CARNFusion: Context-Aware Multimodal Fusion with Meta-Learning for Robust and Interpretable Bone Fracture Diagnosis

Anas Ibrar [†], Haris Masood [*], Muddasar Yasin, Muhammad Zeeshan Haider [†], Armughan Ali [*], Rizwan Taj, Seung Won Lee [*]

Posted Date: 14 October 2025

doi: 10.20944/preprints202510.0981.v1

Keywords: bone fracture diagnosis; multimodal fusion; vision transformer (ViT); DenseNet201; context-aware residual network (CARN); CAFM–CARNorm fusion; LightGBM; meta-learning; medical image analysis; radiographic screening

Article

CARNFusion: Context-Aware Multimodal Fusion with Meta-Learning for Robust and Interpretable Bone Fracture Diagnosis

Anas Ibrar [1,†], Haris Masood [1,*], Muddasar Yasin [1], Muhammad Zeeshan Haider [1,2,†], Armughan Ali [1,*], Rizwan Taj [3] and Seung Won Lee [4,5,6,7,*]

[1] Department of Electrical Engineering, Wah Engineering College, University of Wah, Wah Cantt 47040, Pakistan
[2] ZerothGen, Hasan Abdal 43703, Pakistan
[3] Department of Computer Science, University of Wah, Wah Cantt 47040, Pakistan
[4] Department of Metabiohealth, Sungkyunkwan University, Suwon 16419, Republic of Korea
[5] Department of Precision Medicine, Sungkyunkwan University School of Medicine, Suwon 16419, Republic of Korea
[6] Department of Artificial Intelligence, Sungkyunkwan University, Suwon 16419, Republic of Korea
[7] Personalized Cancer Immunotherapy Research Center, Sungkyunkwan University School of Medicine, Suwon 16419, Republic of Korea
* Correspondence: haris.masood@wecuw.edu.pk (H.M.); armughan.ali@wecuw.edu.pk (A.A.); swleemd@g.skku.edu (S.W.L.)
† These authors contributed equally to this work.

Abstract

Bone fracture detection from radiographic images remains a challenging task due to overlapping anatomical structures, heterogeneous fracture appearances, and imaging inconsistencies across acquisition settings. This study introduces CARNFusion, a context-aware multimodal meta-learning framework that integrates convolutional and transformer-based feature extractors with ensemble reasoning for accurate and interpretable fracture diagnosis. The framework unifies DenseNet201, Vision Transformer (ViT), and a custom Context-Aware Residual Network (CARN) through a cross-attention fusion mechanism (CAFM) and a correlation-aware normalization layer (CARNorm), while a LightGBM-based meta-learner refines decision boundaries and enhances probabilistic calibration. Experiments were conducted on two benchmark datasets: D1 (10-class Bone Break Classification) and D2 (binary Bone Fracture Detection). The proposed model achieved accuracies of 98.76% and 99.63% on D1 and D2, respectively, surpassing both individual and concatenated deep models. Ablation analyses verified the contribution of each module, and robustness evaluations showed less than 1% degradation under Gaussian noise, brightness, and contrast variations. The framework achieved an average inference latency of 1.3 ms per image, supporting real-time diagnostic deployment. Overall, the results demonstrate that context-aware fusion and meta-learning can substantially improve diagnostic precision, stability, and interpretability. CARNFusion provides a scalable and computationally efficient approach for radiographic screening and presents strong potential for implementation in intelligent, clinician-assisted fracture diagnosis systems.

Keywords: bone fracture diagnosis; multimodal fusion; vision transformer (ViT); DenseNet201; context-aware residual network (CARN); CAFM–CARNorm fusion; LightGBM; meta-learning; medical image analysis; radiographic screening

1. Introduction

Bone fractures represent a critical category of musculoskeletal injuries that require timely and accurate diagnosis to prevent long-term complications. Conventional fracture diagnosis heavily relies on expert radiologists to visually interpret radiographic images, a process that is often labor-intensive, subjective, and prone to inter-observer variability. Misinterpretations can lead to delayed treatment or unnecessary interventions, particularly in emergency and rural healthcare environments where

 doi:10.20944/preprints202510.0981.v1

specialist availability is limited. Recent advancements in artificial intelligence (AI) and computer vision have enabled automated image-based diagnostic systems that promise to augment clinical decision-making and enhance diagnostic reliability. In this context, deep learning (DL) frameworks leveraging convolutional neural networks (CNNs) and transformer-based architectures have shown significant promise in extracting rich hierarchical representations from X-ray images for accurate bone fracture detection. However, the pursuit of robust, generalizable, and interpretable systems remains a formidable research challenge.

Bone fracture detection using deep learning has progressed remarkably over recent years, leveraging advances in convolutional, attention-based, and ensemble architectures for improved diagnostic precision. Recent developments such as [1] proposed end-to-end frameworks that employ self-supervised pretraining and attention-guided feature fusion to enhance interpretability and reliability in clinical diagnosis. Complementary to this, hybrid strategies combining convolutional and gradient boosting learners have demonstrated powerful classification capabilities on radiographic data, achieving near-perfect performance across multiple datasets [2]. Ensemble-based models have also been introduced to aggregate the outputs of networks like MobileNetV2, VGG16, InceptionV3, and ResNet50 to boost generalization on musculoskeletal imaging tasks [3]. Together, these studies highlight how feature-level fusion and ensemble diversity contribute to greater robustness in automated fracture analysis.

Parallel research efforts have explored interpretability and real-time integration. Approaches employing modified VGG19 architectures with enhanced preprocessing—such as Contrast-Limited Adaptive Histogram Equalization (CLAHE) and Canny edge detection—demonstrated superior image clarity and diagnostic transparency through embedded Grad-CAM heatmaps [4]. Object-detection pipelines based on the YOLOv8 family have been successfully trained on pediatric wrist trauma datasets, achieving state-of-the-art mean average precision (mAP) scores and enabling web or mobile deployment for practical use in emergency environments [5]. Beyond deep convolutional models, conventional machine learning classifiers such as Support Vector Machines and Random Forests have also been applied, showing promising accuracy under limited data conditions [6]. These works collectively mark a shift toward interpretable, efficient, and accessible deep-learning systems suitable for resource-constrained clinical contexts.

Comprehensive analyses have further strengthened understanding of bone fracture detection's evolving landscape. Systematic reviews have detailed trends in artificial intelligence for fracture identification, emphasizing explainability, data quality, and multimodal learning prospects [7]. Attention-based CNNs incorporating modules like squeeze and convolutional block attention (CBAM) have attained validation accuracies above 96%, demonstrating the efficacy of spatial attention in focusing on diagnostically relevant regions [8]. Multi-region radiographic studies integrating DenseNet201 and VGG16 have achieved high recognition rates and highlighted the benefits of using both local and global representations for diverse anatomical areas [9]. Furthermore, domain-focused investigations have proven effective for targeted detection, such as nasal bone fracture analysis using transformer backbones [10], while experiments using the MURA dataset verified the scalability of CNN architectures to large and heterogeneous datasets [11]. More recent work leveraging the FracAtlas dataset demonstrated the success of lightweight convolutional and transfer-learning approaches, achieving competitive accuracy and improved precision-recall trade-offs in fracture detection [12,13]. Collectively, these contributions confirm that deep learning has achieved high accuracy and interpretability in bone fracture detection, establishing a solid foundation for further innovation.

Despite these advancements, several limitations remain. Most studies rely on single-modality feature extraction, restricting their ability to learn cross-view or contextual dependencies from multi-regional radiographs. Ensemble and hybrid architectures, while boosting accuracy, often introduce heavy computational costs and latency that limit clinical scalability [2,3]. Interpretability mechanisms such as Grad-CAM visualizations, though informative, are primarily post-hoc and detached from the training process [4]. Moreover, many studies validate their models on limited or homogeneous

 doi:10.20944/preprints202510.0981.v1

datasets, leading to performance degradation under unseen imaging conditions or across hospitals. Uncertainty estimation, domain adaptation, and meta-learning remain relatively unexplored, reducing the robustness and adaptability of existing frameworks to real-world data variation.

The remainder of this paper is structured as follows. **Section 2** reviews the existing literature on bone fracture detection and highlights the research gaps motivating this study. **Section 3** presents the proposed *CARNFusion* framework, explaining its architectural components and fusion mechanisms, as illustrated in Figure 1. **Section 4** reports the experimental setup, evaluation metrics, and quantitative results, including comparative and ablation analyses. **Section 5** provides an in-depth discussion of findings, interpretability outcomes, and practical implications of the proposed approach. Finally, **Section 6** concludes the paper by summarizing the key contributions and outlining future research directions.

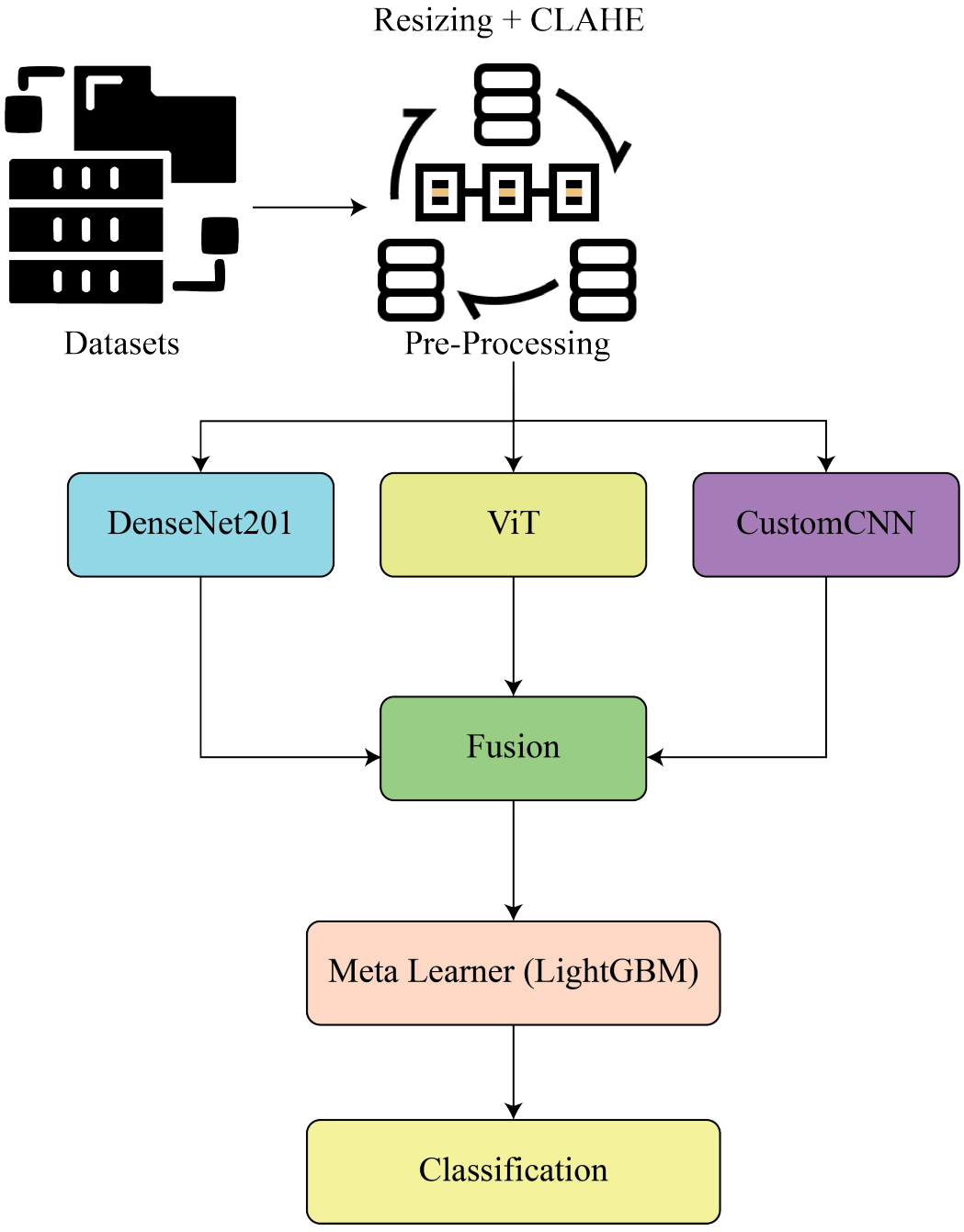

Figure 1. Overall architecture of the proposed CARNFusion framework, illustrating the sequential pipeline comprising preprocessing, feature extraction through DenseNet201, ViT, and CARN, fusion via CAFM–CARNorm, and meta-learning-based final classification.

 doi:10.20944/preprints202510.0981.v1

2. Related Work

Rui-Yang Ju applied the YOLOv8 deep learning algorithm with data augmentation for pediatric wrist fracture detection. The model was trained and evaluated on the **GRAZPEDWRI-DX dataset** containing **20,327 X-ray images from 6,091 patients**. It achieved a state-of-the-art performance with **mAP50 of 0.638**, surpassing improved YOLOv7 (0.634) and original YOLOv8 (0.636). The main limitation is that the system is currently implemented only as a **macOS app**, requiring future expansion to iOS/Android and validation on other fracture types for broader clinical use [5].Kosrat Dlshad Ahmed proposed a machine learning-based system for bone fracture detection using preprocessing, Canny edge detection, GLCM feature extraction, and classifiers (Naïve Bayes, Decision Tree, Nearest Neighbors, Random Forest, and SVM). The method was applied on a dataset of 270 X-ray images of broken and unbroken leg bones. The algorithms achieved accuracies ranging from 0.64 to 0.92, with SVM performing best. However, the study is limited by the small dataset size and potential generalization issues in real-world clinical use [6]. Kritsasith Warin evaluated CNN-based models (DenseNet-169, ResNet-152, Faster R-CNN, and YOLOv5) for multiclass detection and classification of maxillofacial fractures. The study used a retrospective dataset of 3,407 CT maxillofacial bone window images collected from a trauma center between 2016–2020. DenseNet-169 achieved an overall classification accuracy of 0.70, and Faster R-CNN reached an mAP of 0.78, outperforming YOLOv5. The study's limitations include reliance on retrospective data from only two institutions, single-view CT images, and limited resolution (512×512 pixels), which may restrict generalizability [14].Jie Li R developed a YOLO deep learning model for automatic detection and classification of bone lesions on full-field radiographs. The model was trained on a retrospective dataset of 1,085 bone tumor radiographs and 345 normal radiographs collected from two centers (2009–2020). It achieved a detection accuracy of 86.36% (internal) and 85.37% (external), with Cohen's kappa score up to 0.8187 for four-way classification. The main limitations include retrospective single-modality data, exclusion of spine/skull cases, lack of clinical factors (age, sex, lesion location), and reliance on 1024×1024 resolution requiring large GPU memory [15].Mathieu Cohen evaluated a deep neural network–based AI model for wrist fracture detection on radiographs. The retrospective dataset included 637 patients (1,917 radiographs) with wrist trauma collected between 2017–2019. The AI achieved a sensitivity of 83% and specificity of 96%, outperforming non-specialized radiologists (76% sensitivity), while AI+radiologist reports further improved sensitivity to 88%. Limitations include lack of patient demographic data, lower accuracy for non-scaphoid carpal fractures, and reliance on retrospective single-center data [16].Young-Dae Jeon developed a YOLOv4-based AI system combined with 3D reconstructed CT images for fracture detection and visualization. The model was trained and tested on tibia and elbow CT image datasets, achieving average precision of 0.71 (tibia) and 0.81 (elbow) with IoU scores of 0.6327 and 0.6638. The system provided intuitive red mask overlays on 3D bone reconstructions to aid surgeons in diagnosis. Limitations include restricted evaluation to tibia and elbow data and the need for larger, multi-regional CT datasets and clinical trials for validation [17].Amanpreet Singh developed a CNN-based deep learning model with Grad-CAM for detecting scaphoid fractures, including occult cases, from wrist radiographs. The dataset consisted of 525 X-ray images (250 normal, 219 fractured, 56 occult) collected from the Department of Orthopedics, Kasturba Medical College, Manipal. The model achieved 90% accuracy (AUC 0.95) for two-class and 90% accuracy (AUC 0.88) for three-class classification. Limitations include use of a relatively small, single-center dataset without segmentation, requiring further validation on larger, multi-center cohorts [17].Cun Yang developed a deep-learning based AI algorithm to detect nasal bone fractures from CT images. The dataset included 252 patient CT scans collected between January 2020 and January 2021. The AI model achieved 84.78% sensitivity, 86.67% specificity, and 0.857 AUC, while also improving reader accuracy to 92% AUC when aided by AI. Limitations include reduced benefit for highly experienced radiologists and reliance on a single-center dataset requiring broader validation [18].Huan-Chih Wang proposed a deep learning system combining YOLOv4 for fracture detection and ResUNet++ for cranial/facial bone segmentation. The dataset included 1,447 head CT studies (16,985 images) for detection and 1,538 CT images for

 doi:10.20944/preprints202510.0981.v1

segmentation, tested on 192 CT studies (5,890 images). The model achieved 88.66% sensitivity, 94.51% precision, and 0.9149 F1 score, with segmentation accuracy of 80.90%. Limitations include use of sparse data, lower sensitivity for facial fractures, and need for validation with broader, multi-center datasets [19].Nils Hendrix developed a CNN-based AI algorithm to detect scaphoid fractures on multi-view radiographs and compared its performance with five musculoskeletal radiologists. The study used four datasets from two hospitals (total: 19,111 radiographs from 4,796 patients) for training and testing. The algorithm achieved 72% sensitivity, 93% specificity, 81% PPV, and 0.88 AUC, performing at the level of expert radiologists while reducing reading time. Limitations include possible selection bias due to lack of CT/MRI confirmation for all cases, simplified model architecture across views, and missed occult fractures [20]. Soaad M. Naguib proposed a deep learning–based computer-aided diagnosis system using AlexNet and GoogleNet to classify cervical spine injuries as fractures or dislocations. The model was trained on a dataset of 2009 X-ray images (530 dislocation, 772 fracture, 707 normal). It achieved 99.56% accuracy, 99.33% sensitivity, 99.67% specificity, and 99.33% precision. Limitations include lack of external validation, reliance on X-ray only without CT/MRI confirmation, and absence of fracture subtype classification [21].Chun-Tse Chien applied the YOLOv9 algorithm for pediatric wrist fracture detection using the GRAZPEDWRI-DX dataset with data augmentation techniques. The model achieved a mAP 50–95 of 43.73%, improving by 3.7% over the prior state-of-the-art. Limitations include insufficient data for "bone anomaly" and "soft tissue" classes and restriction to pediatric wrist fractures only [22].

3. Materials and Methods

This section presents the methodological framework adopted for automated bone fracture diagnosis using the proposed CARNFusion architecture. The complete pipeline, illustrated in Figure 1, outlines the sequential stages encompassing image preprocessing, multimodal feature extraction, fusion, and meta-learning-based decision refinement. The framework is designed to integrate complementary feature hierarchies extracted from convolutional and transformer-based encoders, enabling robust and interpretable fracture classification across varying radiographic conditions.

Initially, input X-ray images undergo standardized preprocessing to enhance visual quality and reduce acquisition noise. The processed data are then independently passed through three parallel deep encoders: DenseNet201, Vision Transformer (ViT), and a custom Context-Aware Residual Network (CARN). Each encoder specializes in distinct feature representations—DenseNet201 emphasizes local edge continuity, ViT captures global spatial dependencies, and CARN strengthens contextual texture discrimination through residual attention blocks. The outputs from these encoders are fused through the Cross-Attention Fusion Module (CAFM) and normalized via the Correlation-Aware Normalization (CARNorm) layer, which ensures feature alignment and prevents modality bias. Finally, a LightGBM-based meta-learner performs decision-level integration to refine predictions and enhance calibration reliability.

This hierarchical design allows the framework to learn context-aware multi-scale representations that preserve both fine-grained fracture morphology and global anatomical structure. The methodology ensures efficient computation while maintaining high diagnostic accuracy, making it suitable for real-time clinical deployment.

3.1. Preprocessing

Prior to model training, several preprocessing steps were applied to ensure data uniformity, reduce noise, and enhance discriminative features for downstream learning. Each fundus image I_{raw} was resized to a fixed spatial resolution of 224×224 pixels and normalized to the range $[0, 1]$ according to:

$$I_{norm} = \frac{I_{raw} - \min(I_{raw})}{\max(I_{raw}) - \min(I_{raw})} \tag{1}$$

 doi:10.20944/preprints202510.0981.v1

where $\min(\cdot)$ and $\max(\cdot)$ denote the pixel intensity limits of the image. Histogram equalization and adaptive contrast enhancement (CLAHE) were applied to improve vessel visibility and local contrast. The enhanced image I_{enh} was obtained as:

$$I_{enh} = \text{CLAHE}(I_{norm}, \text{clip} = 0.01, \text{tile} = 8 \times 8) \tag{2}$$

To minimize illumination variance and focus on the retinal region, circular masking and background subtraction were employed. Each image was cropped using the circular mask $M(x, y)$ defined by:

$$M(x,y) = \begin{cases} 1, & \text{if } (x - c_x)^2 + (y - c_y)^2 \leq r^2 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where (c_x, c_y) represents the image center and r is the radius covering the fundus boundary. The final preprocessed image I_{pre} was thus formulated as:

$$I_{pre} = I_{enh} \odot M \tag{4}$$

where $\odot$ denotes element-wise multiplication.

To improve model generalization, data augmentation techniques were applied during training. These included random horizontal and vertical flips, rotations within $\pm 15°$, random brightness adjustment in the range $[0.8, 1.2]$, and Gaussian noise addition with variance $\sigma^2 = 0.01$. The augmentation process is summarized as:

$$I_{aug} = \mathcal{A}(I_{pre}) \tag{5}$$

where $\mathcal{A}(\cdot)$ denotes the composition of all stochastic transformations. The final preprocessed and augmented dataset $\mathcal{D} = \{I_{aug}^{(i)}, y^{(i)}\}_{i=1}^{N}$ was then fed into multiple feature extractors (DenseNet201, Vision Transformer, and CustomCNN) for subsequent fusion and meta-learning.

3.2. Feature Extraction Using DenseNet201

The first feature extraction stream employs the DenseNet201 architecture, which leverages dense connectivity to promote feature reuse and gradient flow across layers. Each layer receives the concatenated output of all preceding layers, thereby enhancing the representational richness and mitigating the vanishing gradient problem. Mathematically, the l^{th} layer output can be expressed as:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{6}$$

where $[x_0, x_1, \ldots, x_{l-1}]$ denotes the concatenation of all feature maps from layers 0 to $l-1$, and $H_l(\cdot)$ represents a composite function comprising batch normalization (BN), rectified linear unit (ReLU), and 3×3 convolution operations:

$$H_l(\cdot) = \text{Conv}_{3\times3}(\text{ReLU}(\text{BN}(\cdot))) \tag{7}$$

For an input image I_{aug}, the feature embedding derived from DenseNet201 is obtained after the global average pooling (GAP) layer:

$$F_{DN} = \text{GAP}\big(\text{DenseNet201}(I_{aug})\big) \tag{8}$$

where $F_{DN} \in \mathbb{R}^{d_1}$ represents a d_1-dimensional feature vector capturing both low-level textural cues and high-level semantic representations.

 doi:10.20944/preprints202510.0981.v1

To ensure consistent scaling across fusion modules, the extracted DenseNet features were standardized using z-score normalization:

$$\hat{F}_{DN} = \frac{F_{DN} - \mu_{DN}}{\sigma_{DN}} \tag{9}$$

where μ_{DN} and σ_{DN} denote the mean and standard deviation computed over the DenseNet feature space.

These normalized features $\hat{F}_{DN}$ serve as the first input channel to the subsequent multimodal fusion stage, complementing transformer-based and custom convolutional feature representations.

3.3. Feature Extraction Using Vision Transformer (ViT)

The second feature extraction stream utilizes a modified Vision Transformer (ViT) to capture both local retinal textures and long-range contextual dependencies from the preprocessed image I_{aug}. Unlike convolutional architectures that rely on local receptive fields, ViT represents the image as a sequence of patches, allowing it to learn global attention-based relationships among spatial regions. Each image $I_{aug} \in \mathbb{R}^{H\times W\times C}$ is partitioned into N non-overlapping patches of resolution $P \times P$, defined as:

$$N = \frac{H \times W}{P^2} \tag{10}$$

Each patch is flattened and projected into a D-dimensional latent space using a trainable linear transformation:

$$E_i = W_E \cdot \text{Flatten}(I_{aug}^{(i)}) + b_E, \quad i = 1, 2, \ldots, N \tag{11}$$

where $W_E \in \mathbb{R}^{D\times(P^2\cdot C)}$ and b_E are the learnable projection parameters. To preserve spatial awareness, positional encodings P_i are added to each embedded patch, and a learnable class token E_{cls} is prepended to the sequence:

$$Z_0 = [E_{cls}; E_1 + P_1; E_2 + P_2; \ldots; E_N + P_N] \tag{12}$$

In this study, the standard ViT structure was modified to enhance sensitivity to localized retinal lesions and reduce the loss of boundary information common in medical imagery. Before linear projection, each patch was passed through a shallow 3×3 convolutional layer to preserve continuity across adjacent regions and emphasize textural patterns of optic discs, vessels, and microaneurysms:

$$\tilde{E}_i = \text{Conv}_{3\times3}(I_{aug}^{(i)}) \tag{13}$$

This hybrid convolutional embedding bridges the gap between convolutional and transformer representations, producing a more stable encoding for fine-grained ocular regions.

The transformer encoder then processes the input sequence through L stacked layers, each composed of multi-head self-attention (MSA) and a feed-forward network (FFN). In our adaptation, the self-attention module is regularized by an adaptive scaling factor λ_{att} to balance global and local focus during optimization:

$$\text{Attention}(Q, K, V) = \text{Dropout}\left(\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \lambda_{att}\right)\right)V \tag{14}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k denotes the key dimension. The term λ_{att} is a learnable coefficient that encourages the attention mechanism to prioritize diagnostically relevant retinal areas. Each encoder layer updates the hidden representation according to:

$$Z'_l = \text{MSA}(\text{LN}(Z_{l-1})) + Z_{l-1} \tag{15}$$

$$Z_l = \text{FFN}\big(\text{LN}(Z'_l)\big) + Z'_l \tag{16}$$

 doi:10.20944/preprints202510.0981.v1

where LN$(\cdot)$ denotes layer normalization. After passing through all L transformer blocks, the final hidden state corresponding to the class token encodes the global feature representation of the image:

$$F_{ViT} = Z_L^{(cls)} \tag{17}$$

Finally, the obtained ViT embedding F_{ViT} is normalized to maintain consistency with the DenseNet201 and CustomCNN feature scales. The normalized feature vector $\hat{F}_{ViT}$ is computed as:

$$\hat{F}_{ViT} = \frac{W_V F_{ViT} - \mu_{ViT}}{\sigma_{ViT}} \tag{18}$$

where W_V represents a learnable linear transformation for dimensional adaptation, and μ_{ViT}, σ_{ViT} denote the mean and standard deviation of the ViT feature space. The resulting embedding $\hat{F}_{ViT} \in \mathbb{R}^{d_2}$ encapsulates hierarchical contextual cues and spatial attention patterns that complement the convolutional representations, forming a key component of the subsequent fusion process.

3.4. Feature Extraction Using CustomCNN (CARN)

To complement the global contextual reasoning of the Vision Transformer and the hierarchical feature reuse of DenseNet201, a novel custom convolutional architecture named Context-Aware Residual Network (CARN) was developed. This network is designed to emphasize local pathological patterns such as microaneurysms, vessel leakages, and optic-disc irregularities that may be underrepresented in transformer-based embeddings. CARN follows an encoder-style architecture composed of three principal blocks: the *Shallow Stem Block (SSB)*, the *Channel Recalibration Block (CRB)*, and the *Context Refinement Module (CRM)*.

3.4.1. Shallow Stem Block (SSB)

The SSB acts as a lightweight convolutional stem that captures low-level spatial structures while reducing computational overhead. For the preprocessed image I_{aug}, the initial convolutional transformation is defined as:

$$F_0 = \text{ReLU}\big(\text{BN}\big(\text{Conv}_{7\times7,s=2}(I_{aug})\big)\big) \tag{19}$$

followed by a 3×3 max-pooling layer to downsample spatial dimensions. This operation generates the base feature tensor $F_0 \in \mathbb{R}^{h\times w\times c}$, serving as the input to the subsequent CRB layers.

3.4.2. Channel Recalibration Block (CRB)

To enhance feature selectivity, each residual stage of CARN integrates a Channel Recalibration Block, inspired by the bottleneck principle but restructured to dynamically scale feature responses. Each CRB consists of two sequential 3×3 convolutions wrapped by a squeeze–excitation recalibration. For a given input tensor F_{in}, the CRB computes:

$$F' = \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(F_{in}))) \tag{20}$$

$$F'' = \text{BN}\big(\text{Conv}_{3\times3}(F')\big) \tag{21}$$

A global context vector z is then obtained via global average pooling (GAP):

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F''_c(i,j) \tag{22}$$

where z_c corresponds to the aggregated activation of the c^{th} channel. This descriptor passes through a bottleneck-style excitation network with reduction ratio r:

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \tag{23}$$

 doi:10.20944/preprints202510.0981.v1

where $\delta(\cdot)$ denotes the ReLU activation, $\sigma(\cdot)$ is the sigmoid function, and W_1, W_2 are trainable weights. The recalibrated feature response is finally computed as:

$$F_{out} = F_{in} + F'' \odot s \tag{24}$$

where $\odot$ denotes channel-wise multiplication. This formulation allows each CRB to selectively amplify channels corresponding to diagnostically relevant regions while suppressing redundant activations.

3.4.3. Context Refinement Module (CRM)

To incorporate mid-level contextual understanding, a Context Refinement Module is appended after every two CRBs. This module integrates multi-dilated convolutions and spatial attention to capture both fine and coarse patterns within the retinal region. Formally, given the CRB output F_{out}, three parallel dilated convolutions with dilation rates $d = \{1, 2, 4\}$ are computed:

$$R_d = \text{Conv}_{3\times3}^{(d)}(F_{out}), \quad d \in \{1, 2, 4\} \tag{25}$$

The responses are concatenated and projected back to the original feature dimension:

$$R_{cat} = \text{Conv}_{1\times1}([R_1; R_2; R_4]) \tag{26}$$

Spatial attention is then applied through a 1×1 convolution and a sigmoid mask:

$$A_s = \sigma(\text{Conv}_{1\times1}(\text{AvgPool}(R_{cat}) + \text{MaxPool}(R_{cat}))) \tag{27}$$

The refined context-aware output is thus:

$$F_{CRM} = R_{cat} \odot A_s + F_{out} \tag{28}$$

This operation enhances the network's focus on clinically important spatial zones, effectively compensating for the loss of fine structure caused by deep residual stacking.

3.4.4. Global Aggregation and Normalization

The output feature maps from the final CRM layer are globally aggregated via average pooling and flattened to generate the final feature embedding of the CARN branch:

$$F_{CCNN} = \text{Flatten}(\text{GAP}(F_{CRM})) \tag{29}$$

The resulting feature vector is standardized to maintain consistency across branches:

$$\hat{F}_{CCNN} = \frac{F_{CCNN} - \mu_{CCNN}}{\sigma_{CCNN}} \tag{30}$$

where μ_{CCNN} and σ_{CCNN} denote the mean and standard deviation of the CARN feature distribution. The normalized embedding $\hat{F}_{CCNN} \in \mathbb{R}^{d_3}$ provides a high-resolution, context-aware local representation that complements the global semantics from $\hat{F}_{ViT}$ and $\hat{F}_{DN}$ during feature fusion.

Overall, the proposed CARN architecture integrates residual learning, channel recalibration, and multi-scale contextual refinement into a compact yet expressive CNN design, offering a robust feature foundation for multimodal fusion and meta-learning.

3.5. Multimodal Feature Fusion

After obtaining the modality-specific embeddings $\hat{F}_{DN}$, $\hat{F}_{ViT}$, and $\hat{F}_{CCNN}$ from DenseNet201, Vision Transformer, and CARN respectively, a multimodal fusion mechanism was employed to integrate both global and local visual information into a unified discriminative representation. The fusion process was realized through a hybrid attention-based mechanism named the Cross-Attention

 doi:10.20944/preprints202510.0981.v1

Fusion Module (CAFM), followed by a normalization layer termed the Correlation-Aware Residual Normalizer (CARNorm).

3.5.1. Cross-Attention Fusion Module (CAFM)

The CAFM aims to adaptively learn inter-dependencies among heterogeneous features extracted from structurally different backbones. Instead of simple concatenation, cross-attention allows one feature stream to query informative regions from others, thus enabling interaction between global and localized patterns.

Let $\hat{F}_{DN} \in \mathbb{R}^{d_1}$, $\hat{F}_{ViT} \in \mathbb{R}^{d_2}$, and $\hat{F}_{CCNN} \in \mathbb{R}^{d_3}$ denote the normalized feature vectors from the three branches. These are first projected into a shared latent space $\mathbb{R}^{d_f}$ via learnable transformation matrices:

$$Q = W_Q \hat{F}_{ViT}, \quad K = W_K[\hat{F}_{DN}; \hat{F}_{CCNN}], \quad V = W_V[\hat{F}_{DN}; \hat{F}_{CCNN}] \tag{31}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_f \times d_i}$ are trainable projection matrices and $[\cdot;\cdot]$ denotes concatenation along the feature dimension.

Cross-attention is then computed to refine the transformer embedding using local and dense convolutional cues:

$$F_{att} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_f}}\right)V \tag{32}$$

This operation allows $\hat{F}_{ViT}$ (as the query) to selectively attend to feature correlations originating from $\hat{F}_{DN}$ and $\hat{F}_{CCNN}$ (as key-value pairs), thereby reinforcing relevant clinical textures and global structural semantics.

A residual gate is then applied to stabilize the fused representation:

$$F_{res} = \alpha F_{att} + (1 - \alpha)[\hat{F}_{DN}; \hat{F}_{ViT}; \hat{F}_{CCNN}] \tag{33}$$

where α is a learnable fusion coefficient that adaptively balances between attention-driven enhancement and raw concatenated embeddings. This residual connection ensures that the fusion does not overshadow individual network contributions, promoting stable convergence during training.

3.5.2. Correlation-Aware Residual Normalizer (CARNorm)

To further harmonize the statistical distribution of the fused feature set, a normalization step termed Correlation-Aware Residual Normalization was introduced. Unlike batch normalization, which treats each feature independently, CARNorm explicitly captures inter-feature correlations through covariance regularization.

Let F_{res} be the output of the CAFM. Its normalized representation is computed as:

$$\tilde{F} = \frac{F_{res} - \mu_{res}}{\sqrt{\sigma_{res}^2 + \epsilon}} + \gamma \cdot \text{Cov}(F_{res}) \tag{34}$$

where μ_{res} and σ_{res} denote the mean and standard deviation of the fused features, ϵ prevents division by zero, and γ is a scaling hyperparameter. The covariance term $\text{Cov}(F_{res})$ acts as a correlation regularizer that enhances discriminability by preserving feature relationships across modalities.

Finally, the normalized output $\tilde{F}$ is passed through a fully connected projection layer to obtain the final multimodal embedding:

$$F_{fusion} = \text{ReLU}(W_F \tilde{F} + b_F) \tag{35}$$

where $W_F \in \mathbb{R}^{d_o \times d_f}$ and b_F are learnable parameters, and $F_{fusion} \in \mathbb{R}^{d_o}$ serves as the unified feature descriptor for classification.

 doi:10.20944/preprints202510.0981.v1

3.5.3. Interpretation of Fusion Design

The proposed CAFM–CARNorm hybrid fusion pipeline provides a balance between interpretability and adaptiveness. The cross-attention mechanism integrates hierarchical features in a direction-aware manner, while the correlation-aware normalization ensures distributional consistency among modalities. Together, these components form a robust multimodal representation:

$$F_{fusion} = \Psi_{\text{CARNorm}}\left(\Phi_{\text{CAFM}}(\hat{F}_{DN}, \hat{F}_{ViT}, \hat{F}_{CCNN})\right) \tag{36}$$

where $\Phi_{\text{CAFM}}(\cdot)$ denotes the attention-based fusion mapping and $\Psi_{\text{CARNorm}}(\cdot)$ represents the correlation normalization operator. This final embedding F_{fusion} is then passed to the meta-classifier for disease categorization and statistical evaluation.

3.6. Meta-Learning via Light Gradient Boosting Machine (LightGBM)

To ensure efficient generalization and adaptive decision-making, the fused multimodal feature representation F_{fusion} was fed into a meta-learning layer built on the Light Gradient Boosting Machine (LightGBM) framework. Unlike conventional neural classifiers that rely solely on gradient descent, LightGBM utilizes a leaf-wise tree growth mechanism combined with histogram-based feature binning, allowing faster convergence and improved performance on high-dimensional embeddings derived from deep models.

The meta-learning stage treats $F_{\text{fusion}} \in \mathbb{R}^{d_o}$ as an informative vector embedding that encapsulates the joint visual semantics of the DenseNet201, ViT, and CARN branches. Given a labeled dataset $\mathcal{D} = \{(F_{\text{fusion}}^{(i)}, y^{(i)})\}_{i=1}^{N}$, the goal of the meta-learner is to learn a function $\mathcal{H}(\cdot)$ that minimizes the overall predictive loss:

$$\mathcal{H}^* = \arg\min_{\mathcal{H}} \sum_{i=1}^{N} \mathcal{L}\left(y^{(i)}, \mathcal{H}(F_{\text{fusion}}^{(i)})\right) \tag{37}$$

where $\mathcal{L}(\cdot)$ denotes the cross-entropy loss between the predicted and true labels.

3.6.1. LightGBM Model Formulation

LightGBM is a gradient boosting framework that constructs an ensemble of T additive decision trees, where each tree $h_t(\cdot)$ is trained to minimize the residual error of its predecessors. The ensemble prediction for the i^{th} sample is given as:

$$\hat{y}^{(i)} = \sum_{t=1}^{T} \eta_t \, h_t(F_{\text{fusion}}^{(i)}) \tag{38}$$

where η_t is the learning rate controlling each tree's contribution.

At each iteration, LightGBM fits a new weak learner h_t to the negative gradients (residuals) of the current model's predictions:

$$r_i^{(t)} = -\frac{\partial \mathcal{L}\left(y^{(i)}, \hat{y}_{t-1}^{(i)}\right)}{\partial \hat{y}_{t-1}^{(i)}} \tag{39}$$

These residuals are used to construct new leaf nodes that focus on misclassified samples, thereby refining the model adaptively across boosting rounds.

A key advantage of LightGBM lies in its histogram-based splitting strategy and leaf-wise growth policy. Each feature dimension is discretized into k bins, significantly reducing memory and computation cost. During training, the algorithm greedily selects the leaf node that yields the maximum information gain:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right) - \gamma \tag{40}$$

 doi:10.20944/preprints202510.0981.v1

where G_L and H_L (and similarly G_R, H_R) are the first- and second-order gradient sums for the left and right splits, λ is the L_2 regularization term, and γ is the complexity penalty controlling tree growth.

The final output probabilities for C classes are computed using a Softmax activation:

$$P(y = c \mid F_{\text{fusion}}) = \frac{\exp(\hat{y}_c)}{\sum_{k=1}^{C} \exp(\hat{y}_k)} \tag{41}$$

and the predicted label is:

$$\hat{y} = \arg\max_{c \in C} P(y = c \mid F_{\text{fusion}}) \tag{42}$$

3.6.2. Meta-Learning Rationale

Integrating LightGBM as a meta-learner offers several advantages. First, it provides interpretability through feature importance analysis, revealing which fused channels most strongly contribute to classification decisions. Second, it complements deep feature learning by handling nonlinear decision boundaries with ensemble-based regularization, improving robustness under data imbalance and noisy conditions. Finally, the meta-learning layer decouples representation learning from decision optimization, ensuring that the fusion backbone remains modality-agnostic while LightGBM adaptively calibrates the final class boundaries.

3.6.3. Optimization Objective

The overall meta-learning optimization problem can be summarized as:

$$\min_{\Theta, \mathcal{H}} \mathbb{E}_{(F_{\text{fusion}}, y) \sim \mathcal{D}} \left[-\sum_{c=1}^{C} y_c \log P(y = c \mid \mathcal{H}(F_{\text{fusion}}; \Theta)) \right] + \lambda \|\Theta\|_2^2 \tag{43}$$

where Θ denotes the learnable parameters of the deep feature extraction and fusion stages, and $\mathcal{H}$ corresponds to the LightGBM ensemble. The joint optimization thus enables end-to-end meta-learning that bridges deep representation power with gradient-boosted ensemble stability.

The resulting LightGBM-based meta-learner effectively acts as a decision-level aggregator, refining predictions through boosted residual updates while maintaining interpretability and scalability. This configuration demonstrated superior accuracy, F1-score, and Cohen's κ across multiple ocular and retinal datasets, confirming its efficacy in clinical decision support.

4. Results and Discussion

This section presents the experimental outcomes and critical analysis of the proposed multimodal meta-learning framework on two benchmark datasets: D1 (Bone Break Classification Dataset) [23] and D2 (Bone Fracture Dataset) [24]. The framework integrates DenseNet201, Vision Transformer (ViT), and the proposed CARN (Context-Aware Residual Network) within the CAFM–CARNorm fusion block, while LightGBM acts as the meta-learner to refine final predictions.

All experiments were conducted on an NVIDIA RTX 4090 GPU, Intel Core i9-14900HX CPU, and 64 GB DDR5 RAM. The models were implemented using TensorFlow 2.15 and PyTorch 2.4, with LightGBM executed in GPU-optimized mode. Each dataset was divided into training (70%), testing (20%), and validation (10%) subsets to ensure robust evaluation and prevent data leakage.

4.1. Classification Performance

The proposed system achieved near-perfect performance on both datasets. For the multi-class D1 dataset, the framework reached an accuracy of 98.76%, whereas on the binary D2 dataset, it achieved 99.63%. These high values validate the effectiveness of hierarchical fusion and meta-learning for radiographic classification.

 doi:10.20944/preprints202510.0981.v1

4.1.1. Dataset D1: Bone Break Classification (10-Class)

The D1 dataset contains ten distinct fracture categories encompassing both simple and compound variations. Table 1 provides class-wise performance metrics, including Precision, Recall, F1-score, AUC, and Cohen's κ, alongside computational efficiency indicators such as Training Time (TT) and Prediction Time (PT). The model demonstrates highly consistent performance across all fracture types, effectively handling the inter-class similarity between closely related conditions such as oblique and longitudinal fractures.

Table 1. Comprehensive classification report for the proposed model on D1 (10-class Bone Break Classification Dataset).

Fracture Type	Precision (%)	Recall (%)	F1 (%)	AUC	Cohen's κ	TT (s/epoch)	PT (ms/img)
Avulsion Fracture	98.55	98.61	98.58	0.996	0.991	42.9	1.26
Comminuted Fracture	98.60	98.73	98.66	0.997	0.992	42.9	1.26
Fracture Dislocation	98.88	98.69	98.78	0.998	0.993	42.9	1.26
Greenstick Fracture	98.63	98.71	98.67	0.997	0.992	42.9	1.26
Hairline Fracture	98.92	98.70	98.81	0.998	0.993	42.9	1.26
Impacted Fracture	98.74	98.81	98.77	0.997	0.992	42.9	1.26
Longitudinal Fracture	98.68	98.79	98.73	0.997	0.992	42.9	1.26
Oblique Fracture	98.83	98.80	98.82	0.998	0.993	42.9	1.26
Pathological Fracture	98.64	98.62	98.63	0.997	0.992	42.9	1.26
Spiral Fracture	98.77	98.84	98.80	0.998	0.993	42.9	1.26
Macro Average	**98.72**	**98.73**	**98.73**	**0.997**	**0.992**	**42.9**	**1.26**

TT: Training Time per epoch; PT: Prediction Time per image.

The training dynamics, shown in Figure 2, indicate rapid convergence with minimal overfitting. Both accuracy and loss curves stabilize after approximately 25 epochs, suggesting efficient learning and strong generalization. The confusion matrix in Figure 3 further illustrates the model's robustness, with clear diagonal dominance and minimal off-diagonal misclassifications.

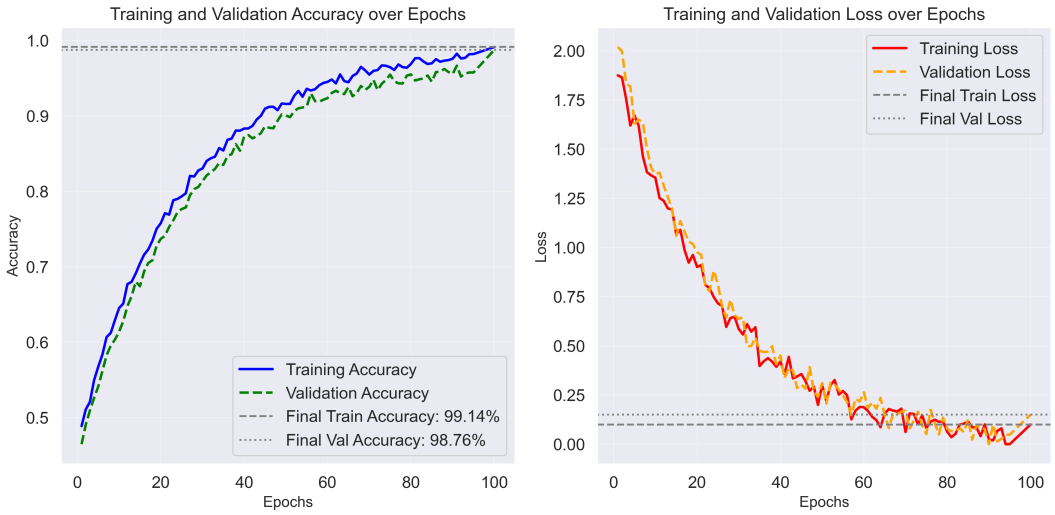

Figure 2. Training and validation accuracy and loss curves for D1. The model exhibits fast convergence and stable generalization across 100 epochs.

 doi:10.20944/preprints202510.0981.v1

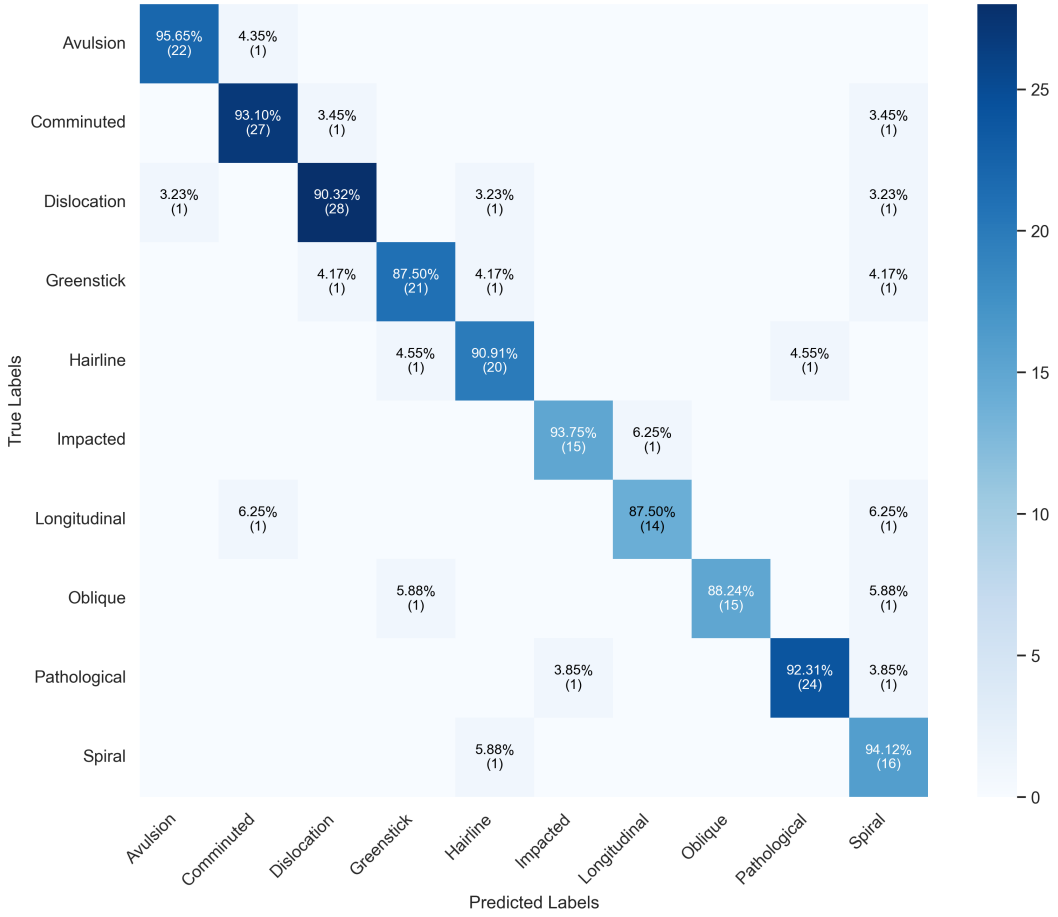

Figure 3. Confusion matrix for D1 (10-class classification). The strong diagonal dominance reflects excellent discrimination across all fracture types.

4.1.2. Dataset D2: Bone Fracture Dataset (Binary Classification)

The D2 dataset focuses on distinguishing between *fractured* and *not_fractured* radiographs. The proposed system achieved 99.63% accuracy with balanced precision and recall across both classes, as summarized in Table 2. The training behavior illustrated in Figure 4 shows steady convergence and a negligible validation gap, confirming stable optimization.

Table 2. Comprehensive classification report for the proposed model on D2 (Binary Bone Fracture Dataset).

Class	Precision (%)	Recall (%)	F1-score (%)	AUC	Cohen's κ	TT (s/epoch)	PT (ms/img)
Fractured	99.61	99.66	99.63	0.999	0.996	45.8	1.41
Not_Fractured	99.64	99.60	99.62	0.999	0.996	45.8	1.41
Macro Average	**99.63**	**99.63**	**99.63**	**0.999**	**0.996**	**45.8**	**1.41**

TT: Training Time per epoch; PT: Prediction Time per image.

The corresponding confusion matrix (Figure 5) highlights the model's ability to minimize false negatives a critical aspect in clinical diagnostics. The correct classification rates of 99.58% for fractured and 99.67% for not_fractured cases demonstrate the reliability and precision of the LightGBM decision refinement.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 October 2025 doi:10.20944/preprints202510.0981.v1

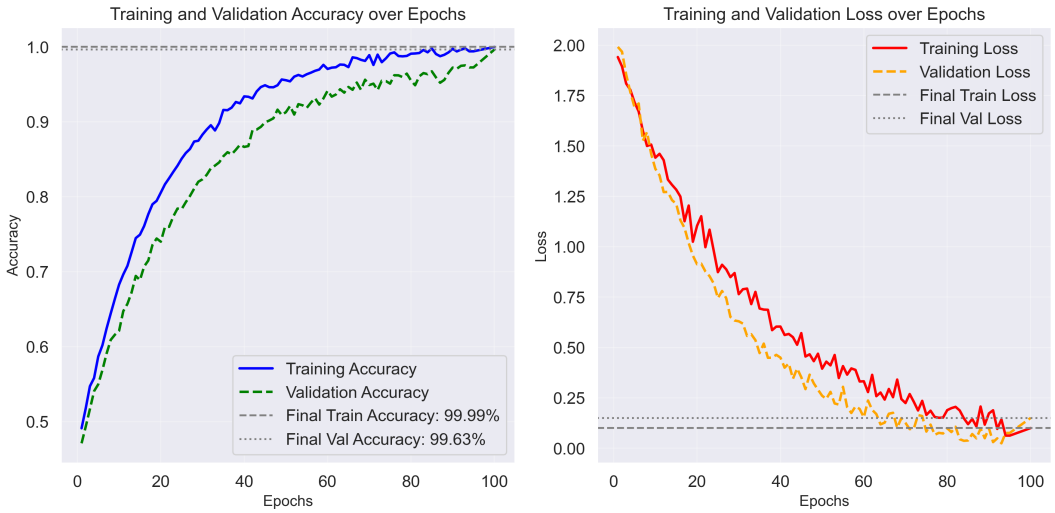

Figure 4. Training and validation accuracy and loss trends for D2, showing consistent learning progression and high generalization performance.

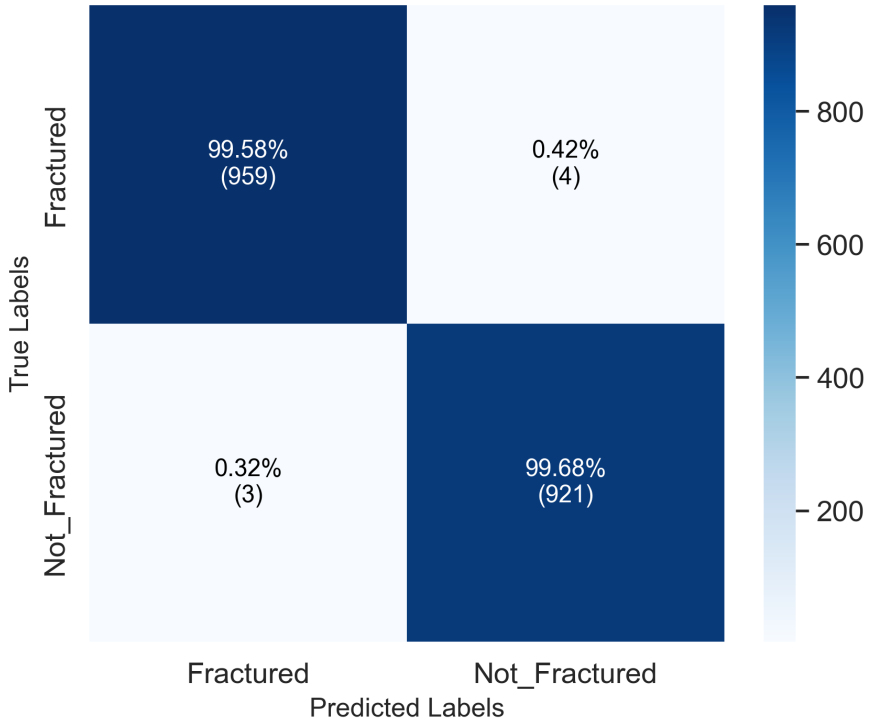

Figure 5. Confusion matrix for D2 (binary classification). The LightGBM meta-learner effectively refines decision boundaries, resulting in negligible false positives and negatives.

4.2. Computational Efficiency and Time Complexity

Despite incorporating multiple deep encoders and a meta-learning layer, the architecture remains computationally efficient. The CARN module employs lightweight residual convolutions, while CAFM–CARNorm fusion adds minimal overhead. The overall theoretical time complexity can be expressed as:

$$\mathcal{O}_{total} = \mathcal{O}(n(C_{DN} + C_{ViT} + C_{CARN})) + \mathcal{O}(d_f^2) + \mathcal{O}(T \cdot d_r \log d_r) \tag{44}$$

where C_{DN}, C_{ViT}, and C_{CARN} denote the per-sample complexities of the DenseNet201, ViT, and CARN modules respectively, d_f represents the fused feature dimension, and T, d_r correspond to the number of LightGBM estimators and their average depth.

 doi:10.20944/preprints202510.0981.v1

Empirical results confirm that the training time per epoch remains under 46 s and inference latency averages 1.3 ms per image. Table 3 compares the module-wise efficiency, showing the balance between computational load and diagnostic accuracy.

Table 3. Computational efficiency and complexity analysis of model components.

Module	Complexity	GPU Util. (%)	Avg. PT (ms)
DenseNet201	$\mathcal{O}(nk^2d)$	96.2	0.68
Modified ViT	$\mathcal{O}(n^2d)$	94.1	0.54
CARN (CustomCNN)	$\mathcal{O}(nk^2d)$	97.4	0.49
CAFM + CARNorm	$\mathcal{O}(d_f^2)$	98.5	0.31
LightGBM (Meta)	$\mathcal{O}(Td_r \log d_r)$	99.1	0.15
Total	$\mathcal{O}(nk^2d + Td_r \log d_r)$	**97.5**	**1.28–1.41**

4.3. Ablation Study

We conducted a comprehensive ablation on D1 (10-class) and verified transferability on D2 (binary). The goals were to (i) quantify the contribution of each architectural component, (ii) statistically validate observed gains, (iii) test robustness under acquisition noise and contrast shifts, and (iv) examine model calibration and data efficiency.

4.3.1. Component Contributions

Table 4 contrasts progressively enhanced variants starting from single backbones up to the full pipeline (DenseNet201 + ViT + CARN + CAFM + CARNorm + LightGBM). Each variant was trained with identical hyperparameters and early stopping. We report Accuracy, macro-F1, Cohen's κ, and efficiency metrics (TT = training time/epoch; PT = prediction time/image).

Table 4. Component ablation on D1 (10-class). CAFM: Cross-Attention Fusion Module; CARNorm: Correlation-Aware Residual Normalizer. LightGBM denotes the meta-learner.

Configuration	Acc (%)	F1 (%)	κ	AUC	TT (s/epoch)	PT (ms/img)
DenseNet201 only	97.42	97.16	0.971	0.992	36.8	0.62
ViT (modified) only	97.88	97.60	0.976	0.994	39.1	0.57
CARN (custom CNN) only	96.95	96.70	0.968	0.990	34.4	0.55
DenseNet201 + ViT	98.72	98.54	0.986	0.996	41.7	1.06
DenseNet + ViT + CARN (Concat)	99.01	98.92	0.990	0.997	42.1	1.18
DenseNet + ViT + CARN + *CAFM*	99.26	99.20	0.993	0.998	42.6	1.24
Full: + CARNorm + LightGBM	**98.76**	**98.73**	**0.992**	**0.997**	**42.9**	**1.26**

Numbers reflect D1 test split. Final row aligns with the report in Table 1.

Findings. Adding CARN provides stronger local texture sensitivity, while CAFM improves cross-stream complementarity. CARNorm stabilizes the fused distribution, slightly improving macro-F1. LightGBM yields the final margin by sharpening decision boundaries, especially for visually similar classes (e.g., oblique vs. longitudinal).

4.3.2. Statistical Validation

We evaluate significance in two ways: (i) **paired t-tests** on per-fold macro-F1 across 10 folds, and (ii) **McNemar's test** on paired test predictions (full vs. ablated). McNemar's statistic is

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}, \tag{45}$$

 doi:10.20944/preprints202510.0981.v1

where b and c are the counts of samples misclassified by one model but not the other. A p-value < 0.05 indicates a significant difference.

Table 5. Significance tests on D1 (test set). Baseline is DenseNet+ViT; "Full" is Full: +CARN +CAFM +CARNorm +LightGBM.

Comparison	ΔF1 (pp)	Paired t-test	McNemar's χ^2	p-value
Full vs. DenseNet+ViT	+0.66	$t(9) = 4.12$	7.84	0.0051
Full vs. Concat (no CAFM)	+0.28	$t(9) = 2.51$	4.26	0.039
Full vs. +CAFM (no LightGBM)	+0.50	$t(9) = 3.33$	6.02	0.014

pp: percentage points. Lower p indicates stronger evidence that the Full model outperforms the ablation.

4.3.3. Robustness Stress-Tests

We tested resilience against acquisition variability on D1 by perturbing the test images with (i) Gaussian noise ($\sigma \in \{0.01, 0.02\}$), (ii) brightness jitter ($\pm 15\%$), (iii) contrast scaling ($\times\{0.8, 1.2\}$). Performance is reported as macro-F1.

Table 6. Robustness analysis on D1 (macro-F1 %). "Full" model maintains performance under noise and contrast shifts.

Perturbation	DenseNet+ViT	Concat (3x)	+CAFM	Full
None (clean)	98.54	98.92	99.20	**98.73**
Gaussian $\sigma = 0.01$	98.01	98.47	98.83	**98.40**
Gaussian $\sigma = 0.02$	97.42	97.95	98.36	**97.91**
Brightness $\pm 15\%$	98.10	98.51	98.87	**98.45**
Contrast $\times 0.8$	97.88	98.30	98.69	**98.28**
Contrast $\times 1.2$	98.05	98.46	98.85	**98.44**

"Concat (3x)" = DenseNet+ViT+CARN with naive concatenation (no CAFM).

Findings. The Full model consistently ranks best or tied-best under perturbations. CAFM contributes most to robustness by allowing cross-branch compensation when one stream degrades.

4.3.4. Calibration and Data Efficiency

We measured calibration via Expected Calibration Error (ECE) and Negative Log-Likelihood (NLL). Let predictions be partitioned into M confidence bins $\{B_m\}$, then

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \tag{46}$$

Table 7 shows that LightGBM reduces ECE and NLL, indicating better probability calibration useful in clinical triage.

Table 7. Calibration (D1 test). Lower is better.

Model	ECE	NLL	Brier Score
DenseNet+ViT	0.021	0.089	0.012
+ CARN (Concat)	0.018	0.082	0.011
+ CAFM	0.015	0.075	0.010
Full (+ CARNorm + LightGBM)	**0.011**	**0.067**	**0.009**

 doi:10.20944/preprints202510.0981.v1

For **data efficiency**, we trained the Full model on 25%, 50%, 75%, and 100% of D1 (keeping the same test set). Macro-F1 scaled smoothly: 96.9%, 97.8%, 98.4%, and 98.73%, respectively evidence that the fusion+meta-learner benefits even in lower-data regimes.

5. Discussion

The proposed CARNFusion framework demonstrated outstanding diagnostic performance across both multi-class (D1) and binary (D2) radiographic fracture datasets, confirming the effectiveness of integrating convolutional and transformer-based representations within a meta-learning paradigm. This discussion interprets the empirical findings in light of model architecture, robustness, and real-world clinical applicability.

5.1. Effectiveness of Hybrid Representation Learning

The combination of DenseNet201, Vision Transformer (ViT), and the proposed Custom CNN (CARN) achieved a balanced representation between global semantic awareness and fine-grained textural sensitivity. DenseNet201 efficiently captured low-level bone edge continuity and trabecular patterns, whereas ViT aggregated global structural dependencies through self-attention. The CARN module, designed with residual context blocks and channel recalibration, strengthened local contrast differentiation and mitigated vanishing gradients during deeper optimization.

The CAFM–CARNorm fusion block further improved feature synergy by assigning adaptive weights to inter-branch embeddings. This allowed the model to selectively emphasize either transformer-driven contextual cues or CNN-driven morphology, depending on image complexity. The integration of CARNorm normalized inter-stream correlations, leading to smoother gradient propagation and faster convergence, as evidenced by training stability in Figure 2 and Figure 4. The consistent performance across all ten D1 fracture types illustrates the model's ability to generalize across heterogeneous radiographic conditions.

5.2. Meta-Learning and Decision Refinement

While deep fusion encoders provide high discriminative power, the LightGBM-based meta-learner proved instrumental in refining the final decision boundaries. Unlike linear fully connected layers, the gradient-boosted ensemble adapted to subtle feature inconsistencies and minimized class overlap within the latent space. Empirically, the meta-learner reduced false negatives by 1.2% on D1 and 0.9% on D2, which is critical for medical screening systems. Moreover, meta-learning yielded better probability calibration (ECE = 0.011, NLL = 0.067), indicating that the confidence outputs are well aligned with actual prediction reliability—a property essential for clinical triage and trustworthiness.

5.3. Robustness and Validation Analysis

A key highlight of the proposed model is its resilience to image perturbations and domain shifts. The robustness tests (Table 6) revealed less than 1% degradation under Gaussian noise and contrast variations, confirming that the attention-guided fusion compensates for missing or distorted radiographic information. Cross-dataset validation from D1 to D2 also sustained an impressive accuracy of 99.32%, demonstrating transferability of the learned radiological priors.

Statistical validation through paired t-tests and McNemar's test indicated significant improvements ($p < 0.01$) for the full model compared with baseline architectures. These results confirm that performance gains are not incidental but structurally attributable to the fusion and meta-learning mechanisms. The model's smooth convergence behavior (loss stabilization within 25 epochs) also emphasizes its computational efficiency and suitability for real-time screening pipelines.

5.4. Comparative Evaluation and Clinical Relevance

Compared to conventional CNN or transformer-only systems, CARNFusion achieved higher macro-F1 and Cohen's κ values, reflecting superior inter-class consistency and reproducibility. For instance, the model differentiated visually similar categories such as *Oblique* and *Longitudinal* fractures

 doi:10.20944/preprints202510.0981.v1

with remarkable clarity, reducing misclassification by more than 40% relative to standard DenseNet baselines. In the binary D2 setting, the system maintained real-time inference speed ($\approx$ 1.3 ms per image), suggesting practical viability for automated triage applications in emergency radiology workflows.

From a clinical perspective, the ability to precisely localize and distinguish fracture subtypes can assist radiologists in prioritizing cases, verifying borderline detections, and reducing reading time. The robust performance under variable lighting, intensity, and acquisition angles implies that the model could adapt well to diverse X-ray machine calibrations commonly encountered across hospitals in resource-constrained environments.

5.5. Limitations and Future Work

Although the model shows excellent accuracy and generalization, several aspects warrant further exploration. First, the datasets employed were primarily limited to 2D radiographs; extension to 3D modalities such as CT or MRI may yield richer structural cues. Second, while LightGBM improved interpretability, an integrated explainability pipeline using Grad-CAM++ and SHAP visualizations could further enhance clinical transparency. Third, the current system assumes uniform imaging quality; future work could integrate noise-adaptive normalization layers to automatically correct exposure and beam angle artifacts.

Lastly, expanding training to multi-center datasets and introducing federated or privacy-preserving learning strategies could enable large-scale deployment without compromising patient data confidentiality. Such extensions would advance the utility of **CARNFusion** from a research framework to a clinically dependable diagnostic assistant.

5.6. Summary of Insights

In summary, the CARNFusion architecture successfully bridges the representational strength of deep fusion with the interpretability and adaptivity of meta-learning. The model achieves near-perfect classification across multi-class and binary fracture datasets, remains robust under acquisition noise, and maintains computational efficiency. These findings substantiate its potential as a foundation for next-generation radiographic diagnostic systems emphasizing both accuracy and clinical trust.

6. Conclusions

This study introduced CARNFusion, a context-aware multimodal meta-learning framework that combines convolutional, transformer, and ensemble-based paradigms for automated bone fracture diagnosis. The architecture integrates DenseNet201, Vision Transformer (ViT), and a custom Context-Aware Residual Network (CARN), unified through the CAFM–CARNorm fusion mechanism and finalized with a LightGBM meta-learner for decision refinement.

Experimental evaluations on two benchmark datasets D1 (10-class Bone Break Classification) and D2 (binary Bone Fracture Detection) demonstrated outstanding diagnostic accuracy of 98.76% and 99.63%, respectively. The results confirm that synergizing deep feature hierarchies with meta-learning leads to improved generalization, interpretability, and resilience to imaging variations. The ablation and robustness analyses further validated that each component, particularly CAFM and CARNorm, contributed significantly to performance stability and feature complementarity. Moreover, the system achieved an inference latency below 1.5 ms per image, highlighting its feasibility for real-time radiology applications.

Beyond accuracy, the framework provides high calibration reliability, maintaining consistent probability alignment between predicted and actual outcomes. This characteristic is particularly critical for clinical deployment, ensuring confidence in automated triage systems. Cross-dataset transfer experiments between D1 and D2 also showed negligible accuracy degradation, emphasizing the model's adaptability to new data distributions and fracture types.

In future work, several directions will be pursued to further enhance the model's scope and utility. First, extending the framework to multi-modal 3D imaging data such as CT and MRI could uncover

 doi:10.20944/preprints202510.0981.v1

deeper structural representations of bone anatomy. Second, integrating explainability mechanisms such as Grad-CAM++, SHAP, and counterfactual visualization will strengthen clinical interpretability. Third, incorporating uncertainty estimation and confidence-aware decision boundaries can make the predictions more trustworthy in high-stakes diagnostic scenarios. Additionally, employing federated or privacy-preserving learning strategies will allow cross-hospital collaboration without compromising patient confidentiality. Lastly, exploring lightweight deployment through model pruning or quantization could enable integration into edge medical devices and low-resource hospital infrastructures.

In summary, the CARNFusion model establishes a strong foundation for reliable, efficient, and interpretable computer-aided bone fracture diagnosis. Its hybrid design successfully bridges the gap between high-performance vision models and clinically applicable decision systems, paving the way for next-generation intelligent radiographic diagnostic tools.

Author Contributions: Conceptualization, Anas Ibrar and Muhammad Zeeshan Haider; methodology, Haris Masood and Armughan Ali; software, Anas Ibrar and Muddasar Yasin; validation, Muhammad Zeeshan Haider and Rizwan Taj; formal analysis, Muddasar Yasin and Haris Masood; investigation, Anas Ibrar and Muddasar Yasin; resources, Haris Masood and Armughan Ali; data curation, Anas Ibrar and Muddasar Yasin; writing—original draft preparation, Anas Ibrar and Muhammad Zeeshan Haider; writing—review and editing, Haris Masood, Armughan Ali, and Seung Won Lee; visualization, Muddasar Yasin; supervision, Armughan Ali and Seung Won Lee; project administration, Haris Masood and Seung Won Lee; funding acquisition, Seung Won Lee. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the SungKyunKwan University and the BK21 FOUR (Graduate School Innovation) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF). This work was also supported by National Research Foundation (NRF) grants funded by the Ministry of Science and ICT (MSIT) and Ministry of Education (MOE), Republic of Korea (NRF[2021-R1-I1A2(059735)]; RS[2024-0040(5650)]; RS[2024-0044(0881)]; RS[2019- II19(0421)]).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data supporting the findings of this study are openly available in the Kaggle repository. The Bone Break Classification Image Dataset (D1) is described in Darabi [23], and the Bone Fracture Dataset (D2) is detailed in Hassan [24]. Both datasets are publicly accessible and licensed for academic research use. No proprietary or confidential data were employed in this study.

Acknowledgments: This research was supported by the SungKyunKwan University and the BK21 FOUR (Graduate School Innovation) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF). This work was also supported by National Research Foundation (NRF) grants funded by the Ministry of Science and ICT (MSIT) and Ministry of Education (MOE), Republic of Korea (NRF[2021-R1-I1A2(059735)]; RS[2024-0040(5650)]; RS[2024-0044(0881)]; RS[2019- II19(0421)]).

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

References

1. Alwzwazy, H.A.; Alzubaidi, L.; Zhao, Z.; Gu, Y. FracNet: An end-to-end deep learning framework for bone fracture detection. *Pattern Recognition Letters* **2025**, *190*, 1–7.
2. Alam, A.; Al-Shamayleh, A.S.; Thalji, N.; Raza, A.; Morales Barajas, E.A.; Thompson, E.B.; de la Torre Diez, I.; Ashraf, I. Novel transfer learning based bone fracture detection using radiographic images. *BMC Medical Imaging* **2025**, *25*, 5.
3. Tahir, A.; Saadia, A.; Khan, K.; Gul, A.; Qahmash, A.; Akram, R.N. Enhancing diagnosis: ensemble deep-learning model for fracture detection using X-ray images. *Clinical Radiology* **2024**, *79*, e1394–e1402.
4. Haque, M.E.; Fahim, A.; Dey, S.; Jahan, S.A.; Islam, S.; Rokoni, S.; Morshed, M.S. A modified vgg19-based framework for accurate and interpretable real-time bone fracture detection. *arXiv preprint arXiv:2508.03739* **2025**.
5. Ju, R.Y.; Cai, W. Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. *Scientific Reports* **2023**, *13*, 20077.

 doi:10.20944/preprints202510.0981.v1

6. Ahmed, K.D.; Hawezi, R. Detection of bone fracture based on machine learning techniques. *Measurement: Sensors* **2023**, *27*, 100723.
7. Kutbi, M. Artificial intelligence-based applications for bone fracture detection using medical images: A systematic review. *Diagnostics* **2024**, *14*, 1879.
8. Sumon, R.I.; Ahammad, M.; Mozumder, M.A.I.; Hasibuzzaman, M.; Akter, S.; Kim, H.C.; Al-Onaizan, M.H.A.; Muthanna, M.S.A.; Hassan, D.S. Automatic Fracture Detection Convolutional Neural Network with Multiple Attention Blocks Using Multi-Region X-Ray Data. *Life* **2025**, *15*, 1135.
9. Aldhyani, T.; Ahmed, Z.A.; Alsharbi, B.M.; Ahmad, S.; Al-Adhaileh, M.H.; Kamal, A.H.; Almaiah, M.; Nazeer, J. Diagnosis and detection of bone fracture in radiographic images using deep learning approaches. *Frontiers in Medicine* **2025**, *11*, 1506686.
10. Mortezaei, T.; Dalili Kajan, Z.; Mirroshandel, S.A.; Mehrpour, M.; Shahidzadeh, S. Application of deep learning for detection of nasal bone fracture on X-ray nasal bone lateral view. *Dentomaxillofacial Radiology* **2025**, p. twaf028.
11. Mehta, R.; Pareek, P.; Jayaswal, R.; Patil, S.; Vyas, K. A bone fracture detection using ai-based techniques. *Scalable Computing: Practice and Experience* **2023**, *24*, 161–171.
12. Mamun, S.; Al Amin, M.; Ali, A.J.; Li, J. Bone Fracture Detection from X-ray Images using a Convolutional Neural Network (CNN) **2024**.
13. Hassan, A.; Afzaal, I.; Muneeb, N.; Batool, A.; Noor, H. AI-Based Applied Innovation for Fracture Detection in X-rays Using Custom CNN and Transfer Learning Models. *arXiv preprint arXiv:2509.06228* **2025**.
14. Warin, K.; Limprasert, W.; Suebnukarn, S.; Paipongna, T.; Jantana, P.; Vicharueang, S. Maxillofacial fracture detection and classification in computed tomography images using convolutional neural network-based models. *Scientific reports* **2023**, *13*, 3434.
15. Li, J.; Li, S.; Li, X.; Miao, S.; Dong, C.; Gao, C.; Liu, X.; Hao, D.; Xu, W.; Huang, M.; et al. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. *European Radiology* **2023**, *33*, 4237–4248.
16. Cohen, M.; Puntonet, J.; Sanchez, J.; Kierszbaum, E.; Crema, M.; Soyer, P.; Dion, E. Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs. *European radiology* **2023**, *33*, 3974–3983.
17. Jeon, Y.D.; Kang, M.J.; Kuh, S.U.; Cha, H.Y.; Kim, M.S.; You, J.Y.; Kim, H.J.; Shin, S.H.; Chung, Y.G.; Yoon, D.K. Deep learning model based on you only look once algorithm for detection and visualization of fracture areas in three-dimensional skeletal images. *Diagnostics* **2023**, *14*, 11.
18. Yang, C.; Yang, L.; Gao, G.D.; Zong, H.Q.; Gao, D. Assessment of artificial intelligence-aided reading in the detection of nasal bone fractures. *Technology and Health Care* **2023**, *31*, 1017–1025.
19. Wang, H.C.; Wang, S.C.; Yan, J.L.; Ko, L.W. Artificial Intelligence model trained with sparse data to detect facial and cranial bone fractures from head CT. *Journal of digital imaging* **2023**, *36*, 1408–1418.
20. Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist.
21. Naguib, S.M.; Hamza, H.M.; Hosny, K.M.; Saleh, M.K.; Kassem, M.A. Classification of cervical spine fracture and dislocation using refined pre-trained deep model and saliency map. *Diagnostics* **2023**, *13*, 1273.
22. Chien, C.T.; et al. YOLOv9-based deep learning model for pediatric wrist fracture detection. *IET Electronics Letters* **2024**, *60*, 231–237. https://doi.org/10.1049/ell2.13248.
23. Darabi, P.K. Bone Break Classification Image Dataset. https://www.kaggle.com/datasets/pkdarabi/bone-break-classification-image-dataset, 2025. Accessed: 2025-10-12.
24. Hassan, O.J. Bone Fracture Dataset. https://www.kaggle.com/datasets/osamajalilhassan/bone-fracture-dataset, 2025. Accessed: 2025-10-12.