

XBFD-Net: A Unified Explainable Deep Learning Framework for Bone Fracture Detection with Dual Attention and Quantitative Localization

Galiveeti Poornima^{1*}[0000-0003-4289-114X] and Manju Bargavi Ranjeth²[0000-0001-7112-5810]

¹ Lincoln University, Malaysia.

² Department of Computer Science and IT, Jain (Deemed-to-be University), Bangalore
pdf.Galiveeti@lincoln.edu.my

Abstract. Existing deep learning approaches for bone fracture detection from plain radiographs address either detection or explainability in isolation, lacking a unified, clinically deployable framework. We present XBFD-Net, a novel end-to-end architecture that simultaneously addresses fracture detection, transparent localization, and quantitative explainability evaluation — three capabilities not previously combined in a single system. Firstly, we introduced a Dual Attention Fusion (DAF) module that utilizes the channel attention to suppress irrelevant areas of the image and spatial attention to focus on anatomical structures relevant to fracture recognition during the feature extraction stage. Unlike previous studies that evaluate explainability only qualitatively, we propose a quantitative localization benchmark based on Intersection over Union (IoU) and Dice coefficient with respect to radiologist-annotated fracture masks. XBFD-Net outperforms multiple popular methods, including Grad-CAM, Score-CAM, and attention-only baselines by up to 15.2 IoU points in detection accuracy (93.4% recall, AUC: 0.96) and localization performance (IoU: 56.8%, Dice: 72.4%) across datasets against datasets. Qualitative analysis by the radiologists demonstrated that heatmaps delineated cortical disruptions (eg, small hairline fractures) accurately.

Keywords: Bone fracture detection; explainable AI; dual attention; Grad-CAM; IoU; Dice; localization; clinical decision support; deep learning; radiology.

1 Introduction

Bone fractures are among the most critical and time-sensitive injuries diagnosed in emergency radiology. Accurate interpretation of plain radiographs—the primary imaging modality for fracture assessment—requires expertise that is unevenly distributed across healthcare settings. Deep learning models have demonstrated promising detection performance, yet their adoption in clinical environments remains limited by two well-documented barriers: the black-box nature of model predictions and the absence of standardized, quantitative methods to verify that model attention aligns with clinically relevant anatomy.

A survey of existing literature finds a fragmented research landscape: optimization of detection models focuses on classification accuracy and not evaluating localization; explainability methods are applied in a post-hoc manner and only qualitatively assessed; benchmarking frameworks vary between studies, making reproducibility challenging. All existing systems have been limited to the integration of detection, attention-driven additive feature augmentation, and quantitative explainability assessment—but without a unified framework that bestows reproducibility for use in clinical decision-support applications.

In this paper, we propose a method, XBFD-Net (eXplainable Bone Fracture Detection Network), to fill the gap. The primary contributions are:

- DAF: A Dual Attention Fusion (DAF) Module that integrates channel and spatial attention mechanisms directly on the classification backbone to improve detection accuracy and feature interpretability.
- A quantitative localization evaluation protocol using IoU and Dice coefficient, benchmarked against radiologist-annotated fracture regions — the first such evaluation in the bone fracture detection literature.
- A unified architecture serving as both a high-performance classifier and a clinical decision support tool, with heatmap outputs designed for direct radiologist consumption.

XBFD-Net outperforms four baselines—ResNet-50, DenseNet-121, Vision Transformer, and the strongest prior explainability method (Attention-only)—on both detection and localization metrics, while preserving detection performance across fracture types, including subtle hairline fractures.

2 Related Work

2.1 Fracture Detection with Deep Learning

The highest detection accuracies (86--92%) on single-region fracture datasets [1, 2] have been achieved via transfer learning from ImageNet-pretrained CNNs. More recently, Vision transformers (ViTs) have been adapted to musculoskeletal imaging, obtaining slightly improved performance (90.7%) owing to their global self-attention mechanism; at the expense of interpretability and computation [3]. Importantly, none of these systems assesses the spatial correspondence of their feature maps with actual fracture sites.

2.2 Explainability in Medical Imaging

This is the main difference between related post-hoc explainability methods, which, like Grad-CAM [4], Score-CAM [5], and LIME [6], give class-discriminative heatmaps but an operator on a frozen model with no impact on prediction. Then, attention-based architectures that build the attention

maps into the models calculate this in a more aligned fashion with clinical features [7]. However, all existing studies on explainability in fracture detection are qualitatively radiologist-scored optimizer-based highlights and therefore are inherently subjective. This is the first work on pixel-level IoU and Dice metrics comparison to annotated masks about localization quality.

2.3 Research Gaps Addressed

Thus, the motivation for XBFD-Net is three-fold: (i) lack of a holistic detection-explainability framework; (ii) lack of quantitative localization metrics in fracture AI evaluation; (iii) without a prior system as a clinical decision support tool compared to an evaluator alone.

3 XBFD-Net Architecture

3.1 Backbone

DenseNet-121 is adopted as the backbone of the feature extractor for XBFD-Net, which performs commendably in fine-grained texture recognition that is critical to cortical bone evaluation. It is a decoder: The backbone is pre-trained on ImageNet, and then the final two dense blocks are unfrozen for training. Images that are fed in are resized to 224x224 pixels and have pixel intensities between 0 and 1.

3.2 Novel Contribution — Dual Attention Fusion (DAF) Module

The DAF module is inserted after the final dense block and before global average pooling. It consists of two sequentially applied attention sub-modules:

Channel Attention Sub-module: Given the feature map $F \in \mathbb{R}^{(H \times W \times C)}$, at this level, global average pooling and global max pooling are performed along the spatial dimension to provide two channel descriptors. They are fed through a shared two-layer MLP (reduction ratio $r = 16$), and the outputs of these are summed up through sigmoid activation to return a channel attention vector $M_c \in \mathbb{R}^C$, which means that the input feature map is recalibrated as $F' = M_c \cdot F$, suppressing non-diagnostic channels (e.g., soft tissue signal) and amplifying bone-structural channels.

Spatial Attention Sub-module: At this level, global average pooling and global max pooling are performed along the spatial dimension to provide two-channel descriptors. They are fed through a shared two-layer MLP (reduction ratio $r = 16$), and the outputs of these are summed up through sigmoid activation to return a channel attention vector $M_c \in \mathbb{R}^C$, which means that the input feature map is recalibrated as $F' = M_c \cdot F$, suppressing non-diagnostic channels (e.g., soft tissue signal) and amplifying bone-structural channels.

The DAF output F'' is directly used to generate the Grad-CAM heatmap, ensuring that the explainability map reflects the attention-refined features

4

rather than raw backbone activations. This produces tighter, more anatomically precise localization compared to standard Grad-CAM applied to a non-attention backbone.

3.3 Classification Head

F'' is then fed into global average pooling, which is followed by a dropout ($p = 0.5$) layer and a fully connected sigmoid output for binary fracture classification. Training uses binary cross-entropy loss, Adam optimizer ($\text{lr} = 1e[-4]$), batch size: 32, and early stopping on validation loss with patience 10.

3.4 Quantitative Localization Evaluation Protocol

We had a board-certified radiologist annotate pixel-level fracture boundary masks on our subset of 500 radiographs to test whether attention maps highlight fracture regions. Predicted heatmaps thresholded at 50% of peak activation. Three measures are computed:

- **Dice Coefficient:** $2|P \cap G| / (|P| + |G|)$, measuring F1-like spatial overlap.
- **Point Error (px):** Euclidean distance between the centroid of P and the centroid of G , measuring localization precision.
- **Intersection over Union (IoU):** $|P \cap G| / |P \cup G|$, where P = predicted mask, G = ground-truth mask.

4 Experimental Setup

Datasets: We use the Kaggle Bone Fracture X-ray dataset (10,580 x-rays; broken / not broken over limbs, hip, spine, knee) as the main evaluation corpus. Cross-dataset generalization testing is advanced by the Stanford MURA dataset (40,561 images; 7 upper-extremity regions). All splits are stratified by label and region: 70/15/15% 5-fold cross-validation is applied throughout.

Baselines: Four baselines are evaluated: ResNet-50 (fine-tuned, no attention), DenseNet-121 (fine-tuned, no attention), Vision Transformer (ViT-B/16, fine-tuned), and an Attention-only variant of XBFD-Net without the DAF module (using only Grad-CAM post-hoc). XBFD-Net (full DAF + quantitative protocol) is compared against all four.

5 Results

5.1 Detection Performance

Table 1. Fracture detection performance (5-fold CV). XBFD-Net achieves consistent gains across all metrics.

Model	Accuracy (%)	Sensitivity (%)	AUC
ResNet-50	89.2	86.4	0.91
DenseNet-121	90.1	87.6	0.92

			5
Vision Transformer (ViT-B/16)	90.7	88.1	0.93
XBFD-Net (Proposed)	93.4	91.9	0.96

5.2 Localization Performance

Table 2. Quantitative localization results against radiologist-annotated fracture masks (n = 500). Lower Point Error is better.

Method	IoU (%)	Dice (%)	Point Error (px)
Grad-CAM	41.6	58.7	18.4
Score-CAM	45.9	61.2	15.7
Attention-only	49.3	65.1	13.6
XBFD-Net (DAF + Grad-CAM)	56.8	72.4	9.8

XBFD-Net achieves an IoU of 56.8% and Dice of 72.4%, surpassing the best prior method (Attention-only) by 7.5 IoU points and 7.3 Dice points. Point Error is reduced from 13.6 px to 9.8 px, representing a 28% improvement in localization precision. These gains are attributable to the DAF module's integration of attention directly into the feature pathway, rather than applying explainability post-hoc on a non-attention backbone.

5.3 Key Observations

- **Explainability does not degrade detection:** XBFD-Net obtains the best per-class and mean detection accuracy (93.4% @0.5 M-AP is irrespective of the AT in appearance domain, while 79.2 - 41.7%) outperforming other attention models as well as with respect to localization scores alone dis-proving a hill belief that attention mechanism sacrifices classification per-formance for the sake of localization scores { $\text{Height}(\hat{x}) > \text{Min Height}(x)$ } + { $\text{Width}(\hat{x}) > \text{Min Width}(x)$ }, x – predicted boxes; o – horizontal offset (+ve value = left centre).
- **Hairline fracture sensitivity:** XBFD-Net lowers false negatives on hairline fracture by 22.3% (relative to DenseNet-121 baseline) due to the cortical dis-continuities imaged being focused upon using the spatial attention module.
- **Clinical alignment:** In a radiologist's review of 200 heatmaps, 91.4% were rated as clinically informative, and 87.2% were rated as aligned with the radi-ologist's own area of diagnostic focus.

6 Discussion

When attention is embedded into the architecture versus a post-hoc application, XBFD-Net shows that explainability and detection performance are not competing ob-jectives. The DAF module generates heatmaps that are both more accurate (higher IoU/Dice) and interpretable (lower point error, higher radiologist agreement) than any post-hoc method tested.

The introduction of a quantitative localization evaluation protocol is a methodological contribution independent of the architecture. By providing IoU and Dice benchmarks against annotated fracture masks, this paper establishes a reproducible standard for comparing explainability methods in fracture detection — a standard that has been conspicuously absent from the

literature. We make the annotation protocol and evaluation code publicly available to encourage adoption.

Limitations: The localization evaluation is based on 500 annotated images from a single radiologist; multi-reader annotation with inter-rater reliability assessment is planned for the journal version of this work. Additionally, XBFD-Net is currently evaluated on binary classification; extension to fracture type and severity grading is a priority for future work.

7 Conclusion

We introduced XBFD-Net, a novel and unified explainable deep learning framework for bone fracture detection with three contributions: the Dual Attention Fusion (DAF) module, a quantitative localization evaluation protocol for model interpretability assessment, and an efficient output pipeline of a clinical decision support system. Our framework exhibits 93.4% detection accuracy, 0.96 AUC performance, and state-of-the-art localization (IoU 56.8%; Dice 72.4%) against radiologist-annotated fracture masks. XBFD-Net fills the three gaps in the current literature: detection-explainability separation, absence of quantitative localization metrics, and lack of clinical workflow integration, providing a relevant foundation for safe auditable AI deployment in emergency radiology.

References

- [1] Meena, T., & Roy, S. (2022). Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. *Diagnostics*, 12(10), 2420.
- [2] Uysal, F., Hardalaç, F., Peker, O., Tolunay, T., & Tokgöz, N. (2021). Classification of shoulder X-ray images with deep learning ensemble models. *Applied Sciences*, 11(6), 2723.
- [3] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021*.
- [4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 336–359.
- [5] Wang, H., et al. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. *CVPR Workshops 2020*.
- [6] Sheu, R.-K., & Pardeshi, M. S. (2022). A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction, and scoring system. *Sensors*, 22(20), 8068.
- [7] Shi, W., Tong, L., Zhu, Y., & Wang, M. D. (2021). COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks. *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2376–2387.
- [8] Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of

explainable AI techniques in healthcare. *Sensors*, 23(2), 634.

- [9] Parvin, S., & Rahman, A. (2024). A real-time human bone fracture detection and classification from multi-modal images using a deep learning technique. *Applied Intelligence*, 54(19), 9269–9285.
- [10] Magnéli, M., et al. (2023). Deep learning classification of shoulder fractures on plain radiographs of the humerus, scapula, and clavicle. *PLOS ONE*, 18(8), e0289808.