, 2024, https://doi.org/10.1016/j.bspc.2024.105995.

[27] D. Sutradhar, N. M. Fahad, M. A. K. Raiaan, M. Jonkman, and S. Azam, "Cervical spine fracture detection utilizing YOLOv8 and deep attention-based vertebrae classification ensuring XAI," *Biomedical Signal Processing and Control*, vol. 101, p. 107228, 2025, https://doi.org/10.1016/j.bspc.2024.107228.

[28] C.-H. Tsai, K.-C. Lin, Y.-Y. Chen, P.-C. Chen, Y.-S. Lo, and T.-Y. Ho, "AI assisted diagnosis using DEANet to improve correct diagnosis of iliac wing fracture and ischial spine fracture.," *PubMed*, vol. 128, pp. 106625–106625, 2025, https://doi.org/10.1016/j.clinbiomech.2025.106625.

[29] H. A. Alwzwazy, L. Alzubaidi, Z. Zhao, and Y. Gu, "FracNet: An end-to-end deep learning framework for bone fracture detection," *Pattern Recognition Letters*, vol. 190, pp. 1–7, 2025, https://doi.org/10.1016/j.patrec.2025.01.034.

[30] W. Wei *et al*., "YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images," *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 1, p. 101309, 2025, https://doi.org/10.1016/j.jrras.2025.101309.

[31] A. M. Elsheikh and A. A. Elhag, "Machine learning-based analysis of multi-region bone fracture detection and classification using biomedical images," *Alexandria Engineering Journal*, vol. 128, pp. 186–199, 2025, https://doi.org/10.1016/j.aej.2025.05.074.

[32] K. Kanagaraj, A. Oliver, R. Ranjith, and P. N. Jeipratha, "Bone fracture detection and classification using node-level capsule graph neural network with X-ray images of broken and unbroken bones," *Biomedical Signal Processing and Control*, vol. 110, p. 108302, 2025, https://doi.org/10.1016/j.bspc.2025.108302.

[33] P. K. Darabi, "Bone Break Classification Image Dataset," *Kaggle.com*, 2024. https://www.kaggle.com/datasets/pkdarabi/bone-break-classification-image-dataset.

[34] Y. Pamungkas, E. Triandini, W. Yunanto, and Y. Thwe, "Impact of Hyperparameter Tuning on ResNet-UNet Models for Enhanced Brain Tumor Segmentation in MRI Scans," *International Journal of Robotics and Control Systems*, vol. 5, no. 2, pp. 917–936, 2025, https://doi.org/10.31763/ijrcs.v5i2.1802.

[35] T. B. N.-Tat, T. Q. Hung, P. T. Nam, and V. M. Ngo, "Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities," *Alexandria Engineering Journal*, vol. 119, pp. 558–586, 2025, https://doi.org/10.1016/j.aej.2025.01.090.

[36] J. B. Ruhland, I. Masoudian, and D. Heider, "Enhancing deep neural network training through learnable adaptive normalization," *Knowledge-Based Systems*, vol. 326, p. 113968, 2025, https://doi.org/10.1016/j.knosys.2025.113968.

[37] J. Debnath *et al*., "LMVT: A hybrid vision transformer with attention mechanisms for efficient and explainable lung cancer diagnosis," *Informatics in Medicine Unlocked*, vol. 57, pp. 101669–101669, 2025, https://doi.org/10.1016/j.imu.2025.101669.

[38] Y. Pamungkas, E. Triandini, W. Yunanto, and Y. Thwe, "Enhancing Diabetic Retinopathy Classification in Fundus Images using CNN Architectures and Oversampling Technique", *J Robot Control (JRC)*, vol. 6, no. 1, pp. 413–425, 2025, https://doi.org/10.18196/jrc.v6i1.25331.

[39] E. Aktas, N. Ceylan, E. Y. Bilgin, E. Bilgin, and L. Ince, "Evaluation of calcaneal inclusion angle in the diagnosis of pes planus with pretrained deep learning networks: An observational study," *Medicine*, vol. 104, no. 31, pp. e43639–e43639, 2025, https://doi.org/10.1097/md.0000000000043639.

[40] D. Hussain and Y. H. Gu, "Exploring the Impact of Noise and Image Quality on Deep Learning Performance in DXA Images," *Diagnostics*, vol. 14, no. 13, pp. 1328–1328, 2024, https://doi.org/10.3390/diagnostics14131328.

[41] R. Khan, N. Alzaben, Y. I. Daradkeh, M. Y. Lee, and I. Ullah, "Bilateral collaborative streams with multi-modal attention network for accurate polyp segmentation," *Scientific Reports*, vol. 15, no. 1, 2025, https://doi.org/10.1038/s41598-025-15401-1.

[42] M. U. Saeed, W. Bin, J. Sheng, and H. M. Albarakati, "An Automated Multi-scale Feature Fusion Network for Spine Fracture Segmentation Using Computed Tomography Images," *Deleted Journal*, vol. 37, no. 5, pp. 2216–2226, 2024, https://doi.org/10.1007/s10278-024-01091-0.

[43] K. Han, Q. Wang, M. Zhu, and X. Zhang, "PVTReID: A Quick Person Reidentification-Based Pyramid Vision Transformer," *Applied sciences*, vol. 13, no. 17, pp. 9751–9751, 2023, https://doi.org/10.3390/app13179751.

[44] J. Ju, Z. Qu, H. Qing, Y. Ding, and L. Peng, "Evaluation of Artificial Intelligence-based diagnosis for facial fractures, advantages compared with conventional imaging diagnosis: a systematic review and meta-analysis," *BMC Musculoskeletal Disorders*, vol. 26, no. 1, 2025, https://doi.org/10.1186/s12891-025-08842-2.

[45] C. Rainey, J. McConnell, C. Hughes, R. Bond, and S. McFadden, "Artificial Intelligence for Diagnosis of Fractures on Plain radiographs: a Scoping Review of Current Literature," *Intelligence-Based Medicine*, vol. 5, p. 100033, Apr. 2021, https://doi.org/10.1016/j.ibmed.2021.100033.

[46] S. A. Hicks *et al*., "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, p. 5979, 2022, https://doi.org/10.1038/s41598-022-09954-8.

[47] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification," *PLOS ONE*, vol. 19, no. 12, p. e0316019, 2024, https://doi.org/10.1371/journal.pone.0316019.

[48] L. Farah, J. Murris, I. Borget, A. Guilloux, N. Martelli, and S. Katsahian, "Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence–Based Health Technologies: What Healthcare Stakeholders Need to Know," vol. 1, no. 2, pp. 120–138, 2023, https://doi.org/10.1016/j.mcpdig.2023.02.004.

[49] Y. Zhang, D. Hong, D. McClement, O. Oladosu, G. Pridham, and G. Slaney, "Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging," *Journal of Neuroscience Methods*, vol. 353, p. 109098, 2021, https://doi.org/10.1016/j.jneumeth.2021.109098.

[50] I. D. Mienye *et al*., "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Informatics in Medicine Unlocked*, vol. 51, p. 101587, 2024, https://doi.org/10.1016/j.imu.2024.101587.

[51] N. Shifa, M. Saleh, Y. Akbari, and S. Al Maadeed, "A review of explainable AI techniques and their evaluation in mammography for breast cancer screening," *Clinical Imaging*, vol. 123, p. 110492, 2025, https://doi.org/10.1016/j.clinimag.2025.110492.

[52] B. Kocak *et al*., "Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations," *European Journal of Radiology Artificial Intelligence*, p. 100030, 2025, https://doi.org/10.1016/j.ejrai.2025.100030.

[53] W. Albattah and R. U. Khan, "Impact of imbalanced features on large datasets," *Frontiers in Big Data*, vol. 8, 2025, https://doi.org/10.3389/fdata.2025.1455442.

[54] L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, and E. Vezzetti, "Vision Transformer for femur fracture classification," *Injury*, vol. 53, no. 7, pp. 2625–2634, 2022, https://doi.org/10.1016/j.injury.2022.04.013.

[55] M. Ahmad, M. H. F. Butt, M. Mazzara, S. Distefano, A. M. Khan, and H. A. Altuwaijri, "Pyramid Hierarchical Spatial-Spectral Transformer for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–10, 2024, https://doi.org/10.1109/jstars.2024.3461851.

[56] M. R. Ahmed *et al*., "Hierarchical Swin Transformer Ensemble with Explainable AI for Robust and Decentralized Breast Cancer Diagnosis," *Bioengineering*, vol. 12, no. 6, pp. 651–651, 2025, https://doi.org/10.3390/bioengineering12060651.

[57] M. Ennab and H. Mcheick, "Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models," *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, p. 12, 2025, https://doi.org/10.3390/make7010012.

[58] J. Zhang *et al*., "Advances in attention mechanisms for medical image segmentation," *Computer Science Review*, vol. 56, p. 100721, 2025, https://doi.org/10.1016/j.cosrev.2024.100721.

[59] C. M. Tsai and J.-D. Lee, "Dynamic Ensemble Learning with Gradient-Weighted Class Activation Mapping for Enhanced Gastrointestinal Disease Classification," *Electronics*, vol. 14, no. 2, p. 305, 2025, https://doi.org/10.3390/electronics14020305.

[60] D. Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, "Reviewing CAM-Based Deep Explainable Methods in Healthcare," *Applied Sciences*, vol. 14, no. 10, p. 4124, 2024, https://doi.org/10.3390/app14104124.

[61] S. A. El-Ghany, M. A. Mahmood, and A. A. A. El-Aziz, "FracFusionNet: A Multi-Level Feature Fusion Convolutional Network for Bone Fracture Detection in Radiographic Images," *Diagnostics*, vol. 15, no. 17, pp. 2212–2212, 2025, https://doi.org/10.3390/diagnostics15172212.