Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Hierarchical Transformer with Residual CGTM Blocks and Gated Dilated Convolutional Refinement for Binary Bone Fracture Classification from Radiographic Images

Md Fakrul Alam
Master of Science in Information Studies,Trine University, USA
Corresponding Author: Md Fakrul Alam, **E-mail**: Shaikot9999@gmail.com

| ABSTRACT

Automated bone fracture classification from radiographic images remains challenging because fracture evidence is often subtle, spatially confined, and sensitive to projection angle and anatomical variation. While convolutional models and standard transformer-based architectures have advanced radiographic image analysis, they may not jointly capture fine-grained local structural disruptions and the broader contextual dependencies required for reliable fracture recognition. In particular, existing methods do not sufficiently unify localized fracture-sensitive feature extraction, cross-window contextual interaction, and multi-scale structural refinement within a single classification framework. To address this gap, this study proposes a hierarchical transformer-based network for binary bone fracture classification built on residual CGTM blocks. The architecture integrates window-based multi-head self-attention, shifted-window multi-head self-attention, and a Gated Dilated ConvFFN within a four-stage representation hierarchy, with adaptive patch merging between stages and attention pooling for discriminative feature aggregation. The model was evaluated on a bone fracture radiograph dataset using dedicated training, validation, and held-out test subsets, with performance stability assessed through 10-fold cross-validation. Comparative evaluation was conducted against MaxViT and EVA-02 under identical experimental conditions. The proposed model achieved a mean validation accuracy of 99.276% and a held-out test accuracy of 99.25%, together with a ROC-AUC of 99.99% and an average precision of 99.99%, outperforming MaxViT at 97.99% test accuracy and EVA-02 at 93.355% mean validation accuracy. These results indicate that jointly modeling local structural sensitivity and cross-window contextual interaction within a hierarchical transformer framework offers a meaningful advantage for radiographic fracture classification, though broader validation remains necessary before clinical use.

| KEYWORDS

bone fracture classification; hierarchical transformer; residual CGTM blocks; gated dilated convolutional refinement; radiographic image analysis; window-based self-attention

| ARTICLE INFORMATION

ACCEPTED: 10 April 2026 **PUBLISHED:** 25 May 2026 **DOI:** 10.32996/jcsts.2026.8.5.18

1.Introduction

Bone fractures are among the most frequently encountered injuries in clinical practice and require prompt radiographic assessment to guide appropriate treatment [1,2]. Plain radiography remains the primary imaging modality for initial fracture evaluation because it is widely available, relatively inexpensive, and provides rapid structural information across a broad range of clinical settings[3,4]. As the demand for radiographic interpretation continues to grow, automated fracture classification from X-ray images has attracted increasing attention as a means of supporting consistent and timely diagnostic decisions [5,6]. Binary fracture classification, which distinguishes radiographs showing fracture-related structural disruption from those representing normal anatomy, provides a focused and clinically relevant computational problem [7,8,9]. Addressing this task reliably through automated methods is technically meaningful, particularly in settings where specialist availability is limited and diagnostic throughput is high.

Radiographic fracture recognition is inherently difficult, even for experienced clinicians [10]. Fracture lines may be faint, incomplete, low in contrast, or partially obscured by overlapping anatomical structures, and their appearance can change substantially depending on the projection angle, bone morphology, and acquisition orientation [11]. The diagnostically relevant features that distinguish a fractured from a non-fractured radiograph often consist of thin cortical interruptions, slight step-offs, irregular trabecular patterns, or narrow discontinuities that occupy only a small portion of the full image [12]. These local findings cannot be interpreted in isolation. Their clinical significance depends on surrounding bone contours, joint configuration, and the broader spatial relationship between structures within a given projection. This dual requirement of fine-grained sensitivity to compact local irregularities combined with awareness of the surrounding anatomical context makes automated fracture classification from radiographs a technically demanding problem that places specific constraints on the representational capacities of deep learning systems[13].

Deep learning has substantially advanced the field of medical image analysis, and both convolutional neural networks and transformer-based architectures have been applied to radiographic classification tasks with considerable success [14,15]. Convolutional models exploit hierarchical local receptive fields to extract spatially structured features efficiently, and they have demonstrated strong performance across a wide range of imaging benchmarks [16,17]. Their dependence on progressively stacked local operations, however, can limit direct modeling of long-range structural dependencies, which may be important when local fracture evidence must be interpreted relative to more distant anatomical structures or broader projection-dependent context [18]. Transformer-based models address part of this limitation by enabling flexible token-level interactions across the full spatial extent of the image [19,20]. Yet generic attention formulations do not always preserve spatially delicate local detail with sufficient fidelity, particularly when the relevant evidence is sparse, edge-like, and confined to a narrow region [21,22]. Hierarchical vision transformers improve on this by operating across multiple spatial scales, but standard attention pipelines within such architectures may still lack the spatial refinement mechanisms needed to respond consistently to subtle osseous irregularities in radiographic images.

These considerations point to a concrete methodological gap. Existing approaches do not sufficiently unify localized fracture-sensitive feature extraction, cross-window contextual interaction, and multi-scale structural refinement within a single classification framework. For radiographic bone fracture analysis, this unification is particularly important because fracture evidence is often confined to a narrow spatial region while its interpretation depends on surrounding anatomy and view-specific structural relationships. An architecture suited to this task should support fine-grained attention within local spatial windows, enable communication across neighboring windows so that structural patterns can be interpreted in their spatial context, incorporate spatially enriched feed-forward processing that extends sensitivity beyond the immediate attention window, and progressively abstract features across hierarchical stages without losing the localized structural detail present at finer scales. Generic convolutional or transformer pipelines applied without task-specific adaptation may not simultaneously fulfill all of these representational requirements.

To address this gap, the present study proposes a hierarchical transformer-based network for binary bone fracture classification built around residual CGTM blocks. The architecture operates as a four-stage feature hierarchy on patch-partitioned radiographic inputs, performing progressive spatial downsampling with channel expansion across stages. Each residual CGTM block integrates window-based multi-head self-attention for localized structural modeling, shifted-window multi-head self-attention to promote cross-window contextual interaction, and a Gated Dilated ConvFFN for spatially enriched feed-forward refinement. The dilated convolutional component within the feed-forward stage expands the effective receptive field without collapsing spatial resolution, and the gating mechanism supports adaptive feature propagation based on the incoming representation state. Adaptive patch merging handles stage transitions through learned hierarchical downsampling and feature consolidation, while attention pooling at the final stage aggregates spatial tokens into a compact representation before classification. This design was motivated by the specific representational demands of fracture-sensitive radiographic analysis rather than by the application of a general-purpose visual backbone without task-specific adaptation.

This study evaluates the proposed architecture on a binary bone fracture radiograph dataset using dedicated training, validation, and held-out test subsets, with 10-fold cross-validation employed to assess performance stability. The model is compared against two competitive transformer-based baselines, MaxViT and EVA-02, under a consistent and controlled experimental framework. The central contribution of this work is a fracture-sensitive hierarchical transformer design that more explicitly integrates localized structural sensitivity, cross-window contextual reasoning, and spatially aware feature refinement within a unified architecture, providing a more task-appropriate framework for automated bone fracture classification from radiographs than generic feature extraction pipelines alone.

2. Related Work

Recent contributions to bone fracture classification from radiographic images reflect a broad range of architectural strategies, from attention-augmented encoder-decoder networks to novel quantum-inspired pipelines. Devi et al. [23] proposed SAG-UNet, which refines the attention gate mechanism of a standard Attention UNet within the decoder upsampling path using batch normalization, ReLU activation, dropout, and 1x1 linear convolution to construct a spatial attention weight map that is propagated through skip connections; applied to ten-class fracture type classification on the Kaggle Bone Break dataset with Growing Neural Gas-based feature quantization as a preprocessing step, SAG-UNet achieved a classification accuracy of 98.78%, substantially improving upon the 89% baseline of the unmodified Attention UNet and demonstrating that targeted architectural refinement of the gating mechanism can yield a meaningful discriminative advantage. Operating from a different methodological premise, Mewada et al. [24] integrated a quantum convolution layer as the initial feature extraction stage of a conventional CNN pipeline, constructing a quantum convolutional neural network evaluated on the Bone Fracture Multi-Region X-ray benchmark, where it achieved an accuracy of 96% and an F1-score of 95.81%; while the approach introduces a novel computational paradigm, its reported performance does not surpass that of well-tuned classical architectures. Elsheikh and Elhag [25] proposed DLCVM-MBFDC, combining Wiener filtering for noise suppression, DenseNet-201 for deep feature extraction, and a stacked autoencoder optimized with the wild gibbon optimizer algorithm for multi-region fracture classification, reporting a superior accuracy of 97.85% on the Kaggle dataset and illustrating that modular optimization of independent pipeline stages can produce competitive classification performance, albeit at the cost of end-to-end representational coherence. Kanagaraj et al. [26] introduced a Node-Level Capsule Graph Neural Network (NCGNN) for binary fracture classification, preprocessing radiographic images with a Distributed Nonlinear Polynomial Graph Filter, extracting textural features including contrast, homogeneity, energy, and dissimilarity through the High-Order Synchroextracting Transform, and optimizing the NCGNN with the Gazelle Optimization Algorithm; while the capsule routing mechanism provides equivariance to spatial transformations, the architecture depends on multi-stage handcrafted feature extraction prior to the classification stage, which constrains the flexibility of the learned representation. Torne et al. [27] conducted a systematic comparison of VGG-16, ResNet-50 paired with a support vector machine, and EfficientNetB0 paired with XGBoost on a ten-class fracture type classification task using X-ray images, reporting that VGG-16 and its ensemble variant with random forest achieved the highest accuracy of 0.95 across all fracture classes; the results confirm the practical value of ensemble augmentation of standard convolutional backbones, though the purely convolutional feature hierarchy does not explicitly address cross-region structural dependencies that may be relevant when fracture evidence must be interpreted in relation to broader anatomical context.

Several recent studies have explored the application of convolutional and transformer-based architectures for radiographic fracture classification, with a shared emphasis on transfer learning, hybrid model design, and systematic backbone comparison. Sekar et al. [28] evaluated a broad set of architectures including ResNet, DenseNet, VGG-19, EfficientNet, Vision Transformers, and domain-adapted hybrid variants such as ResNet-Hybrid and DenseNet-Hybrid on a bone fracture categorization task, finding that the DenseNet-Hybrid configuration achieved near-perfect accuracy owing to its domain-specific structural adaptations, while Vision Transformers demonstrated superior global relational modeling relative to EfficientNet and standard SVM-based approaches showed substantially lower performance due to their reliance on manually defined feature representations; these findings confirm that hybrid architectures incorporating domain adaptation can outperform both generic convolutional backbones and conventional machine learning pipelines in fracture classification settings. Alex [29] applied a pre-trained ResNet-18 to binary fracture classification on a multi-region dataset of 10,580 radiographic images, achieving a validation accuracy of 97.59% with rotation and horizontal flip augmentation, and demonstrated that lightweight convolutional backbones under transfer learning can deliver competitive binary discrimination performance while maintaining computational efficiency suitable for resource-constrained clinical environments. Ahammed et al. [30] systematically evaluated DenseNet-169, VGG, ResNet, Xception, and EfficientNet as fixed feature extractors paired with logistic regression, random forest, XGBoost, and a custom feed-forward network as classifiers, reporting that DenseNet-169 combined with the custom feed-forward classifier achieved the highest binary classification accuracy of 99%, a ROC-AUC of 99.99%, and equivalent precision, recall, and F1-score of 998%, establishing that the selection of both the feature extractor and the downstream classifier exerts a large influence on final classification performance and that DenseNet-based representations are particularly well suited to fracture-discriminative feature learning when paired with an appropriate classification head. Islam et al. [31] compared Vision Transformer (ViT-16) and convolutional architectures including CNN, ResNet-50, and ResNet-34 on the MURA musculoskeletal radiograph dataset of 40,019 images for binary fracture classification, finding that ViT-16 achieved the highest reported accuracy of 88%, followed closely by a standard CNN at 87% and ResNet-50 at 86%; while these results confirm that vision transformers are applicable to radiographic fracture classification, the observed performance levels are notably lower than those achieved by specialized convolutional or hybrid designs on comparable tasks, indicating that generic ViT formulations applied without task-specific architectural adaptation do not recover sufficient local structural sensitivity to match fracture-specialized convolutional models.

A further cluster of recent studies has investigated hybrid feature extraction pipelines, backbone architecture comparisons, and end-to-end detection frameworks as complementary strategies for improving automated fracture analysis from radiographic images. Alam et al. [32] proposed MobLG-Net, which combines MobileNet as a transfer learning-based spatial feature extractor with a light gradient boosting machine (LGBM) to generate class probability representations, subsequently evaluated against KNN, LGBM, logistic regression, and random forest classifiers with optimized hyperparameters and cross-validation; the LGBM and logistic regression classifiers trained on the MobLG-Net features achieved a binary fracture detection accuracy of 99%, demonstrating that routing compact convolutional representations through tree-based classifiers can produce competitive performance without the computational overhead of large-scale end-to-end training. Aldhyani et al. [33] evaluated VGG-16, ResNet-152V2, and DenseNet-201 for multi-region fracture detection and classification on the Kaggle dataset of 10,580 radiographic images, reporting a validation accuracy of 97% with DenseNet-201 and VGG-16 and identifying data heterogeneity and inter-class visual ambiguity as the primary factors constraining performance in existing convolutional models, thereby highlighting that standard transfer learning without architectural adaptation may not be sufficient to handle the full range of structural variability present in clinical fracture datasets. Alwzwazy et al. [34] proposed FracNet, an end-to-end detection framework incorporating self-supervised pretraining, multi-scale feature fusion, attention mechanisms, and integrated visualization tools, which achieved a reported detection accuracy of 100% across three benchmark datasets and demonstrated high adaptability to new data with minimal retraining requirements; while the framework provides spatial interpretability beyond image-level classification, its formulation as a detection rather than classification system means it is not directly comparable to binary classification pipelines operating on full radiographic images. Isgor and Koklu [35] proposed a lightweight hybrid approach combining MobileNetV2 exclusively for deep feature extraction with five ensemble classifiers, namely Random Forest, Extra Trees, Gradient Boosting, AdaBoost, and Bagging Classifier, evaluated under 10-fold cross-validation on a dataset of 4,840 fractured and 4,623 non-fractured bone X-ray images; the Extra Trees classifier achieved the highest performance with an accuracy of 99.14%, precision of 99.22%, recall of 99.03%, and F1-score of 99.12% within a training time of 199 seconds, confirming that decoupling deep feature extraction from ensemble classification can deliver both high accuracy and computational efficiency, though the fixed feature extractor prevents joint optimization of the representational and discriminative components of the model.

The final group of reviewed contributions spans detection-oriented architectures, classical hybrid classification pipelines, lightweight backbone comparisons, and multi-stage segmentation and classification frameworks, collectively illustrating the methodological breadth of current automated fracture analysis research. Srinivasu et al. [36] investigated the combined effect of hyperparameter tuning and preprocessing augmentation on YOLO V10 for radiographic fracture detection, applying unsharp masking and contrast-limited adaptive histogram equalization prior to training and systematically varying learning rate and epoch count; the optimally configured model achieved a detection accuracy of 0.964 on augmented data with a precision of 0.98, compared to 0.95 on raw images, demonstrating that preprocessing strategy and training configuration exert a substantial and measurable influence on localization-based fracture detection performance independent of architectural design. Scutelnicu et al. [37] provided a systematic review of automatic bone fracture detection techniques encompassing image processing methods, classical machine learning algorithms, CNN-based models, region-based convolutional networks, and transformer-based architectures, identifying data availability, image quality variability, computational complexity, and integration into clinical workflows as the primary challenges confronting current approaches, and highlighting multimodal data fusion, explainable artificial intelligence, and federated learning as the most promising directions for improving diagnostic reliability and system deployability in future research. Padmakala [38] compared Custom CNN, MobileNetV2, and ResNet-50 on a binary orthopedic fracture classification task with images normalized and resized to 224x224 pixels, reporting that MobileNetV2 achieved perfect accuracy, sensitivity, and AUC of 1.000 under the reported experimental conditions, while the Custom CNN attained an accuracy of 97.74% with a sensitivity of 95.50% and AUC of 0.9915, and ResNet-50 substantially underperformed both alternatives at 73.18% accuracy and an AUC of 0.7820, illustrating that backbone selection has a disproportionately large effect on binary fracture classification performance even when all other training conditions are held constant. Saritas et al. [39] employed InceptionV3 within a transfer learning framework to extract multi-layered feature representations from radiographic images, subsequently classifying these representations using SVM and Naive Bayes, with SVM achieving an accuracy of 76.8% and Naive Bayes reaching 67.3%; while the decoupled extraction and classification design is computationally convenient, the reported accuracy levels fall considerably below those achieved by end-to-end trained architectures on comparable fracture classification tasks, suggesting that fixed feature representations extracted without joint optimization may not capture the full discriminative structure required for reliable fracture-sensitive classification. Jothi and Jayanthi [40] proposed a multi-stage pipeline that integrates a dilated mask R-CNN with recurrent Kalman filters and recurrent fuzzy filters for bone fracture segmentation, followed by feature concatenation of deep mask R-CNN outputs with handcrafted descriptors including local binary patterns, histogram of oriented gradients, and SURF, which are then classified by a Self-Attention Recurrent Neural Network; while the framework captures both spatial localization and temporal feature dependencies within a single pipeline, its dependence on

handcrafted intermediate representations and multi-stage optimization constrains the end-to-end representational flexibility of the model and introduces design complexity that may limit generalization across datasets with different imaging characteristics.

3.1 Dataset description

The dataset used in this study supports binary bone fracture classification from radiographic images, with two target classes: Fractured and Not Fractured. Radiographs are partitioned into training, validation, and test subsets, which respectively support model fitting, performance monitoring, and independent held-out evaluation.Figure 1 presents the class distribution across all three subsets. The training subset is the largest and is approximately class-balanced, with a slight predominance of Fractured images. The validation subset is smaller and shows moderate imbalance, with more Not Fractured samples than Fractured samples. The test subset contains 199 Not Fractured and 200 Fractured radiographs, giving a near-balanced composition of 399 samples used for final assessment. The dataset consists predominantly of upper-limb radiographs, with wrist and forearm images constituting the main anatomical content. The images vary in positioning, projection angle, and anatomical alignment, reflecting the kind of diversity encountered in routine radiographic practice. Fracture patterns may present differently depending on the viewing angle, and projection-related structural overlap can alter the local appearance of osseous regions. This variability means that binary classification cannot rely on global appearance alone.Effective discrimination between Fractured and Not Fractured radiographs requires the ability to detect subtle cortical irregularities and local structural discontinuities while also interpreting those findings relative to surrounding bone contours and view-specific anatomical context. These dataset characteristics directly motivate the need for a model that jointly captures fine local fracture cues and broader structural relationships.

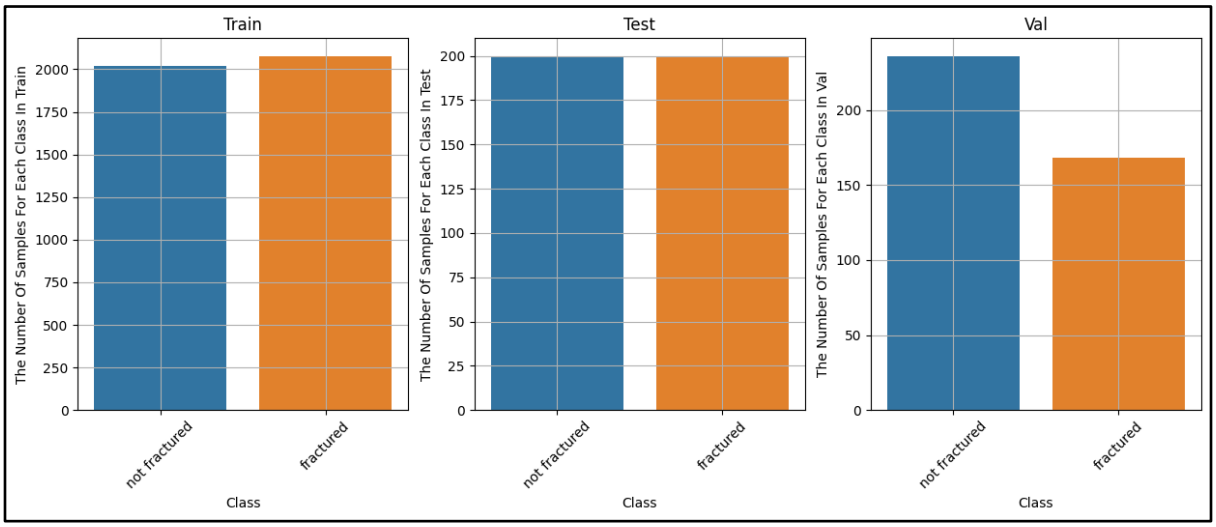

Figure 1:Class distribution of the bone fracture radiograph dataset across training, validation, and test subsets

3.2 Data preprocessing and augmentation

Radiographic bone fracture classification places specific demands on input preprocessing. X-ray images vary in spatial resolution, intensity range, and acquisition orientation, yet this variability is not inherently diagnostic and can destabilize model optimization if left unaddressed. At the same time, preprocessing must remain structurally conservative, because the features that distinguish fractured from non-fractured radiographs often manifest as fine cortical interruptions, narrow discontinuities, or subtle local contour deviations that can be suppressed by aggressive normalization or unconstrained photometric transformation. The pipeline was therefore designed to reduce non-informative acquisition variability in a reproducible way while preserving the fine osseous detail on which reliable fracture discrimination depends.

All radiographs were represented as three-channel inputs to ensure compatibility with pretrained vision architectures that expect this tensor format. Where images were not natively three-channel, the luminance signal was replicated uniformly across all three channels without channel-specific filtering or intensity transformation. Each image was resized to 224 × 224 pixels to establish a uniform spatial resolution for batch-based training. Pixel intensities were then linearly scaled to the range [0, 1], after which channel-wise normalization was applied using the fixed ImageNet statistics: mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. These constants were adopted as fixed external conventions aligned with pretrained backbone expectations and were not estimated from any portion of the present dataset. This distinction is methodologically important: computing normalization statistics from training data and applying them uniformly across subsets would constitute a form of

preprocessing-level information leakage, which the current approach avoids entirely. No contrast enhancement, edge sharpening, denoising, or anatomical cropping was applied, as such operations are unnecessary for the binary classification objective and risk suppressing fracture-relevant structural information.

Stochastic data augmentation was applied exclusively to the training subset to improve model robustness to realistic appearance variability in musculoskeletal radiographic images. The augmentation policy was intentionally kept moderate, because fine fracture signatures such as thin cortical interruptions and narrow structural discontinuities can be distorted or obscured by strong geometric or photometric perturbations. Horizontal flipping was applied with probability 0.5. Rotation was constrained to a maximum of ±10 degrees, translation was limited to within 10 percent of image width and height, and zoom was restricted to the range 0.90 to 1.10. Brightness and contrast were each independently perturbed within ±10 percent to simulate mild acquisition-related intensity differences without producing radiographically implausible inputs. Vertical flipping was excluded because it yields anatomically implausible orientations for musculoskeletal radiographs and could introduce structurally misleading patterns into the training distribution.

Data partitioning into training, validation, and test subsets was completed before any preprocessing or augmentation was applied. Augmented variants were generated entirely within the training subset and were not permitted to cross partition boundaries. Validation and test subsets received only the deterministic operations described above: fixed-size resizing, pixel intensity scaling, and channel-wise normalization using the same fixed ImageNet statistics. No stochastic transformation was applied to these subsets at any stage. Because normalization constants were not derived from the available data, no statistics-based information leakage between subsets was introduced at the preprocessing stage. This ordering and separation preserve the methodological integrity of each subset's role: training for parameter optimization, validation for model selection, and test for independent assessment of generalization performance on data that remained entirely isolated from model development.

3.3 Overall architecture of the proposed bone fracture classification network

The proposed model is formulated as a hierarchical transformer-based architecture for binary bone fracture classification from radiographic images (Figure 2) . Its design is motivated by the observation that fracture recognition depends on two complementary forms of visual reasoning. On the one hand, the model must remain sensitive to subtle local structural irregularities such as fine cortical disruption, narrow discontinuities, and weak fracture-related contour deviations. On the other hand, these local findings must be interpreted in relation to the surrounding anatomical context, since projection, alignment, overlap, and normal morphological variation can alter the apparent visual pattern. This dual requirement motivates an architecture that combines localized feature interaction, cross-region contextual modeling, and progressive hierarchical abstraction within a unified end-to-end framework.Given an input radiograph $x \epsilon R^{H \times W \times 3}$, the network first performs patch partition and patch embedding to convert the image into an initial latent representation. This initialization step can be expressed as :

$$Z_0 = PE(X),$$

where PE(·)denotes the combined patch partition and patch embedding operation. The resulting tokenized representation is then processed through four sequential stages of hierarchical feature transformation. The global processing pipeline begins with the input radiograph, followed by patch partition and patch embedding, then Stage 1, followed by adaptive patch merging and Stage 2, again followed by adaptive patch merging and Stage 3, then adaptive patch merging and Stage 4, and finally attention pooling, a fully connected layer, and softmax classification. The architecture is built around repeated residual CGTM blocks and performs progressive spatial downsampling together with channel expansion, enabling the learning of increasingly abstract fracture-sensitive representations across depth.

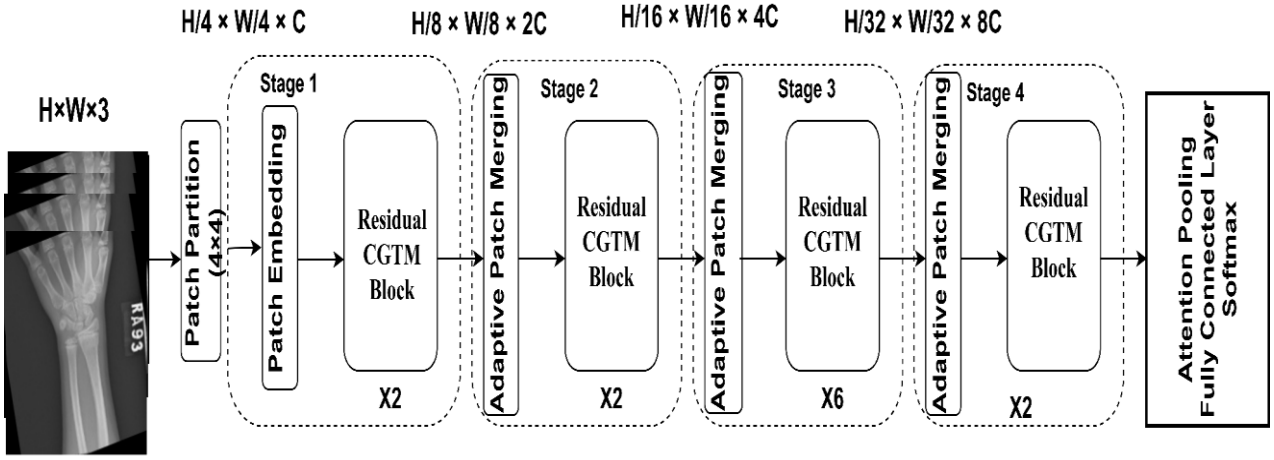

Figure 2: Overall architecture of the proposed hierarchical transformer-based bone fracture classification network

At the methodological level, the model is distinguished by the coordinated integration of window-based multi-head self-attention, shifted-window multi-head self-attention, gated dilated convolutional feed-forward refinement, hierarchical stage-wise representation learning, and attention-based final feature aggregation. This combination is technically meaningful for fracture classification because local attention windows help preserve short-range structural dependencies that are relevant to fine osseous irregularities, while shifted-window interaction facilitates information exchange across neighboring spatial regions that may jointly define a fracture pattern. The gated dilated convolutional feed-forward component further enriches the representation by introducing spatial refinement with broader contextual sensitivity than a plain feed-forward transformation. As a result, the proposed model should not be viewed as a generic vision transformer applied directly to radiographs, but rather as a fracture-sensitive hierarchical design intended to retain localized structural detail while also modeling contextual anatomical dependencies that influence fracture interpretation.

$$Z_s = F_s(Z_{s-1}) \, , s \in \{1,2,3,4\},$$

where $F_s(\cdot)$ denotes the feature transformation associated with the s-th hierarchical stage. This staged progression allows the network to move from relatively fine-grained structural encoding toward increasingly abstract contextual representations while preserving the hierarchical organization needed for fracture-sensitive analysis.After the final stage, the learned high-level representation is aggregated through an attention pooling module before being mapped to the final binary decision through a fully connected layer and softmax classifier. This final design choice is consistent with the overall objective of emphasizing the most discriminative fracture-relevant features rather than relying on uniform global averaging alone. The present subsection has introduced the model at the system level and established the logic of its end-to-end processing pipeline. The following subsections describe the residual CGTM block, the hierarchical stage transitions, and the attention pooling classification head in greater detail.

3.4 Residual CGTM block for fracture-sensitive feature modeling

The residual CGTM block forms the principal computational unit of the proposed architecture and is designed to support fracture-sensitive representation learning within radiographic images. Its structure is motivated by the observation that reliable fracture classification requires simultaneous sensitivity to fine local structural irregularities and awareness of their relationship to surrounding anatomical context. To address this requirement, the block combines localized window-level attention, shifted cross-window interaction, and spatially aware feed-forward refinement within a residual pre-normalized formulation. Layer normalization is used before each transformation to stabilize feature statistics, while residual learning preserves information flow and facilitates optimization in deeper hierarchical stages. In this way, the block is constructed to exploit the complementarity between attention-based relational modeling and convolutional structural refinement, which is particularly relevant for subtle osseous discontinuities that may occupy only a limited spatial region.

Let Z denote the input feature tensor to a residual CGTM block. The first sub-block applies LayerNorm followed by window-based multi-head self-attention and a residual update. This step is expressed as

$$\hat{Z} = Z + WMSA(LN(Z)),$$

where LN(·) denotes layer normalization and WMSA(·) denotes window-based multi-head self-attention. The role of this operation is to model localized feature interactions within non-overlapping spatial windows. Such window-restricted attention is well suited to radiographic fracture analysis because many relevant cues, including thin cortical interruptions, short fracture lines, and local contour irregularities, are spatially confined and require fine-grained structural discrimination rather than global averaging. By restricting attention to local windows, the block preserves sensitivity to compact structural evidence while maintaining computational tractability across hierarchical feature maps.Following this attention-based update, the block performs a second normalized residual refinement through the Gated Dilated ConvFFN:

$$Z' = \hat{Z} + GDCFFN(LN(\hat{Z})),$$

where GDCFFN(·) denotes the Gated Dilated ConvFFN transformation. This module serves as a spatially aware feed-forward refinement stage that complements the relational modeling introduced by local self-attention. The use of dilated convolution expands the effective receptive field without requiring excessive spatial downsampling, thereby allowing the block to incorporate broader contextual cues around localized fracture-relevant regions. At the same time, the gating operation supports adaptive feature modulation, enabling the network to regulate how strongly refined features are propagated based on the incoming representation. In the context of fracture classification, this refinement is meaningful because a local discontinuity may need to be interpreted together with neighboring cortical alignment, adjacent bone boundaries, and surrounding anatomical structure. The Gated Dilated ConvFFN therefore enhances the block's ability to preserve subtle local evidence while enriching it with structurally relevant context.Although localized window attention is effective for short-range structural modeling, its fixed partitioning can limit direct information exchange across neighboring windows. To address this limitation, the residual CGTM design incorporates a complementary shifted-window attention stage after the initial W-MSA and feed-forward refinement. Specifically, the intermediate representation Z' is normalized and processed by shifted-window multi-head self-attention, followed by residual addition:

$$\underline{Z} = Z' + SWMSA(LN(Z')),$$

where SWMSA(·) denotes shifted-window multi-head self-attention. By shifting the attention partition, this operation promotes cross-window interaction and allows information from adjacent regions to be integrated more effectively. This is particularly important in radiographic fracture analysis because a diagnostically relevant structure may span a window boundary or require joint interpretation of neighboring cortical segments. After this contextual interaction, a final normalized Gated Dilated ConvFFN refinement is applied with residual learning:

$$Z'' = \underline{Z} + GDCFFN(LN(\underline{Z})),$$

yielding the output feature tensor Z'' of the block. Together, Eqs. (3) to (6) describe a sequence in which intra-window dependency modeling, spatial refinement, inter-window contextual exchange, and additional structural refinement are integrated within a unified residual unit. The combined use of W-MSA and SW-MSA enables the block to encode both local and neighboring structural relationships, while repeated Gated Dilated ConvFFN refinement helps maintain spatial sensitivity beyond what a standard transformer feed-forward layer would provide.The methodological contribution of the residual CGTM block lies in this deliberate integration of local attention, shifted contextual interaction, and gated dilated convolutional refinement within a single hierarchical transformer unit tailored to fracture-sensitive radiographic analysis. Rather than relying solely on generic self-attention or purely convolutional processing, the block is specifically motivated by the need to represent subtle osseous irregularities while preserving broader structural continuity across the image. This design is appropriate for wrist and forearm radiographs, where clinically relevant fracture evidence may be both spatially localized and context dependent. The residual CGTM block therefore serves as the foundation for the stage-wise representation learning strategy of the full network. The next subsection describes how these blocks are organized across the hierarchical stages and how adaptive patch merging supports progressive feature abstraction.

3.5 Hierarchical multi-stage representation learning and adaptive patch merging

The proposed network employs a four-stage hierarchical representation strategy to transform the patch-embedded radiograph into progressively more abstract and fracture-sensitive feature representations (Figure 3) This stage-wise organization is important for bone fracture classification because the target visual evidence spans multiple levels of representation. Fine cortical discontinuities, narrow fracture lines, and subtle contour perturbations must be preserved at relatively high spatial resolution, whereas reliable classification also depends on progressively broader contextual integration across surrounding bone structure and anatomical arrangement. A hierarchical design therefore provides a natural mechanism for moving from local structural

encoding toward higher-level semantic abstraction without collapsing the representation prematurely.

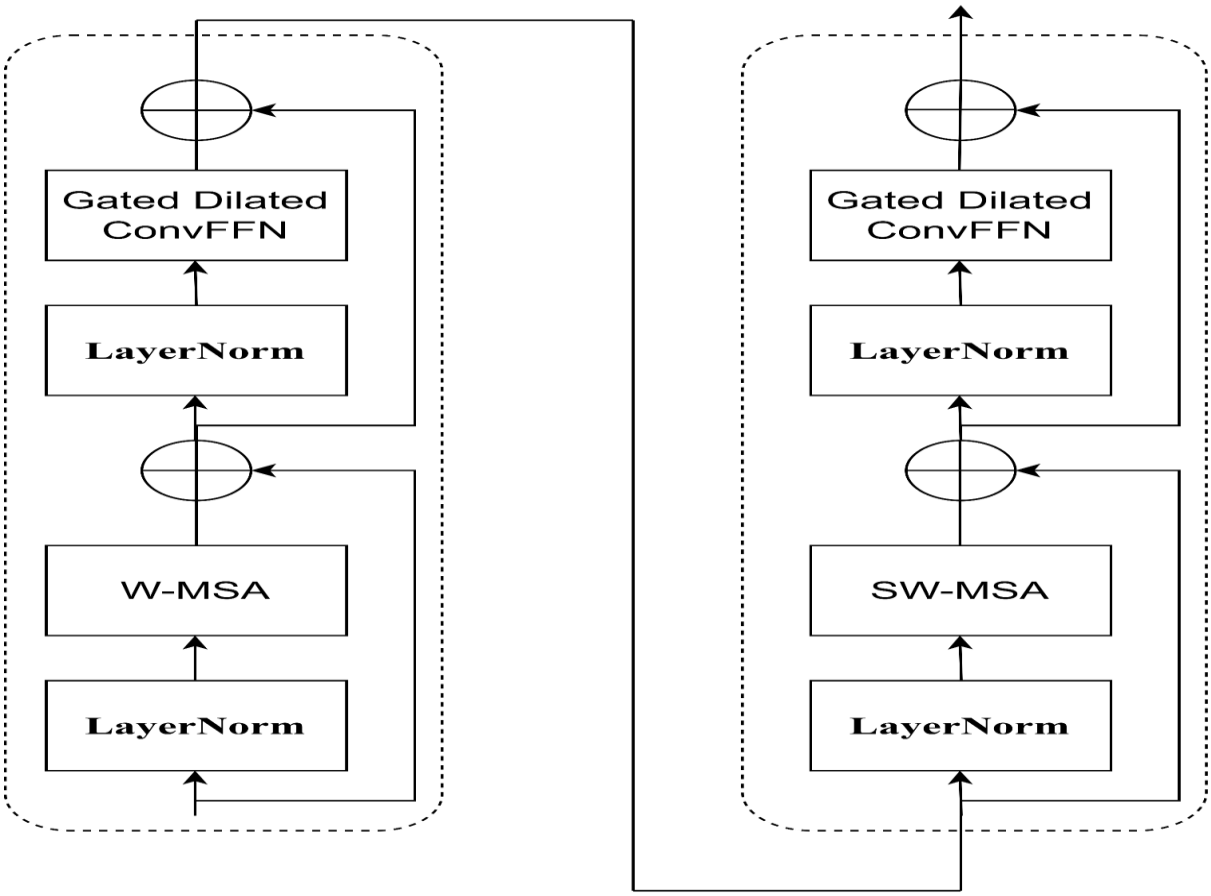

Figure 3 Internal structure of the residual CGTM block showing the W-MSA and SW-MSA sub-blocks with Gated Dilated ConvFFN refinement

Within this hierarchy, Stage 1 operates at a feature resolution of H/4 × W/4 × C and contains 2 residual CGTM blocks. Stage 2 reduces the spatial resolution to H/8 × W/8 × 2C and also contains 2 residual CGTM blocks. Stage 3 further transforms the representation at H/16 × W/16 × 4C using 6 residual CGTM blocks, while Stage 4 produces the highest-level representation at H/32 × W/32 × 8C with 2 residual CGTM blocks. This progression reflects a structured balance between detail preservation and abstraction. The earlier stages retain finer spatial information and are therefore better positioned to encode localized fracture-sensitive cues such as cortical interruption or local osseous irregularity. By contrast, the later stages operate on more compact but higher-dimensional feature maps, allowing the model to represent broader structural relationships and more abstract class-discriminative patterns associated with the overall radiographic appearance of fractured and non-fractured cases.

The transition between consecutive stages is achieved through adaptive patch merging, which performs learned hierarchical downsampling and feature consolidation. Rather than treating stage transitions as simple resolution reduction alone, the model uses this operation to jointly compress spatial structure and expand channel capacity, thereby enabling deeper stages to process richer semantic representations at lower spatial cost. The stage transition can be written as:

$$Z_{s+1}{}^{in} = PM_s(Z_s{}^{out})$$

where $Z_s{}^{out}$ denotes the output feature tensor of stage s, $Z_{s+1}{}^{in}$ denotes the input tensor to the next stage, and $PM_s(\cdot)$ represents the stage-specific adaptive patch merging operation. This formulation highlights that each inter-stage transition performs structured feature propagation rather than independent reinitialization. In the context of fracture classification, such hierarchical compression is useful because deeper contextual modeling becomes increasingly efficient when the spatial representation has been condensed while its discriminative capacity has been enhanced through channel expansion.

Inside each stage, the feature representation is refined through repeated application of the residual CGTM block. Let $Z_s{}^{in}$ and $Z_s{}^{out}$ denote the input and output of stage s, respectively. The repeated block composition can be expressed as:

$$Z_s^{\text{out}} = B_s^{\text{ns}}(Z_s^{in})$$

where B_s denotes the residual CGTM block and $n_s \in \{2, 2, 6, 2\}$ gives the number of block repetitions in stages 1 through 4. This repeated transformation is particularly meaningful for the third stage, which contains six residual CGTM blocks and therefore provides the deepest intermediate-level processing in the hierarchy. Such extended transformation depth is well motivated for fracture-sensitive representation learning because many radiographic fracture patterns require more than simple local detection. They often require iterative refinement, contextual reconciliation, and disambiguation from normal anatomical variation or projection-induced appearance changes. The deeper third stage is therefore positioned to serve as a key intermediate representation layer in which local structural evidence and broader contextual cues are repeatedly integrated before the final high-level abstraction of the fourth stage.

Taken together, the four-stage hierarchy, adaptive patch merging, and non-uniform block repetition pattern form a central methodological component of the proposed architecture. The design supports early preservation of localized structural detail, intermediate refinement of fracture-sensitive contextual relationships, and later abstraction into compact discriminative representations suitable for final classification. In this sense, the hierarchical organization is not merely a generic architectural choice, but a task-motivated mechanism for representing radiographic fracture patterns across multiple spatial and semantic scales. The final subsection describes how the resulting high-level representation is aggregated through attention pooling and mapped to the binary prediction space.

3.6 Attention pooling and final classification formulation

The output of Stage 4 constitutes the final high-level fracture-sensitive representation produced by the hierarchical backbone. At this point, the feature map encodes both localized structural evidence and broader contextual information accumulated through successive residual CGTM transformations and adaptive stage transitions. However, direct uniform aggregation of these features may not preserve the most informative responses equally well, particularly in radiographic fracture classification where decisive evidence may be concentrated in limited spatial regions rather than distributed evenly across the entire image. For this reason, the proposed model employs attention pooling to generate a weighted global representation in which informative final-stage tokens contribute more strongly to the downstream decision process than less relevant responses.

Let $\{z_i\}_{i=1}^{N}$ denote the set of final-stage feature vectors after Stage 4, where each z_i corresponds to a spatial token or location-specific representation in the last hierarchical feature map. Attention pooling computes a global feature vector g as a weighted combination of these representations:

$$g = \sum_{i=1}^{N} \alpha_i z_i \quad , \qquad \sum_{i=1}^{N} \alpha_i = 1$$

where α_i denotes the learned attention weight associated with the i-th token. This formulation allows the aggregation mechanism to emphasize discriminative fracture-relevant responses before classification rather than relying on uniform averaging across all spatial positions. Such learnable weighting is especially appropriate for radiographic fracture analysis because the visual evidence that determines the final class label may be localized to a narrow cortical region, a small discontinuity, or a restricted zone of structural irregularity. Attention pooling therefore complements the preceding hierarchical backbone by converting its spatially distributed high-level representation into a compact summary that remains sensitive to the most informative structural cues.The pooled representation is then passed to a fully connected layer, followed by softmax activation to produce the final class probabilities for the two-category classification task. The prediction function is written as:

$$\hat{y} = softmax(Wg + b)$$

where W and b denote the trainable weight matrix and bias vector of the final classifier, respectively, and $\hat{y}$ represents the predicted probability distribution over the Fractured and Not Fractured classes. Through this formulation, the network maps the weighted global feature representation into a binary decision space while preserving differentiability across the entire pipeline. The model is optimized end-to-end for bone fracture classification, allowing the attention pooling module, the classifier, and the hierarchical feature extractor to be learned jointly according to the target task.Taken together, the proposed architecture integrates residual CGTM blocks for structured local and shifted contextual interaction, Gated Dilated ConvFFN refinement for spatially enriched feature transformation, hierarchical multi-stage learning with adaptive patch merging for progressive abstraction, and attention-based final aggregation for discriminative classification. This coordinated design constitutes the central methodological contribution of the study, as it provides a fracture-sensitive classification framework that is better aligned with the structural and contextual demands of radiographic bone fracture recognition than a generic feature extraction pipeline

alone. With the architecture formulation established, the following section describes the training configuration, optimization strategy, and experimental evaluation protocol used to assess the proposed model.

3.7 Baseline models for comparative evaluation

Evaluating the proposed architecture requires comparison against reference models that are genuinely competitive in visual representation learning. Reporting classification accuracy in isolation is insufficient to establish whether the proposed design offers a meaningful modeling advantage; what is needed is a structured comparison against established backbones capable of learning both local structural features and broader contextual dependencies from radiographic images. To this end, two transformer-oriented architectures, MaxViT and EVA-02, were selected as baselines for binary bone fracture classification.MaxViT was chosen as the primary comparative baseline because it integrates convolutional inductive bias with multi-axis attention within a hierarchical processing framework, making it well suited to the dual requirements of fracture detection: identifying fine-grained cortical discontinuities and capturing broader anatomical context across the image. This combination makes MaxViT a more demanding benchmark than conventional convolutional networks, and it provides a meaningful reference point for assessing whether the proposed residual CGTM-based design offers additional discriminative value beyond what a strong general-purpose visual architecture already provides. In the present study, MaxViT achieved a mean training accuracy of 98.591% and a mean validation accuracy of 98.445%, with a test accuracy of 97.99% on the held-out evaluation set. These results confirm that MaxViT constitutes a genuinely competitive baseline rather than a weak comparator.

EVA-02 was included to provide a second reference point from a distinct architectural lineage. As a large-scale vision transformer backbone, EVA-02 represents a high-capacity approach to visual recognition that does not incorporate the same local window interaction, shifted contextual modeling, or gated dilated feed-forward refinement that characterise the proposed network. This architectural contrast is methodologically useful, since it permits an assessment of whether the explicit design choices in the proposed model confer a practical advantage in fracture-sensitive representation learning relative to a competitive backbone that operates under a different representational paradigm. In the fold-wise evaluation, EVA-02 achieved a mean training accuracy of 93.599% and a mean validation accuracy of 93.355%.

To ensure that observed performance differences reflect architectural properties rather than differences in experimental conditions, all baseline models were evaluated using the same binary classification task, identical data partitions, and the same 10-fold cross-validation protocol applied to the proposed method. This consistent experimental setup reduces the risk that any observed performance variation is attributable to differences in data exposure or evaluation formulation.Together, MaxViT and EVA-02 constitute a methodologically appropriate and technically credible reference set. MaxViT represents a high-performing hierarchical architecture that already integrates multi-scale visual modeling, while EVA-02 provides a complementary comparison from a large-scale transformer backbone with a distinct representational strategy. The combination of these two baselines enables a principled assessment of whether the proposed hierarchical CGTM-based network offers a genuine and consistent advantage for radiographic fracture classification.

3.8 Training configuration and implementation details

A carefully specified training configuration is essential for reproducible and methodologically fair evaluation in radiographic bone fracture classification. In this study, all compared models were trained under a consistent experimental framework so that performance differences could be interpreted primarily in relation to architectural design rather than inconsistent optimization or implementation choices. This is particularly important for fracture classification from X-ray images, where performance can be influenced by input standardization, optimization stability, regularization strength, and the interaction between model capacity and subtle structural image cues. For this reason, the proposed hierarchical transformer model and the comparative baselines, MaxViT and EVA-02, were implemented using the same core training and data-handling configuration.

All radiographs were represented as 3-channel inputs and resized to a fixed spatial resolution of 224 × 224 pixels. This input size was selected to maintain compatibility with modern vision backbones while preserving sufficient structural detail for binary fracture classification. Pixel intensities were rescaled to the range [0, 1], followed by channel-wise normalization using the ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. The use of this normalization scheme ensured a standardized input distribution aligned with pretrained feature extraction conventions commonly adopted in contemporary vision models. Importantly, the same input representation and normalization protocol were applied consistently to the proposed model, MaxViT, and EVA-02 so that the comparative evaluation remained methodologically aligned across architectures.

Optimization was performed using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay coefficient of 1×10^{-4}. All models were trained with a batch size of 16 over 100 epochs. The learning rate followed a cosine annealing

schedule to support stable convergence across the full training horizon, preceded by a linear warmup phase spanning the first 5 epochs. The warmup phase was incorporated to limit early optimization instability arising from large gradient updates at initialization, allowing the learning dynamics to adjust gradually before the cosine decay took effect. Mixed-precision training was enabled throughout to improve computational efficiency and reduce memory consumption, which is of practical importance when training transformer-based architectures on radiographic inputs processed at the target resolution. Gradient clipping was applied with a maximum gradient norm of 1.0 to prevent destabilizing parameter updates during training. A dropout rate of 0.30 was applied within the final classification head as a regularization mechanism, reducing the risk of overfitting to training-set-specific feature co-adaptations while preserving the discriminative capacity of the learned representation.

The study addressed a binary fracture-classification task using a two-class softmax output formulation. Model training was guided by a class-weighted cross-entropy loss, with class weights computed exclusively from the training split. This choice was used to account for mild distribution variation across the available data subsets while avoiding leakage of information from validation or test data into the optimization process. The same classification objective and the same core training configuration were applied to the proposed residual CGTM-based hierarchical model and to both baseline architectures, MaxViT and EVA-02. This consistent protocol was important for fair comparative evaluation because it limited confounding due to architecture-specific tuning differences and ensured that the observed performance patterns reflected the representational properties of the models under a common training regime.

The implementation environment was designed to support reproducibility and practical experimental execution. All experiments were implemented in Python 3.10 using PyTorch 2.1 and torchvision 0.16 with CUDA 12.1 support. The experimental environment was based on Ubuntu 22.04 LTS. Training and evaluation were conducted on a single NVIDIA GPU with at least 12 GB VRAM, accompanied by a modern multi-core processor and sufficient system memory for batch-based radiographic classification experiments. To promote controlled experimentation and reproducibility, the random seed was fixed to 42 throughout the training pipeline.

4.1 Results of the proposed model

The proposed model demonstrated consistently strong performance across both fold-wise evaluation and held-out testing. Over the 10-fold training and validation procedure, it achieved a mean training accuracy of 99.384% and a mean validation accuracy of 99.276%, indicating highly stable optimization behavior across repeated data partitions. On the held-out test set, the model attained an accuracy of 99.25%, together with a ROC-AUC of 99.99% and an average precision of 99.99% on the precision-recall curve. These results collectively indicate that the proposed architecture maintained strong generalization on unseen radiographic samples without evidence of severe overfitting under the adopted training regime.

Table 1 presents the fold-wise training and validation accuracies across all 10 evaluation folds. Training accuracy remained uniformly high, ranging from 99.19% to 99.56%, while validation accuracy ranged from 99.08% to 99.47%. This narrow interval reflects limited fold-to-fold fluctuation despite natural variation across sampled partitions. The corresponding standard deviations were 0.123 for training accuracy and 0.127 for validation accuracy, confirming highly consistent performance across repeated evaluation folds. The small difference between mean training and mean validation accuracy further suggests controlled generalization behavior. Together, the fold-wise pattern indicates that the proposed architecture learned stable fracture-sensitive representations in a reproducible manner rather than benefiting from a single favorable partition.

Table 1 Fold-wise training and validation accuracy of the proposed model across 10 evaluation folds

Fold	Training	Validation
1	99.48	99.36
2	99.41	99.29
3	99.27	99.18

4	99.52	99.44
5	99.34	99.21
6	99.19	99.08
7	99.45	99.33
8	99.38	99.26
9	99.24	99.14
10	99.56	99.47

Figure 4 presents the confusion matrix obtained on the held-out test set. The model produced only three misclassifications out of 399 evaluated samples. All 199 Not Fractured cases were correctly classified with zero false positives, while 197 of 200 Fractured cases were correctly identified and 3 were misclassified as Not Fractured. This error distribution is consistent with the reported test accuracy of 99.25%, corresponding to 396 correct predictions. The matrix exhibits a strongly diagonal structure with minimal off-diagonal entries, indicating near-complete class separation and a highly localized error pattern confined to a small subset of Fractured cases.

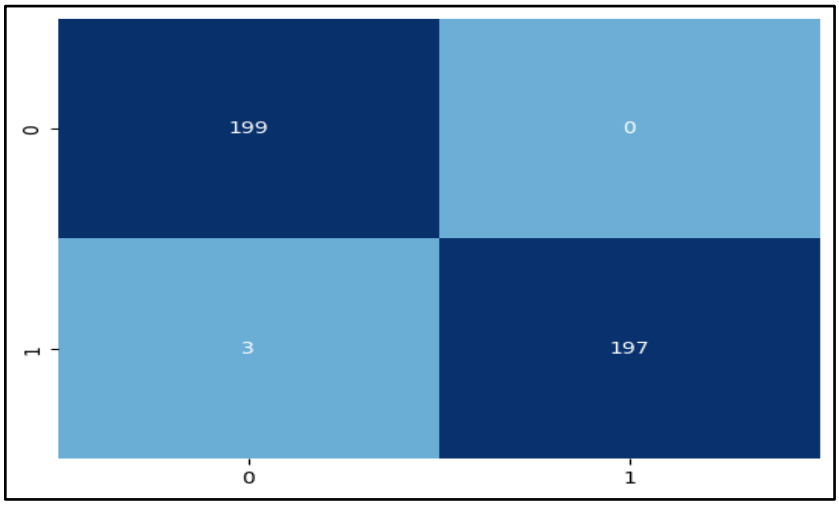

Figure 4:Confusion matrix of the proposed model on the held-out test set

Figures 5 and 6 present the ROC curve and precision-recall curve evaluated on the held-out test set, respectively. The ROC curve is concentrated near the upper-left corner of the plot, reflecting excellent threshold-independent class separability, and yielded an AUC of 99.99%. The precision-recall curve remained close to the upper boundary across nearly the full recall range, producing an average precision of 99.99%. This behavior confirms that the classifier maintained very high precision across a wide range of operating thresholds, which is particularly informative in settings where a single accuracy value may obscure discrimination characteristics at varying decision boundaries. Together, these curves demonstrate that the strong held-out accuracy reflects favorable ranking behavior across thresholds rather than performance at a single operating point.

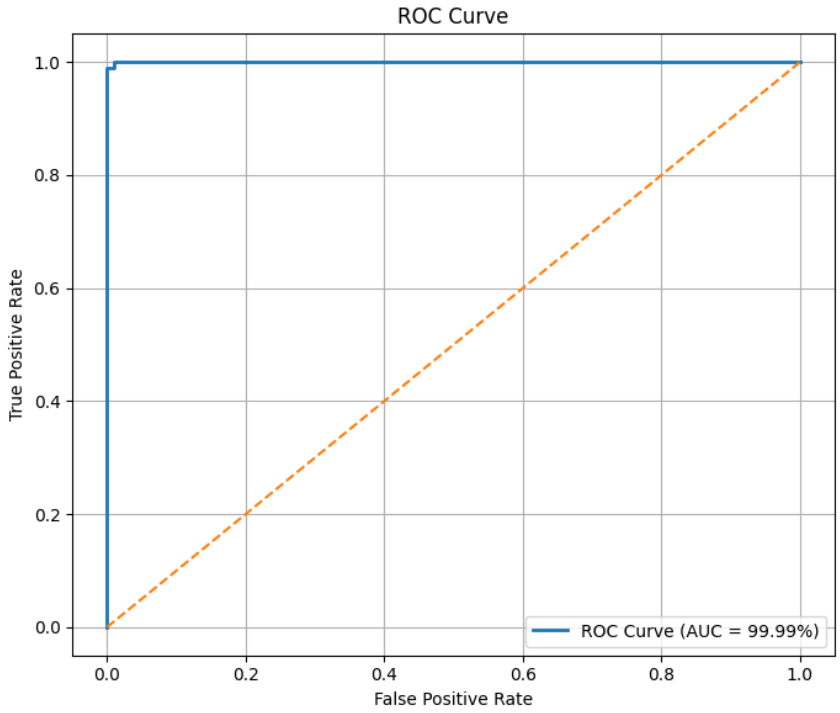

Figure 5:ROC curve of the proposed model on the held-out test set (AUC = 99.99%)

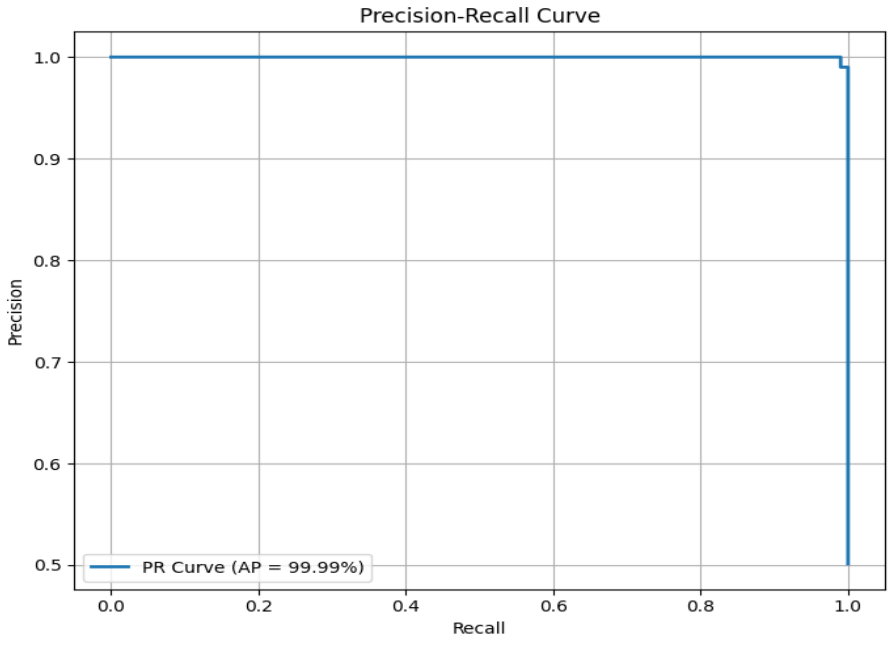

Figure 6 :Precision-recall curve of the proposed model on the held-out test set (AP = 99.99%)

Figure 7 presents the class-wise classification report on the held-out test set. For the Fractured class, precision was 98.51%, recall was 100.00%, and F1-score was 99.25%. For the Not Fractured class, precision was 100.00%, recall was 98.50%, and F1-score was 99.24%. The macro-average precision, recall, and F1-score were 99.26%, 99.25%, and 99.25%, respectively. These values indicate that both classes were recognized with very high and nearly symmetric reliability. The model achieved complete recall for the Fractured class, meaning no fractured case was missed in the held-out evaluation, while maintaining perfect precision for the Not

Fractured class. The near-identical F1-scores across both categories confirm that the classifier did not exhibit a systematic bias toward either class.

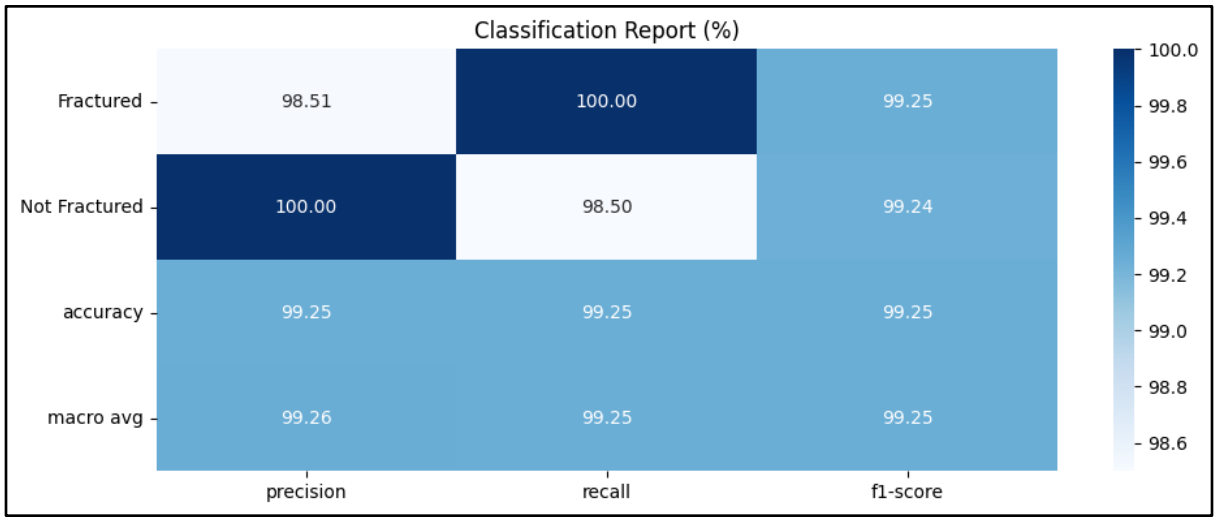

Figure 7 :Classification report of the proposed model on the held-out test set

4.2 Results of the MaxViT model

The MaxViT model achieved strong and stable performance across both fold-wise validation and held-out evaluation, confirming its suitability as a competitive benchmark for binary bone fracture classification from radiographic images. Over the 10-fold training and validation procedure, MaxViT obtained a mean training accuracy of 98.591% and a mean validation accuracy of 98.445%. On the held-out test set, the model achieved an accuracy of 97.99%. These results indicate that MaxViT maintained consistent learning behavior across repeated folds while also demonstrating strong generalization on unseen radiographic samples.Table 2 presents the fold-wise training and validation accuracies across all 10 evaluation folds. Training accuracy ranged from 98.31% to 98.77%, while validation accuracy ranged from 98.18% to 98.66%. This narrow interval indicates limited fold-to-fold variability and suggests that model performance was not driven by a small number of unusually favorable partitions. The corresponding standard deviations were 0.147 for training accuracy and 0.151 for validation accuracy, supporting the stability of the learned representation across repeated evaluation runs. The difference between mean training and mean validation accuracy remained small, indicating controlled generalization behavior without obvious overfitting under the adopted training protocol.

Table 2 Fold-wise training and validation accuracy of the MaxViT model across 10 evaluation folds

Fold	Training	Validation
1	98.74	98.61
2	98.68	98.52
3	98.55	98.39
4	98.77	98.66
5	98.49	98.34
6	98.31	98.18

7	98.63	98.47
8	98.58	98.42
9	98.44	98.29
10	98.72	98.57

Figure 8 presents the confusion matrix obtained on the held-out test set. MaxViT correctly classified 193 of 199 Not Fractured samples and 198 of 200 Fractured samples. The residual errors comprised 6 Not Fractured cases misclassified as Fractured and 2 Fractured cases misclassified as Not Fractured. This bidirectional error pattern is consistent with the reported held-out accuracy of 97.99%, corresponding to 391 correct predictions out of 399 evaluated samples. The off-diagonal entries indicate that a limited degree of class confusion persisted in both directions, reflecting the visual ambiguity between certain Fractured and Not Fractured radiographs in the held-out evaluation.

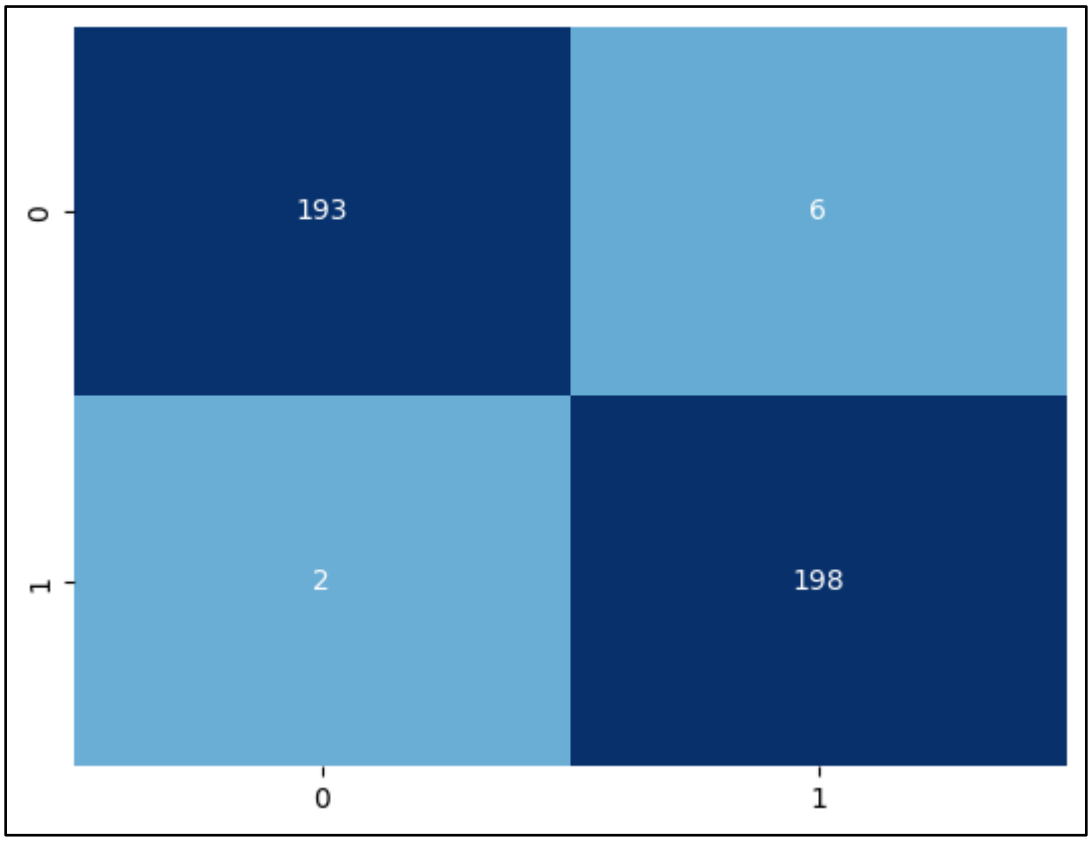

Figure 8 :Confusion matrix of MaxViT on the held-out test set

Figure 9 presents the class-wise classification report on the held-out test set. For the Not Fractured class, precision was 98.97%, recall was 96.98%, and F1-score was 97.96%, based on a support of 199 samples. For the Fractured class, precision was 97.06%, recall was 99.00%, and F1-score was 98.02%, based on a support of 200 samples. The overall accuracy was 97.99%, and the macro-average precision, recall, and F1-score were 98.01%, 97.99%, and 97.99%, respectively. The precision-recall tradeoff differed modestly between the two classes: Not Fractured showed higher precision but lower recall, while Fractured showed higher recall but slightly lower precision. Despite this asymmetry, the F1-scores across both classes remained closely matched, and the overall profile indicates a balanced performance without a systematic bias toward either category.

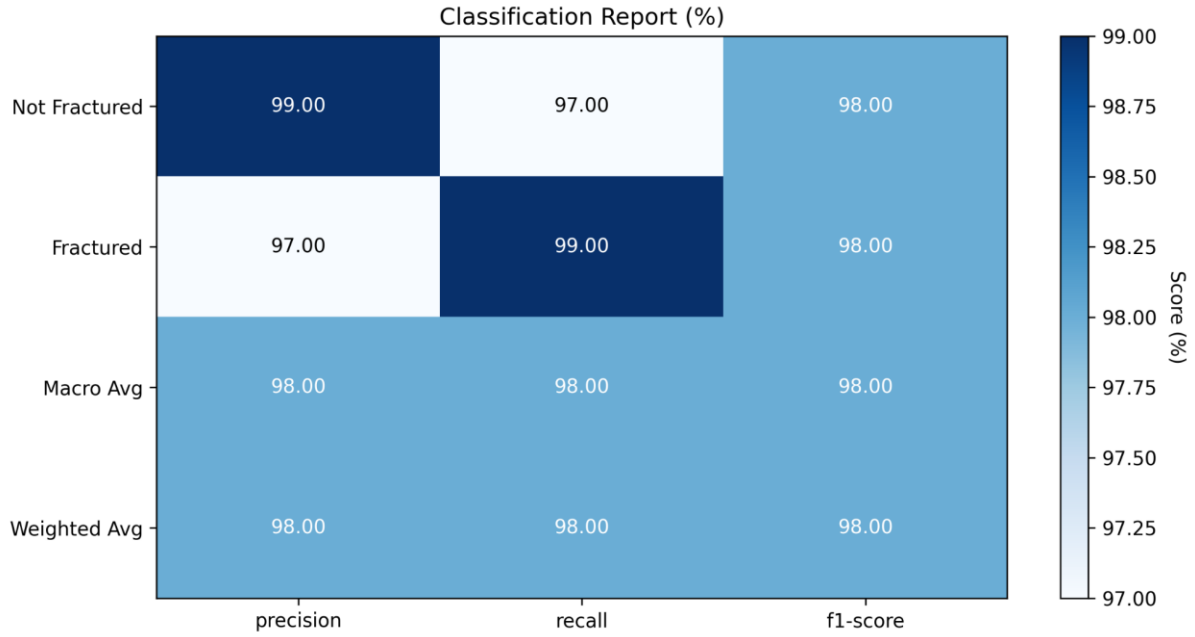

Figure 9 :Classification report of MaxViT on the held-out test set

4.3 Results of the EVA-02 model

The EVA-02 model demonstrated stable fold-wise training and validation behavior across the 10-fold evaluation procedure. Averaged over all folds, the model achieved a mean training accuracy of 93.599% and a mean validation accuracy of 93.355%. Although these values are lower than those observed for the strongest models in the study, they still indicate a consistent optimization pattern under the defined experimental setting. From a results perspective, EVA-02 therefore serves as a meaningful transformer-based reference model for binary bone fracture classification from radiographic images.The fold-wise training results reported in Table 3 show that EVA-02 maintained relatively compact performance across all 10 folds. Training accuracy ranged from a minimum of 93.21% to a maximum of 93.89%, indicating that the model converged to a similar level of performance across repeated partitions of the data. The corresponding training standard deviation was 0.220, which further supports the observation of limited fold-to-fold variability. This numerically constrained range suggests that the training behavior of EVA-02 was reproducible and not strongly dependent on a small subset of favorable folds. In other words, the model exhibited stable optimization dynamics rather than irregular or highly volatile learning across the repeated runs.

Table 3 Fold-wise training and validation accuracy of the EVA-02 model across 10 evaluation folds

Fold	Training	Validation
1	93.82	93.61
2	93.74	93.48
3	93.57	93.26

4	93.89	93.70
5	93.46	93.19
6	93.21	92.98
7	93.68	93.43
8	93.51	93.27
9	93.34	93.08
10	93.77	93.55

A similarly consistent pattern was observed for validation performance. The validation accuracy ranged from 92.98% to 93.70% across the 10 folds, with a validation standard deviation of 0.236. This range remained compact, indicating that generalization behavior was also stable under cross-validation. Moreover, the difference between the mean training accuracy and mean validation accuracy was small, which suggests consistent train-validation behavior and no obvious sign of severe overfitting within the adopted training regime. The validation pattern therefore indicates that EVA-02 was able to retain a relatively steady performance level across different data partitions, even though its absolute accuracy remained below that of the strongest architectures evaluated in this work.

4.4 Statistical analysis and comparative evaluation

To assess comparative model behavior under a consistent experimental setting, statistical analysis was conducted using four complementary views: descriptive statistics of ten-fold validation accuracy, descriptive statistics of fold-wise training accuracy, pairwise significance testing with effect-size analysis, and ranking by mean validation accuracy. This structure allows the comparison to move beyond single-point performance reporting and instead examine central tendency, variability, pairwise separation, and overall ordering of the evaluated models. Across these analyses, the proposed model achieved the strongest mean validation accuracy and the most favorable overall statistical profile among the compared architectures.

Table 4 summarizes the descriptive statistics of ten-fold validation accuracy and therefore reflects model behavior on validation folds rather than held-out test evaluation. The proposed model achieved the highest mean validation accuracy at 99.276%, followed by MaxViT at 98.445% and EVA-02 at 93.355%. The associated standard deviations were 0.127, 0.151, and 0.236, respectively, indicating that the proposed model also showed the lowest fold-to-fold dispersion among the three methods. Its 95% confidence interval ranged from 99.185% to 99.367%, which was both narrow and centered at the highest performance level. MaxViT showed a 95% confidence interval of 98.337% to 98.553%, while EVA-02 showed a wider interval of 93.186% to 93.524%, consistent with its lower mean and larger standard deviation. The minimum and maximum validation accuracies followed the same ordering: the proposed model ranged from 99.08% to 99.47%, MaxViT from 98.18% to 98.66%, and EVA-02 from 92.98% to 93.70%. Taken together, Table 4 indicates that the proposed model not only achieved the best average validation performance, but did so with highly stable fold-wise behavior.

Table 4 Descriptive statistics of ten-fold validation accuracy for the proposed model and baseline models

Model	Mean (%)	SD	95% CI	Min	Max
Proposed Model	99.276	0.127	99.185 to 99.367	99.08	99.47
MaxViT	98.445	0.151	98.337 to 98.553	98.18	98.66
EVA-02	93.355	0.236	93.186 to 93.524	92.98	93.70

Table 5 provides the corresponding descriptive statistics for fold-wise training accuracy. Here again, the proposed model achieved the highest mean value at 99.384%, followed by MaxViT at 98.591% and EVA-02 at 93.599%. The standard deviations remained low across all three models, with values of 0.123, 0.147, and 0.220, respectively, indicating stable optimization over repeated folds. The 95% confidence interval for the proposed model ranged from 99.296% to 99.472%, while MaxViT ranged from 98.486% to 98.696% and EVA-02 from 93.442% to 93.756%. The minimum and maximum fold-wise training accuracies were 99.19% and 99.56% for the proposed model, 98.31% and 98.77% for MaxViT, and 93.21% and 93.89% for EVA-02. These training summaries indicate that all three models converged consistently under the shared optimization setting, but the proposed model retained the strongest overall optimization profile.

Table 5 Descriptive statistics of fold-wise training accuracy for the proposed model and baseline models

Model	Mean (%)	SD	95% CI	Min	Max
Proposed Model	99.384	0.123	99.296 to 99.472	99.19	99.56
MaxViT	98.591	0.147	98.486 to 98.696	98.31	98.77
EVA-02	93.599	0.220	93.442 to 93.756	93.21	93.89

A direct comparison of Tables 1 and 2 provides further insight into generalization behavior. For the proposed model, the mean training accuracy was 99.384% and the mean validation accuracy was 99.276%, representing only a small reduction from training to validation. A similarly modest gap was observed for MaxViT, with 98.591% training and 98.445% validation, and for EVA-02, with 93.599% training and 93.355% validation. These relatively small differences suggest controlled generalization behavior across all three architectures, without an obvious indication of severe overfitting in the fold-wise summaries. However, the proposed model distinguished itself by maintaining both the highest training mean and the highest validation mean, while also preserving the lowest variability among the compared models in the validation setting. MaxViT remained a strong competitive baseline with high and stable performance, whereas EVA-02 showed a consistently lower but still internally stable result profile.

Table 6 reports pairwise statistical comparison using mean differences, confidence intervals, paired tests, Wilcoxon p-values, and Cohen's dz effect sizes. The reported comparisons indicate statistically significant differences across all three model pairs. For the proposed model versus MaxViT, the mean difference was +0.831 percentage points, with a 95% confidence interval of 0.792 to 0.870. For the proposed model versus EVA-02, the mean difference was +5.921 percentage points, with a 95% confidence interval of 5.832 to 6.010. For MaxViT versus EVA-02, the mean difference was +5.090 percentage points, with a 95% confidence interval of 5.027 to 5.153. All of these intervals are strictly positive and do not cross zero, which supports a consistent directional advantage of the higher-performing model in each pairwise comparison. The Wilcoxon p-value was reported as 0.001953 for all three comparisons, further supporting statistical separation under a non-parametric paired analysis. One numerical issue should, however, be acknowledged explicitly: in the first comparison, the reported paired t-statistic for the proposed model versus MaxViT was 48.076, whereas the corresponding p-value was listed as 3.65×10^{-2}. This p-value appears numerically inconsistent

with a test statistic of that magnitude and may reflect a typographical reporting issue. Even with that inconsistency noted, the overall evidence in Table 6, including the positive confidence interval, the Wilcoxon result, the very large reported effect size, and the consistent fold-wise superiority, still supports a meaningful advantage of the proposed model over MaxViT and EVA-02 within the reported experimental setting.

Table 6 Pairwise statistical comparison of model performance using paired significance tests and effect size analysis

Comparison	Mean difference(%)	95% CI	Paired t	p-value	Wilcoxon p	Cohen's dz	Significant
Proposed vs MaxViT	0.831	0.792 to 0.870	48.076	3.65×10^{-2}	0.001953	15.203	Yes
Proposed vs EVA-02	5.921	5.832 to 6.010	149.756	1.34×10^{-16}	0.001953	47.357	Yes
MaxVit vs EVA-02	5.090	5.027 to 5.153	182.252	3.65×10^{-12}	0.001953	57.633	Yes

The reported Cohen's dz values are also informative from a practical standpoint. The proposed model versus MaxViT comparison yielded dz = 15.203, the proposed model versus EVA-02 yielded dz = 47.357, and MaxViT versus EVA-02 yielded dz = 57.633. These are extremely large effect sizes in the context of paired fold-wise comparisons and indicate that the observed differences were not only statistically detectable but also substantial in magnitude within the present evaluation framework. The largest separation was observed between the proposed model and EVA-02, while the proposed model also maintained a clear and non-trivial advantage over MaxViT. Although such large values should be interpreted within the bounded and highly accurate performance regime of this specific experiment, they nonetheless reinforce the conclusion that the ordering among the models was not marginal.

Table 7 provides a compact ranking based on mean validation accuracy. Under this criterion, the proposed model was ranked first with a mean validation accuracy of 99.276%, MaxViT was ranked second with 98.445%, and EVA-02 was ranked third with 93.355%. This ranking is fully consistent with the descriptive validation statistics in Table 5 and with the direction of the pairwise comparisons in Table 7. As a summary device, the ranking table is useful because it condenses the comparative performance landscape into a simple ordinal view while remaining aligned with the more detailed statistical evidence reported elsewhere.

Table 7 Ranking of the proposed model and baseline models based on mean validation accuracy

Rank	Model	Mean Validation (%)
1	Proposed Model	99.276
2	MaxViT	98.445
3	EVA-02	93.355

4.5 Interpretability and error analysis of the proposed model

To examine the decision-making behaviour of the proposed model and identify the sources of residual classification errors, two complementary analyses were performed: a false-positive error analysis using representative misclassified samples from the held-out test set, and a Grad-CAM-based interpretability study using correctly classified fracture cases.

Figure 10 presents three false-positive cases in which the model assigned the Fractured label to radiographs whose ground-truth class is Not Fractured. The three cases are associated with Not Fractured class probabilities of 0.0481, 0.4095, and 0.0098, respectively, indicating that the model expressed substantially different levels of certainty across these errors. The first and third cases represent high-confidence misclassifications, where the model's posterior probability for the Fractured class exceeded 0.95 and 0.99, respectively, suggesting that the relevant image regions strongly activated fracture-discriminative features despite the absence of a true fracture. The second case is comparatively more ambiguous, with a Not Fractured probability of 0.4095, placing the prediction near the decision boundary and indicating that the model was less committed to the incorrect label. Visual inspection of the radiographs in Figure 10 reveals that all three cases exhibit subtle structural characteristics, including trabecular irregularities, cortical density variations, and overlapping bony contours, that can visually resemble fracture-associated image patterns. Such presentations are consistent with well-documented sources of confusion in radiographic fracture detection, where imaging geometry, patient positioning, or pre-existing degenerative changes introduce local discontinuities that superficially mimic fracture lines. The concentration of false positives in this structural category suggests that residual model errors are not random but reflect a specific and diagnostically plausible failure mode.

Figure 10:False-positive examples produced by the proposed model during bone fracture classification

Figure 11 presents Grad-CAM visualizations for two correctly classified Fractured cases, constructed to assess whether the proposed model attends to diagnostically relevant anatomical regions during inference. For each case, Figure 11 displays a triplet composed of the original radiograph, the corresponding Grad-CAM activation heatmap, and the overlay of the heatmap on the input image. The two selected cases achieved predicted Fractured class probabilities of 0.9948 and 1.0000, confirming that both were classified with high confidence. The Grad-CAM heatmaps in Figure 11 reveal that activation gradients are concentrated in the perilesional region of the bone, with the highest response localised to the area surrounding the fracture site. This activation pattern aligns with the primary visual criteria employed in radiological fracture diagnosis, including cortical discontinuity, periosteal reaction, and disruption of trabecular architecture. The spatial coherence of the highlighted regions across both cases indicates that the model does not rely on peripheral image artefacts, background tissue, or non-diagnostic structures, but instead learns representations that are directly grounded in the bone morphology relevant to fracture identification. The convergence of high classification confidence with anatomically plausible Grad-CAM activations provides evidence that the proposed model captures clinically meaningful features and does not achieve discriminative performance through shortcut correlations in the training data.

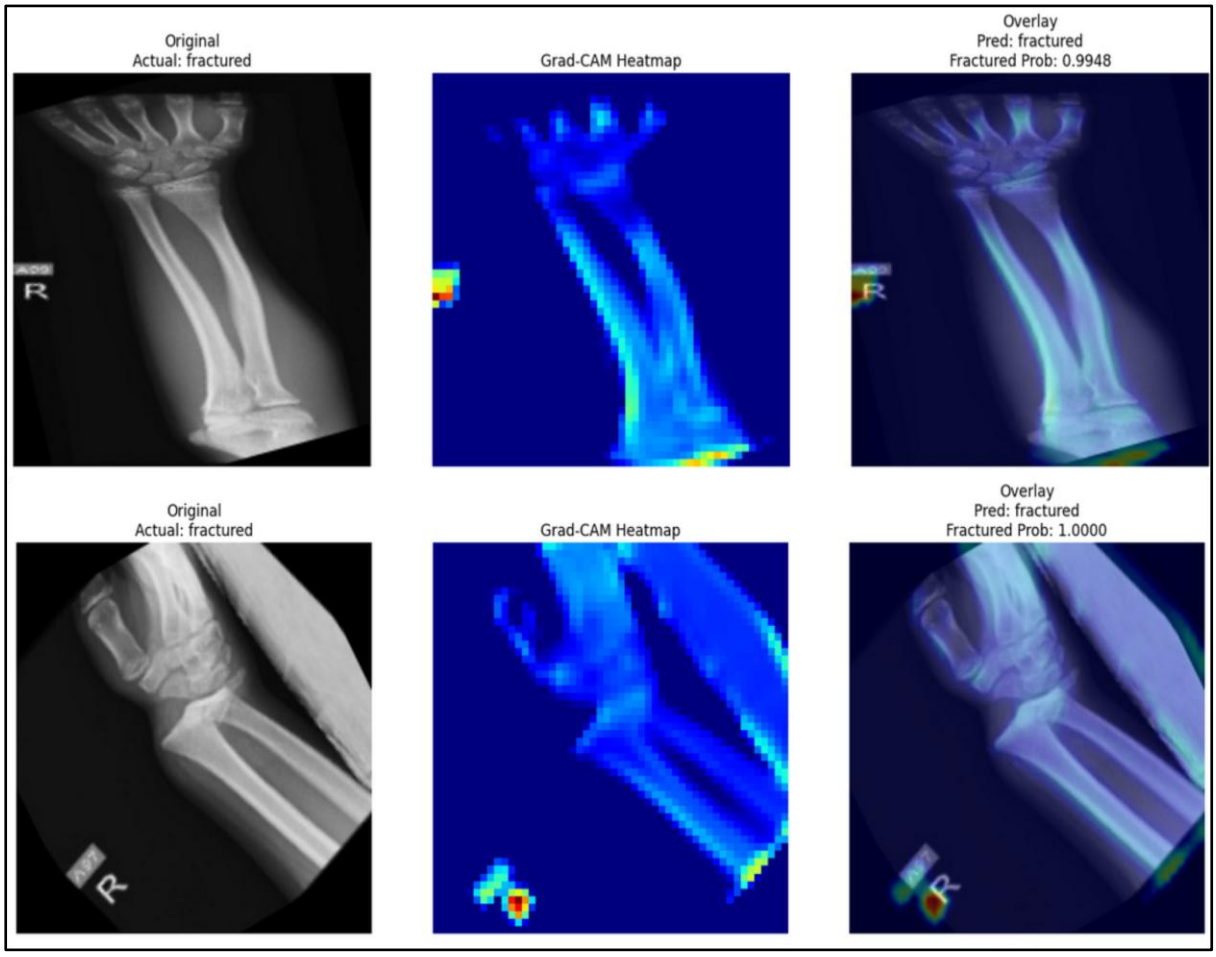

Figure 11:Grad-CAM visualization of correctly classified fractured radiographs produced by the proposed model

Each row shows the original radiograph, the Grad-CAM heatmap, and the corresponding overlay for a correctly classified Fractured case. The predicted Fractured probabilities are 0.9948 and 1.0000, indicating highly confident positive predictions. In both examples, the activation is concentrated in localized osseous regions of the wrist and distal forearm rather than being uniformly distributed across the full image. This figure provides qualitative evidence that the proposed model relies on structurally relevant image regions during fracture-positive prediction, while not implying exact clinical lesion localization.

5. Limitations and future work

The findings reported in this study should be interpreted in the context of several methodological limitations that constrain the scope of the current conclusions and define productive directions for subsequent research.The most consequential limitation is the reliance on a single publicly available dataset. While the dataset provided a sufficient foundation for model development, internal validation, and comparative evaluation, it represents a narrow sample of the radiographic variability encountered in routine clinical practice. Fracture appearance depends on anatomical location, fracture morphology, projection angle, superimposed structures, image acquisition quality, and patient-specific bone characteristics, and the present dataset cannot be assumed to capture this full spectrum of variation. Furthermore, the evaluation did not include external validation on an independent cohort collected under different imaging protocols or from a distinct patient population. Without such cross-dataset assessment, the degree to which the reported performance generalises beyond the present experimental setting remains uncertain. Future work should therefore prioritise multi-centre data collection and cross-institutional validation to produce a more reliable estimate of model robustness under realistic clinical variability.A second limitation concerns the scope of the classification problem. The proposed model addresses binary fracture detection and does not extend to fracture type identification, severity grading, skeletal region specificity, or fracture localisation within the image. These are clinically relevant distinctions that binary classification cannot capture. Extending the framework to support multi-class fracture type categorisation, region-specific models, or fracture detection formulations would substantially increase clinical utility and would

require architectural and training adaptations beyond those described in the present study.A third limitation relates to model interpretability. The Grad-CAM visualisations presented in this study offer qualitative evidence that the proposed model attends to perilesional bone regions during fracture-positive predictions. However, these visualisations are spatially coarse, are derived from gradient aggregation rather than explicit spatial supervision, and were not assessed by radiologists for diagnostic correspondence. They should therefore be understood as indicative of general attention behaviour rather than as precise fracture localisation or clinically validated reasoning. A rigorous interpretability evaluation would require structured radiologist review, spatial agreement metrics against ground-truth fracture annotations, and systematic failure-mode characterisation across a representative case spectrum.A fourth limitation is the absence of uncertainty quantification. The proposed model produces point estimates of class probability without any accompanying measure of prediction confidence or epistemic uncertainty. In a clinical context, the ability to identify cases for which the model is uncertain and to flag these for expert review is as important as classification accuracy on confident predictions. Integrating uncertainty estimation mechanisms, such as Monte Carlo dropout or deep ensemble methods, would improve the practical utility of the framework and support more responsible deployment in high-stakes settings.

Future research should address these limitations through several complementary directions. Validation on larger and more diverse datasets, including cases collected from multiple clinical sites with varying imaging equipment and acquisition protocols, would provide a more stringent test of the generalisability of the proposed architecture. Extension to fracture type classification and fracture localisation would broaden the clinical relevance of the framework. Multimodal integration, incorporating computed tomography or magnetic resonance imaging alongside radiographic input, could further enrich the representational basis for fracture assessment, particularly in cases where plain radiography is diagnostically ambiguous. Incorporating uncertainty quantification into the model output and conducting prospective clinical validation in collaboration with radiologists would constitute important steps toward establishing the practical applicability of the proposed approach beyond the controlled experimental setting of the present study.

6. Conclusion

This study presented a hierarchical transformer-based architecture for binary bone fracture classification from radiographic images, motivated by the inherent difficulty of the task: fracture evidence is often subtle, spatially localised, and simultaneously dependent on fine osseous detail and broader anatomical context. The proposed network addressed this dual requirement through residual CGTM blocks that integrate window-based multi-head self-attention, shifted-window multi-head self-attention, and Gated Dilated ConvFFN refinement, supported by adaptive patch merging across four hierarchical stages and attention pooling for discriminative feature aggregation at the classification head.Evaluated under a 10-fold cross-validation protocol, the proposed model achieved a mean training accuracy of 99.384% and a mean validation accuracy of 99.276%. On the held-out test set, it reached an accuracy of 99.25%, a ROC-AUC of 99.99%, and an average precision of 99.99%, outperforming both MaxViT and EVA-02 under identical experimental conditions. The Grad-CAM analysis further demonstrated that high-confidence fracture-positive predictions were supported by activation gradients concentrated in perilesional bone regions, consistent with the diagnostically relevant structural features targeted by the proposed design rather than background image content or peripheral tissue.These findings collectively support the conclusion that the proposed hierarchical CGTM-based architecture learns fracture-discriminative representations effectively within the binary radiographic classification setting. The combination of local window attention, shifted contextual interaction, and gated dilated convolutional refinement within a single unified framework appears to be well matched to the representational demands of radiographic fracture analysis, and the results suggest that such task-oriented architectural design can provide a meaningful advantage over strong general-purpose visual backbones.

Author Contributions: Md Fakrul Alam conceived and designed the study, developed the proposed hierarchical transformer architecture incorporating residual CGTM blocks and gated dilated convolutional refinement, prepared and processed the radiographic image dataset, implemented and trained the models, conducted the comparative experiments and 10-fold cross-validation, performed the statistical and performance analyses, interpreted the results, prepared the figures and tables, and drafted, reviewed, and approved the final manuscript.

Data Availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Clinical Trial Number: Not applicable.

Human Ethics and Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Consent to Participate: Not applicable.

Ethics: Not applicable.

Funding:This research received no external funding.

REFERENCES

1.Almigdad, Ahmad, Ayman Mustafa, Sattam Alazaydeh, Mu'men Alshawish, Mohammad Bani Mustafa, and Hamza Alfukaha. "Bone fracture patterns and distributions according to trauma energy." *Advances in Orthopedics* 2022, no. 1 (2022): 8695916.
2.McRae, Ronald, and Max Esser. *Practical Fracture Treatment E-Book: Practical Fracture Treatment E-Book*. Elsevier Health Sciences, 2008.
3.Akinmade, Akinola, Innocent Ikem, Oluwagbemiga Ayoola, Elkanah Orimolade, and Adeolu Adeyeye. "Comparing ultrasonography with plain radiography in the diagnosis of paediatric long-bone fractures." *International Orthopaedics* 43, no. 5 (2019): 1143-1153.
4.Schwarzenberg, Peter, Salim Darwiche, Richard S. Yoon, and Hannah L. Dailey. "Imaging modalities to assess fracture healing." *Current osteoporosis reports* 18, no. 3 (2020): 169-179.
5.Parvez, Rasel, Md Anwar Hossain, Showrov Azam, AKM Bahalul Haque, Sadman Sadik Khan, and Sadekur Rahman. "From Classical to Colloquial: Leveraging LLMs for Sadhu–Cholit Register Identification in Bangla." In *International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, pp. 473-486. Atlantis Press, 2026.
7.Hossain, Md Sahid, Kallol Chakraborty Shekhor, Md Anwar Hossain, and Md Abedur Rahman. "Diagnosis Everywhere: Lightweight AI That Detects Disease from Scans and Records on Limited Hospital Hardware." *Frontiers in Computer Science and Artificial Intelligence* 2, no. 2 (2023): 82-94.
8.Chittajallu, Sai Manvitha, Navya Lakshmi Deepthi Mandalaneni, Dhanush Parasa, and Shahana Bano. "Classification of binary fracture using CNN." In *2019 global conference for advancement in technology (GCAT)*, pp. 1-5. IEEE, 2019.
9.Tanzi, Leonardo, Enrico Vezzetti, Rodrigo Moreno, and Sandro Moos. "X-ray bone fracture classification using deep learning: a baseline for designing a reliable approach." *Applied Sciences* 10, no. 4 (2020): 1507.
10.Shekhor, Kallol Chakraborty, Md Sahid Hossain, Md Abedur Rahman, and Md Anwar Hossain. "Multi-Scale Attention-Fusion Classification with an LLM-Driven Clinical Recommendation Assistant for Brain Tumor MRI." *Journal of Medical and Health Studies* 4, no. 2 (2023): 121-139.
11.Al-Taki, Muhyeddine, and Imad S. Nahle. "Bones and Fractures." In *Hamilton Bailey's Physical Signs*, pp. 147-165. CRC Press, 2016.
12.Dogra, A.S., Ruiz, A.L., Thompson, N.S. and Nolan, P.C., 2000. Dia-metaphyseal distal tibial fractures—treatment with a shortened intramedullary nail: a review of 15 cases. *Injury*, *31*(10), pp.799-804.
13.Rahman, M.S., Shekhor, K.C., Hossain, M.A. *et al.* Self supervised iBOT vision transformer framework for data efficient white blood cell classification. *Discov Artif Intell* **6**, 466 (2026). https://doi.org/10.1007/s44163-026-01185-4
14.Hossain, M. S., Islam, M. A., Ali, M., Chy, &. S., Md Shahdat Hossain (2025). "Blockchain Ai for Supply Chain Transparency Integrating Blockchain with AI to Detect Fraud and Ensure Transparency Across Multi-Tier Global Supply Chains", Business and Social Sciences, 3(1), 1-12, 10457
16.Cheng, Jun, Yuanyuan Li, Zihao Zhao, Dongyan Li, Longzhu Zhong, Shuoxuan Ding, and Xiaopeng Yan. "A hybrid transformer framework with receptive field context aggregation and learnable basis functions for road crack detection." *Measurement Science and Technology* 36, no. 8 (2025): 085406.
17.Wang, Lina, Jiayu Zhang, Ziqing Wang, Shuheng Liu, and Jiangfeng She. "Hierarchical Attention and Dynamic Consistency Filtering for Robust Aerial Multi-view Stereo." *Knowledge-Based Systems* (2025): 114973.
18.Wang, Zezhou, and Sandra Puentes. "Enhancing Bone Fracture Detection with Large Kernel Attention Modules." In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1-7. IEEE, 2025.
19.Chowdhury, M. S. I., Chy, K. S., Islam, M. A., & Chowdhury, M. N. I. (2023). Integrating Blockchain and Data Analytics to Strengthen Financial Traceability and Anti-Fraud Controls. Journal of Computer Science and Technology Studies, 5(3), 189-202. https://doi.org/10.32996/jcsts.2023.5.3.14
20.Xu, Yuming, Weiao Ying, Yi Du, Qing Wang, Juan Zheng, and Zhouliang Yang. "Swin Transformer–based intelligent fracture classification and radiographic assessment to support clinical decision-making." *BMC Medical Informatics and Decision Making* (2026).
21.Shamsabadi, Elyas Asadi, Chang Xu, Aravinda S. Rao, Tuan Nguyen, Tuan Ngo, and Daniel Dias-da-Costa. "Vision transformer-based autonomous crack detection on asphalt and concrete surfaces." *Automation in construction* 140 (2022): 104316.
22.Huang, Jin, Jianyao Gao, Jiazhi Li, Shanyan Gao, Fangxiao Cheng, and Yuzi Wang. "Transformer-based deep learning model for predicting osteoporosis in patients with cervical cancer undergoing external-beam radiotherapy." *Expert Systems with Applications* 273 (2025): 126716.

23. Devi MS, Aruna R, Almufti S, Punitha P, Kumar RL. Bone feature quantization and systematized attention gate UNet-based deep learning framework for bone fracture classification. Intell Data Anal. 2025;29(2):513--538.
24.. Mewada H, Pires IM, Rahevar M, Khatri N. Quantum convolutional neural network for bone fracture classification from X-ray images. Procedia Comput Sci. 2025;256:1143--1150.
25. Elsheikh AM, Elhag AA. Machine learning-based analysis of multi-region bone fracture detection and classification using biomedical images. Alexandria Eng J. 2025;128:186--199.
26. Kanagaraj K, Oliver AS, Ranjith R. Bone fracture detection and classification using node-level capsule graph neural network with X-ray images of broken and unbroken bones. Biomed Signal Process Control. 2025;110:108302.
27. Torne S, Shetty DK, Makkithaya K, Hegde P, Sudhi M, Pullela PK, Tamil Eniyan T, et al. VGG-16, VGG-16 with random forest, ResNet50 with SVM, and EfficientNetB0 with XGBoost -- enhancing bone fracture classification in X-ray using deep learning models. IEEE Access. 2025;13:25568--25577.
28. Sekar L, Karthikeyan M, Samynathan J, Pachiappan K, Iyyanar P, Rajan C. Bone fracture detection based on deep learning techniques. In: Proceedings of the 3rd International Conference on Artificial Intelligence and Machine Learning Applications (AIMLA); 2025:1--7. IEEE.
29. Alex R. Classification of multi-region bone fractures from X-ray images using transfer learning with ResNet18. Int J Artif Intell Med Issues. 2025;3(1):14--21.
30. Ahammed MS, Mridha MMR, Khan MT, Chaudhary MJ, Nakib MFN, Imtiaz S, Ashik MA. Bone fracture classification in X-ray images: a deep learning approach leveraging transfer learning. In: Proceedings of the International Conference on Electrical, Computer and Communication Engineering (ECCE); 2025:1--6. IEEE.
31.Islam MU, Fathima A, Ghosh D. An improved technique for bone fracture classification using visual transformers and convolutional neural networks. SSRN preprint 5191555; 2025.
32.Alam A, Al-Shamayleh AS, Thalji N, Raza A, Morales Barajas EA, Bautista Thompson E, de la Torre Diez I, Ashraf I. Novel transfer learning based bone fracture detection using radiographic images. BMC Med Imaging. 2025;25(1):5.
33. Chowdhury, M. N. I., Chy, K. S., Chowdhury, M. S. I., & Islam, M. A. (2021). Identifying Hidden Fraud Indicators through Longitudinal Analysis of U.S. Financial Misconduct. Journal of Economics, Finance and Accounting Studies , 3(1), 89-99. https://doi.org/10.32996/jefas.2021.3.1.9
34 .Alwzwazy HA, Alzubaidi L, Zhao Z, Gu Y. FracNet: An end-to-end deep learning framework for bone fracture detection. Pattern Recognit Lett. 2025;190:1--7.
35. Isgor B, Koklu M. Lightweight hybrid model for bone fracture detection using MobileNetV2 feature extraction and ensemble learning. J Future Artif Intell Technol. 2025;2(3):521--533.
36. Srinivasu PN, Aruna Kumari GL, Narahari SC, Ahmed S, Alhumam A. Exploring the impact of hyperparameter and data augmentation in YOLO V10 for accurate bone fracture detection from X-ray images. Sci Rep. 2025;15(1):9828.
37. Scutelnicu LA, Forna N, Luca M. An overview of techniques for automatic detection of bone fractures. Procedia Comput Sci. 2025;270:1557--1567.
38. Chy, K. S., Islam, M. A., Chowdhury, M. S. I., Chowdhury, M. N. I., & Islam, M. A. (2022). Enhancing U.S. Financial Compliance and Risk Management through Data-Driven Automation and Anomaly Detection . Journal of Business and Management Studies, 4(3), 231-244. https://doi.org/10.32996/jbms.2022.4.3.22
39. Saritas MM, Kursun R, Koklu M. Detection of bone fractures in X-ray images with machine learning methods using InceptionV3 deep features. 2025. Preprint.
40. Jothi R, Jayanthi K. A deep multi-filtering network based on dilated mask RCNN framework with self-attention RNN model for segmenting and classifying bone fracture. Multimed Tools Appl. 2025;84(25):29815--29841.