

FractureNet: A Region-Aware Ensemble with Adaptive Hybrid Augmentation for Multi-Anatomical Bone Fracture Detection

Galiveeti Poornima^{1*} and Manju Bargavi Ranjeth²

¹ School of Advanced Studies, S-Vyasa University, Bangalore; ² Department of Computer Science and IT, Jain (Deemed-to-be University), Bangalore
pdf.Galiveeti@lincoln.edu.my

Abstract: Two outstanding challenges for automated bone fracture detection from plain radiographs are (1) limited anatomical region representation in training data, and (2) limited model generalization to clinical-grade image variability [4,5]. We propose a new framework, FractureNet, to tackle both challenges simultaneously. Specifically, we present Adaptive Hybrid Augmentation (AHA): a physics-informed augmentation policy that applies adaptive transformations at each region of interest in the image—beam-hardening simulation, trabecular texture synthesis, and stochastic occlusion masking—calibrated to the anatomy depicted in each radiograph. The second contribution is a Region-Aware Ensemble (RAE) approach, which includes DenseNet121, ResNet50, and one lightweight MobileNetV2 branch, integrated via an anatomy-conditioned, learnable gating network to incorporate local knowledge into the main CNN branches. Evaluated on aggregate 51,141 radiographs (including Kaggle Bone Fracture [1], Stanford MURA [2], and RSNA Bone Challenge datasets), FractureNet achieves 93.7% accuracy and an area under the ROC curve (AUC) score of 0.961, outperforming single-architecture baselines by +3.2–6.8 percentage points with AUC as the headline discrimination metric. We show that it is possible to rapidly obtain visualizations of the bone fracture location from both Grad-CAM and LIME visualization maps, providing evidence of clinically coherent localization choices that should aid in improving the trust and auditability of such emerging AI methods in emergency radiology workflows.

Keywords: Bone fracture identification; Deep learning; CNN; Data augmentation; Ensemble methods; Explainable artificial intelligence; Grad-CAM; Medical imaging analysis.

Introduction

Background and purpose: Bone fractures are the most common injury treated in emergency departments worldwide, with an estimated 178 million new fractures occurring each year [1]. On the other hand, radiographs need an accurate and fast interpretation, as a delay in fracture diagnosis or missed diagnosis can lead to complications (malunion, avascular necrosis & prolonged disability). The availability of radiologists is often a limiting factor in many high-throughput trauma environments.

Deep learning on homogeneous datasets has been accelerated by large-scale training of Convolutional Neural Networks (CNNs), achieving near-human performance on specific radiographic tasks, most notably wrist and vertebral fracture detection. However, optimal mass clinical deployment can only be pursued in consideration of two important limitations: (i) standard augmentation pipelines fail to take into account imaging physics-driven variability across different anatomical regions, e.g., cortical bone density in long

bones vs. trabecular patterning in vertebrae; and (ii) single-architecture models are highly task-specific with large heterogeneity across body regions.

In this paper, we introduce FractureNet, a novel general framework with two contributions that directly address these limitations. The rest of the paper is organized as follows: Section 2 presents an overview of related work; Section 3 discusses datasets; Section 4 describes the FractureNet architecture; Section 5 details the experimental design; Section 6 reports results; and Section 7 discusses future directions.

Related Work

Deep Learning for Fracture Detection

Lindsey et al. were among the earliest to use CNNs for fracture detection [2], showing that clinicians can detect occult wrist fractures more accurately with a deep learning model. Later work built on this, with models pre-trained on ImageNet (VGG16 [3], ResNet [4], DenseNet [5]) for fracture classification achieving 85–91% accuracy across seven separate regions using transfer learning. Stanford's MURA dataset [6] was one of the first to allow systematic multi-region evaluation and is also one of the longest-running benchmarks. Recent reviews have documented the explosion of development in compartment-based deep-learning fracture detection over anatomies and imaging modalities [11,12]. This has been used as a basis for object-detection and supervised-learning pipelines to improve wrist and multi-region fracture detection [13,14,15]. Concurrently, transformer-type architectures with scale- and spatial-aware attention are competitive with pure CNN backbones [17].

Data Augmentation in Medical Imaging

General-purpose augmentations (such as random flips, rotations, color jitter) are applied to all data. Nevertheless, it has known limitations with respect to radiographic data: a transformation that is semantically neutral for natural images may encode anatomically implausible transformations for an X-ray. We investigate domain-specific causes of much recent work; e.g., CycleGAN-based style transfer [7] and MixUp variations do not condition on anatomy at all. To the best of our knowledge, none have discussed data augmentation policies conditioned on anatomical regions during inference.

Ensemble Methods in Radiology AI

In medical imaging challenges (e.g., CheXpert and RSNA challenges), model ensembling has consistently improved performance across all folds. The standard methods apply linear combinations with fixed weights or majority voting, an approach shown to improve robustness in musculoskeletal radiograph classification [16]. This work fills a gap not previously examined in multi-anatomical fracture detection: a learnable ensemble gating conditioned on image metadata.

Datasets

Table 1. Dataset summary used in FractureNet evaluation.

Dataset	Images	Regions	Labels
Kaggle Bone Fracture	10,580	Limbs, Hip, Spine, Knee	Fractured / Not Fractured
Stanford MURA	40,561	7 upper-extremity regions	Normal / Abnormal
RSNA Bone Challenge	~9,000 (pediatric)	Wrist	Fracture / No Fracture
Total (de-duplicated)	51,141	11 anatomical regions	Binary per study

Beyond the datasets used in this study, recently released annotated benchmarks such as FracAtlas [18]—comprising 4,083 musculoskeletal radiographs labeled for fracture classification, localization, and segmentation—provide additional resources for evaluating multi-anatomical fracture detection and localization.

FractureNet: Proposed Framework

Novel Contribution 1 — Adaptive Hybrid Augmentation (AHA)

This means that the standard augmentation pipelines apply fixed, region-agnostic transforms. AHA presents a two-step manikin anatomical conditioning mechanism as follows:

Stage 1 — Region Prediction: Given the input radiograph, a lightweight EfficientNet-B0 classifier (pre-trained, frozen) predicts one of 11 anatomy region labels with a top-1 accuracy of 94.1%. It is then used to sample a region-specific augmenting policy.

Stage 2 — Physics-Informed Transforms: A region policy specifies a weighted probability distribution over the subsequent transforms calibrated to the known imaging physics of that region

Beam-hardening simulation — pixel-intensity polynomial warping simulating kVp variation (± 15 kVp range)¶ • e.g. for dense cortical regions (femur, tibia)

Synthesis of material: The material settings for the fabricated C-3 bone are done via trabecular texture generation - a Perlin-noise is overlaid on top of all cancellous samples from these bones, where the use of 3D Minkowski-based shape description and fiber distribution vectors, which are derived using 2D histograms [4], makes this method necessary.

Stochastic occlusion masking: applied with random rectangular masks of between 5 and 15% (of image area) to simulate overlapping soft tissue.

Elastic deformation: Thin-plate spline warping ($\alpha = 34$, $\sigma = 4$) — In pediatric wrist images, we tend to prioritize this technique in our image augmentation tool as it simulates growth-plate variation.

AHA is applied only during training; for inference, the original preprocessed image is used. The proposed strategy tackles the diversity shortcoming present in previous work, and does so while providing aetiologically justifiable augmentations without introducing anatomically unrealistic artifacts.

Novel Contribution 2 — Region-Aware Ensemble (RAE)

RAE is a learnable approach to weighted model predictions, which conditions ensemble weights on the predicted anatomical region token r rather than averaging static weights:

$$\hat{y} = G(r) \cdot [f_{\text{Dense}}(x), f_{\text{Res}}(x), f_{\text{Mobile}}(x)]^T$$

where f_{Dense} , f_{Res} , and f_{Mobile} are the softmax outputs of DenseNet121, ResNet50, and MobileNetV2, respectively, and $G(r) \in \mathbb{R}^3$ is a softmax-normalized weight vector output by a two-layer MLP conditioned on the one-hot region embedding r [105].

The reasoning is that each architecture may have different complementary strengths: DenseNet121 was optimal for vertebral trabecular detail (fine-grained texture patterns); ResNet50 for long-bone shaft fractures; and MobileNetV2 for the best generalization to pediatric wrist cases. RAE makes use of this complementarity in a dynamically resilient manner based on the region of input.

Explainability Module

For each backbone, Grad-CAM heatmaps are generated from the final convolutional layer. An ensemble-level attribution map consists of a summed-out version of all individual Grad-CAM maps weighted by their attention weights, resulting in a single spatially coherent visualization based on the ensemble decision. Superpixel Attribution with LIME: Another kind of attribution for the radiologist review, which gives a complementary explanation of regions.

Experimental Design

Baselines

We evaluate four baselines: (1) base CNN standard, (2) fine-tuned VGG16, (3) fine-tuned DenseNet121, and (4) fine-tuned ResNet50 with standard augmentation. We evaluate FractureNet (AHA + RAE) against all four.

Training Protocol

All models used the Adam optimizer (learning rate = 0.0001), a batch size of 32, and were trained for 50 epochs, with a dropout rate of 0.5, L2 regularization, and early stopping based on validation loss with a patience of 10. Hardware: NVIDIA A100 GPU, CUDA version 12.1. Validation: stratified 5-fold cross-validation ensuring no data leakage between folds.

Evaluation Metrics

Key metrics: accuracy, sensitivity, specificity, F1-score, and AUC-ROC. Secondary metrics: inference latency (ms/image) and model parameter count for evaluating deployment feasibility.

Results and Discussion

Table 2. Performance comparison on the combined test set (5-fold CV mean). Bold denotes the best result.

Model	Accuracy	Sensitivity	Specificity	F1	AUC
Baseline CNN	86.9%	84.2%	89.3%	0.843	0.901
VGG16	87.4%	85.1%	89.7%	0.851	0.913
ResNet50	90.5%	89.0%	92.1%	0.890	0.941
DenseNet121	91.2%	90.3%	92.0%	0.903	0.948
FractureNet (Ours)	93.7%	93.1%	94.4%	0.933	0.961

Overall, FractureNet achieves 93.7% accuracy, which outperforms the strongest specific single-architecture baseline (DenseNet121 at 91.2%) by 2.5 percentage points and the weakest multi-class baseline (baseline CNN at 86.9%) by an impressive margin of 6.8 percentage points. An AUC of 0.961 indicates strong discrimination across the entire operating-point spectrum. An ablation study confirms that AHA and RAE work independently: AHA alone is +1.4% over DenseNet121 (92.6%), and RAE increases the score by a further +0.8%(92.0%); synergistically combined, they yield an additional +2.5%.

A qualitative analysis of Grad-CAM maps, evaluated by a consulting radiologist, established that FractureNet consistently focuses on cortical discontinuities and periosteal reactions — the principal radiographic indicators of fracture — in 91.8% of accurately classified cases, in contrast to 78.3% for the DenseNet121 baseline.

Conclusion

Through two key foundational contributions, Adaptive Hybrid Augmentation (AHA), a physics-informed augmentation approach conditioned to biological and mechanical regions in the anatomy-association space; and Region-Aware Ensemble (RAE) with learnable anatomical gating, we introduced FractureNet —the novel first framework for deep learning-based multi-anatomical bone fracture detection. FractureNet outperforms the state-of-the-art on a large dataset with 51,141 images and is able to generate Clinically Testable Grad-CAM explanations. Next, clinical validation toward promising implementation to be applied in emergency department of the hospital and earlier progress to fracture severity classification.

References

1. GBD 2019 Fracture Collaborators, "Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019", *The Lancet Healthy Longevity*, vol. 2, no. 9, pp. e580–e592, 2021.
2. R. Lindsey et al., "Deep neural network improves fracture detection by clinicians", *PNAS*, vol. 115, no. 45, pp. 11591–11596, 2018.
3. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in *ICLR 2015*, 2015.

4. K. He et al., "Deep residual learning for image recognition", in *CVPR 2016*, pp. 770–778, 2016.
5. G. Huang et al., "Densely connected convolutional networks", in *CVPR 2017*, pp. 4700–4708, 2017.
6. P. Rajpurkar et al., "MURA: Large dataset for abnormality detection in musculoskeletal radiographs", *arXiv preprint*, arXiv:1712.06957, 2017.
7. J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks", in *ICCV 2017*, pp. 2223–2232, 2017.
8. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization", in *ICCV 2017*, pp. 618–626, 2017.
9. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks", in *ICML 2019*, 2019.
10. M. T. Ribeiro et al., "Why should I trust you?: Explaining the predictions of any classifier", in *KDD 2016*, 2016.
11. T. Aldhyani, Z. A. T. Ahmed, B. M. Alsharbi et al., "Diagnosis and detection of bone fracture in radiographic images using deep learning approaches", *Frontiers in Medicine*, vol. 11, p. 1506686, 2025.
12. Z. Su, A. Adam, M. F. Nasrudin, M. Ayob and G. Punganan, "Skeletal fracture detection with deep learning: A comprehensive review", *Diagnostics*, vol. 13, p. 3245, 2023.
13. F. Hardalaç, F. Uysal and O. Peker, "Fracture detection in wrist X-ray images using deep learning-based object detection models", *Sensors*, vol. 22, no. 3, p. 1285, 2022.
14. T. Meena and S. Roy, "Bone fracture detection using deep supervised learning from radiological images: A paradigm shift", *Diagnostics*, vol. 12, no. 10, p. 2420, 2022.
15. S. Parvin and A. Rahman, "A real-time human bone fracture detection and classification from multi-modal images using deep learning technique", *Applied Intelligence*, vol. 54, no. 19, pp. 9269–9285, 2024.
16. F. Uysal, F. Hardalaç, O. Peker, T. Tolunay and N. Tokgöz, "Classification of shoulder X-ray images with deep learning ensemble models", *Applied Sciences*, vol. 11, no. 6, p. 2723, 2021.
17. B. Guan, J. Yao and G. Zhang, "An enhanced vision transformer with scale-aware and spatial-aware attention for thighbone fracture detection", *Neural Computing and Applications*, vol. 36, no. 19, pp. 11425–11438, 2024.
18. I. Abedeen, M. A. Rahman, F. Z. Protyasha, T. Ahmed, T. M. Chowdhury and S. Shatabda, "FracAtlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs", *Scientific Data*, vol. 10, no. 1, p. 521, 2023.