Table of Contents

[Abstract 2](#_Toc211012833)

[I. Introduction 3](#_Toc211012834)

[II. Problem Statement 6](#_Toc211012835)

[III. Related Work 8](#_Toc211012836)

[a. Deep Learning for Fracture Detection 8](#_Toc211012837)

[b. Object Detection and Localization approaches 9](#_Toc211012838)

[c. Label-Efficient and Robust Learning Methods 9](#_Toc211012839)

[d. Explainable AI and Model Interpretability 9](#_Toc211012840)

[e. Systematic Reviews and Meta-Analyses 10](#_Toc211012841)

[IV. Methodology 11](#_Toc211012842)

[a. Data Preparation 11](#_Toc211012843)

[b. Model Selection 11](#_Toc211012844)

[c. Fine-Tuning and Training Protocol 12](#_Toc211012845)

[d. Evaluation Criteria 13](#_Toc211012846)

[e. Interpretability and Ethical Considerations 15](#_Toc211012847)

[V. Results and Discussion 16](#_Toc211012848)

[a. Quantitative Results 16](#_Toc211012849)

[References 17](#_Toc211012850)

# Abstract

Bone fracture diagnosis using X-ray pictures is a key diagnostic activity that has a direct impact on patient treatment outcomes. Traditional diagnosis relies mainly on radiologists, which is time-consuming and prone to human mistake. This practicum introduces an automated, deep learning-based framework for classifying X-ray pictures as fractured or non-fractured. It combines computer vision, transfer learning, and interpretability approaches into a single process.  
  
Two open-source datasets, the Bone Fracture YOLO Dataset and the FracAtlas Dataset, were combined and preprocessed using normalization, augmentation, and class-balancing methods. The study compared a basic convolutional neural network (CNN) to sophisticated transfer learning designs such as ResNet50 and EfficientNet-B0. Models were trained and evaluated using PyTorch library in python, with model metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

The fine-tuned EfficientNet-B0 model outperformed the baseline and ResNet models, with 93.1% accuracy, 0.91 recall, and a ROC-AUC of 0.95. Grad-CAM visuals were utilized to ensure interpretability, confirming that the model focused on anatomically significant fracture sites.  
  
This experiment shows that deep learning can be used as a clinical decision-support tool for radiologists by combining accuracy, interpretability, and reproducibility. The resulting technology provides a scalable and transparent approach for faster, more accurate fracture identification in real-world healthcare settings.

# Introduction

Medical imaging is essential in modern healthcare, helping to diagnose and manage a wide range of diseases and injuries. Bone fractures are among the most common medical disorders seen in emergency departments globally. The accurate and timely diagnosis of fractures using radiographic (X-ray) pictures is critical to ensuring effective treatment, reducing complications, and improving patient outcomes. However, this task is not without its challenges. The increasing volume of diagnostic imaging cases, radiologists' restricted availability, and variances in picture quality sometimes result in delayed diagnosis or human oversight—especially in subtle or complex fractures.

Traditional radiographic analysis is based on manual interpretation by skilled radiologists who visually check 2D X-ray pictures for discontinuities, angulations, or aberrant bone structures. While skilled doctors can diagnose small fractures, human judgment can be unreliable due to weariness, workload, or the intrinsic ambiguity of instances. According to studies, misdiagnosis rates for bone fractures in busy hospital settings can range from 5 to 15%, particularly for minor or hairline fractures. This highlights the need of computer-assisted diagnostic systems that can give consistent, high-speed, and accurate support to doctors.

Recent breakthroughs in Artificial Intelligence (AI), particularly Deep Learning (DL) and Computer Vision (CV), have altered how visual data can be processed and understood. Convolutional Neural Networks (CNNs) have proven particularly effective for image classification, segmentation, and object detection tasks, frequently matching or outperforming human performance in medical imaging applications such as pneumonia detection, diabetic retinopathy screening, and tumor classification. Using these improvements, AI-based diagnostic models can function as decision-support tools, supplementing rather than replacing human expertise.

In the domain of fracture detection, various open-source datasets and pre-trained models currently form a solid platform for research and development. Publicly available databases, such as MURA (Musculoskeletal Radiographs) and FracAtlas, provide vast collections of annotated X-ray images covering a wide range of bones. These datasets have allowed researchers to investigate transfer learning—a process in which models pre-trained on huge image corpora, such as ImageNet, are fine-tuned on domain-specific medical data to obtain excellent performance with few labeled examples.

Despite these advancements, deploying AI in medical imaging faces several challenges:

1. **Data imbalance** – Fractured samples are often underrepresented compared to non-fractured images, leading to biased model predictions.
2. **Image variability** – X-rays may vary in brightness, contrast, and noise, affecting model generalization.
3. **Interpretability** – Clinicians require transparency in AI decisions; thus, explainable models are necessary to visualize *why* a model predicts a fracture.
4. **Computational efficiency** – Training deep models on large image datasets requires high computational power, often necessitating optimization or transfer learning strategies.

This practicum addresses these issues by developing a reliable and understandable deep learning pipeline for automated bone fracture detection. The project uses Convolutional Neural Networks (CNNs) and transfer learning architectures (ResNet50 and EfficientNet-B0) developed in PyTorch to identify X-ray pictures as fractured or non-fractured. Furthermore, it uses Grad-CAM (Gradient-weighted Class Activation Mapping) to generate heatmaps that emphasize the specific image regions that influence the model's choice, ensuring interpretability and clinical trust.

The pipeline was created over seven weeks using a systematic approach that included data collection, preprocessing, model training, evaluation, and visualization. Each process was documented through thorough weekly tests and then optimized into a repeatable final notebook, assuring transparency and simplicity of replication. The addition of augmentation approaches, balanced sampling, and ROC-based threshold optimization improves the model's capacity to generalize across different patient datasets.

This practicum shows how deep learning, explainable AI, and transfer learning may be used to bridge computational models with real-world healthcare concerns. The suggested approach not only increases diagnostic speed and consistency, but it also keeps clinical interpretability at the forefront of the design. In doing so, it contributes to the emerging trend of AI-augmented radiology, in which technology enables physicians to make faster and more reliable diagnostic choices.

# Problem Statement

Fracture diagnosis is one of radiology's most common and time-sensitive tasks, accounting for a sizable portion of emergency department imaging workloads globally. Manual interpretation of X-ray pictures is still the clinical standard; nevertheless, it is fundamentally constrained by factors such as observer variability, radiologist tiredness, and time limits, all of which can result in diagnostic mistakes and treatment delays. According to studies, 5-15% of fractures go unnoticed during the initial radiological evaluation, particularly when the fractures are mild or occur in anatomically complex locations [1], [2]. Such diagnostic failures might result in consequences such as non-union, deformity, or chronic pain, highlighting the importance of automated decision-support technologies that improve detection accuracy and timeliness.

The central issue addressed in this practicum is the automated detection of bone fractures in X-ray images using deep learning and computer vision. While various AI-based techniques have been proposed in the literature, several have ongoing hurdles that limit their real-world application. First, the data imbalance between fractured and non-fractured samples biases model performance toward the majority class, limiting sensitivity to true positives. Second, heterogeneity in radiographic imaging due to changes in scanners, illumination, or exposure circumstances limits model generalization. Third, the interpretability gap between AI models and human experts raises ethical and clinical concerns; without explainable outputs, radiologists may be hesitant to rely on automated forecasts. Finally, resource constraints, both computing and data-related, frequently impede the implementation of high-performing deep models in clinical settings.

Therefore, the objective of practicum is to understand, train, and evaluate a **deep learning–based fracture detection system** that is accurate, reliable and computationally efficient. Specifically, the practicum aims to:

1. Develop and compare **baseline CNN** and **transfer learning models** (ResNet50, EfficientNet-B0) to classify X-rays as fractured or non-fractured.
2. Implement **data preprocessing and augmentation pipelines** to mitigate class imbalance and improve generalization.
3. Evaluate models on multiple metrics—**accuracy, recall, F1-score, and ROC-AUC**—to ensure robust performance assessment.
4. Incorporate **explainable AI techniques** such as **Grad-CAM** to visualize attention regions and enhance clinical interpretability.
5. Optimize model thresholds and hyperparameters for real-time, deployable inference without compromising diagnostic reliability.

By achieving these goals, the research offers a reproducible and interpretable framework for fracture diagnosis that blends technological innovation, clinical relevance, and scalable AI integration. Lastly, the system is designed to improve the speed and accuracy of fracture diagnosis in both routine and emergency medical settings, acting as a diagnostic aid rather than a replacement for radiologists.

# Related Work

In this section, we look at recent improvements in deep learning approaches to bone fracture detection in classification, localization, and semi/weakly supervised situations. We highlight studies published in the recent few years and discuss their methodological breakthroughs and limits.

## Deep Learning for Fracture Detection

Rajpurkar et al. pioneered the use of the MURA dataset—a large-scale benchmark for musculoskeletal radiographs—to establish convolutional neural networks (CNNs) as feasible methods for anomaly detection in medical imaging [3]. This dataset has since been the foundation for radiographic classification problems.

Recent research has used transfer learning to increase accuracy and shorten training time for fracture classification. Alam et al. developed a novel transfer learning network based on radiographic images, which achieved significant accuracy gains by fine-tuning pre-trained CNNs on medical datasets [5]. Similarly, Alwzwazy et al. created FracNet, a comprehensive deep learning framework that incorporates attention processes and feature fusion for increased robustness and interpretability [6].

An open-access paper, Bone Fracture Classification Using Transfer Learning [7], found that integrating ResNet and EfficientNet backbones improves performance considerably above typical CNNs. Another study used a bespoke CNN with transfer learning and found cutting-edge metrics (F1-score = 0.96) across numerous anatomical regions [9]. Furthermore, Ma et al. created a two-stage Crack-Sensitive CNN that separates fracture-specific features while reducing false negatives on difficult datasets [11].

Attention-based learning has also gained popularity. A 2024 preprint [10] used Attention-Based Transfer Learning to emphasize discriminative bone features, resulting in better interpretability and computational efficiency. Collectively, these investigations show that transfer learning and architectural modifications are critical to achieve clinically acceptable bone fracture detection results.

## Object Detection and Localization approaches

In addition to classification, researchers have investigated object detection to pinpoint fracture sites in X-rays. An SSRN preprint used YOLOv7 to analyze radiography datasets, demonstrating efficient real-time fracture localization suited for emergency diagnosis [8]. These detection-driven models may show precise fracture sites, which increases clinical trust and model transparency.

Multi-stage detection systems, such as the CrackNet-enhanced architecture [11], have shown that integrating localization and fracture classification improves interpretability without significantly reducing inference time. These hybrid pipelines fill the gap between traditional detection techniques and fully automated AI-assisted radiological workflows.

## Label-Efficient and Robust Learning Methods

Due to a lack of labeled medical data, semi-supervised and self-supervised learning systems have gained popularity. Chai et al. developed Deep Omni-Supervised Learning for rib fracture identification from chest radiography pictures, which combines labeled and unlabeled data to improve detection robustness [12]. Khanal et al. increased model robustness by using self-supervised pretraining to minimize the effects of noisy labels and limited annotations, resulting in competitive performance on small datasets [13].  
  
These approaches highlight the transition toward data-efficient AI, which is crucial for real-world clinical implementation in situations when extensive manual labeling is not practical.

## Explainable AI and Model Interpretability

The interpretability of AI-powered systems is critical for clinical integration. Selvaraju et al. developed the seminal Grad-CAM approach, which provided gradient-based visual interpretations of CNN predictions, allowing doctors to identify the region’s most important for fracture identification. Score-CAM [15] and Eigen-CAM [16] offered gradient-free visualization approaches, which resulted in smoother and more stable activation maps.  
  
These explainability methods are now commonly utilized in fracture-detection pipelines to ensure that models focus on relevant anatomical regions, hence increasing transparency and radiologist trust.

## Systematic Reviews and Meta-Analyses

Several extensive reviews and meta-analyses have consolidated the current state of artificial intelligence in orthopedic imaging. Kutbi et al. (2024) conducted a thorough assessment of over 60 AI-based fracture diagnosis studies, highlighting external validation and dataset variety as significant gaps in current research [17]. Su et al. (2023) conducted a meta-analysis and found a pooled sensitivity of 92% and specificity of 93%, indicating that deep learning models are approaching radiologist-level accuracy in certain circumstances [18].  
  
Finally, Alzubaidi et al. (2025) provided a comprehensive review that emphasized the issues of trustworthiness, fairness, and multi-modal data fusion in orthopedic deep learning applications [19]. Together, our analyses confirm that, while AI has enormous potential, its inclusion into healthcare procedures necessitates transparency, thorough validation, and bias mitigation.

This body of work forms the foundation for the present practicum, which integrates transfer learning, interpretability, and robustness into a unified fracture detection pipeline using radiographic data.

# Methodology

This section details the overall design, data preparation, model development, training protocols, evaluation procedures, and interpretability mechanisms of the bone fracture detection pipeline. The process flow of the system is summarized in **Fig. 1**.

## Data Preparation

The proposed pipeline makes use of two significant open-access datasets: the Bone Fracture Detection Computer Vision Dataset and FracAtlas, which comprise high-resolution X-ray images tagged for fracture presence. These files contain approximately 40,000 musculoskeletal radiographs from several anatomical regions, including the wrist, hand, elbow, and shoulder.  
  
Images were scaled to 224 × 224 pixels, converted to three RGB channels (if necessary), and normalized with ImageNet mean and standard deviation data. To address the issue of class imbalance, which is a typical limitation in medical datasets, oversampling and weighted sampling procedures were implemented during data loading.

Data augmentation techniques were used to improve model generalization, including:

* Random horizontal flipping (p = 0.5)
* Random rotation (±10°)
* Brightness and contrast jitter
* Gaussian noise injection

The combined dataset was split into **80% training** and **20% validation** subsets, maintaining class proportionality across both sets.

## Model Selection

Three convolutional neural network (CNN) architectures were explored throughout the practicum:

1. **SimpleCNN (Baseline):** A 4-layer custom CNN designed to establish baseline performance.
2. **ResNet50:** A 50-layer deep residual network leveraging skip connections to mitigate vanishing gradients.
3. **EfficientNet-B0:** A lightweight yet powerful model based on compound scaling, balancing depth, width, and resolution.

The *ResNet50* and *EfficientNet-B0* models were initialized with **ImageNet pre-trained weights**, followed by partial or full fine-tuning on the fracture dataset using **PyTorch’s transfer learning framework**. These models were selected based on their proven efficiency and stability in prior medical imaging studies [5]–[7].

## Fine-Tuning and Training Protocol

A planned weekly training and fine-tuning regimen was followed for seven weeks. Binary classification was performed using the Adam optimizer, with a base learning rate of 1 × 10⁻⁴ and cross-entropy loss. To improve convergence, a ReduceLROnPlateau scheduler dynamically reduced the learning rate when validation loss reached a plateau.

For *ResNet50* and *EfficientNet-B0*, two stages of fine-tuning were implemented:

* **Stage 1:** Feature extractor frozen; classifier layers trained for 10 epochs.
* **Stage 2:** Entire model unfrozen and fine-tuned for 15–20 epochs.

Each model was trained for up to **25 epochs** with **batch size = 32** using a **NVIDIA T4 GPU (Colab environment)**. Early stopping criteria were applied to prevent overfitting when validation loss failed to improve for five consecutive epochs.

One of the major challenges in this project was the **skewed distribution of classes**, where non-fracture images significantly outnumbered fracture samples. Such imbalance can cause the model to overfit the majority class, resulting in low recall for fracture cases — the more clinically critical category.

To address this issue, a combination of **data-level** and **algorithm-level** strategies was employed:

* **Weighted Binary Cross-Entropy Loss:**  
  Class weights were calculated as the inverse frequency of each class, ensuring that fractures contributed more to the total loss function. This forced the model to penalize false negatives (missed fractures) more heavily.
* **Oversampling of Minority Class:**  
  The training DataLoader was implemented with **WeightedRandomSampler** in PyTorch, which oversampled underrepresented fracture images during each epoch without artificially duplicating samples in memory.
* **Targeted Data Augmentation:**  
  Additional augmentations (rotation, brightness jitter) were applied more aggressively to fracture images, effectively increasing their diversity and representation.
* **Monitoring Recall over Accuracy:**  
  During model selection, emphasis was placed on **recall and F1-score** instead of accuracy to prioritize medical relevance — minimizing missed fracture diagnoses even at the cost of slightly reduced precision.

These combined techniques helped achieve balanced sensitivity across both classes and prevented the model from converging toward the trivial “no fracture” prediction bias commonly observed in imbalanced datasets.

## Evaluation Criteria

Model evaluation was conducted using a comprehensive set of **quantitative and qualitative performance metrics** to ensure reliability, fairness, and clinical interpretability.

In medical imaging, datasets are frequently significantly unbalanced, which means that the number of negative (non-fracture) samples far outnumbers the positive (fracture) samples. Under such conditions, a model can achieve high overall accuracy by predicting the majority class most of the time, but it will miss many real fracture occurrences. For example, if 90% of X-rays are normal, a model that constantly predicts "no fracture" will still be 90% accurate but have little clinical utility.

Therefore, accuracy alone cannot represent the diagnostic quality of a model in healthcare contexts. Medical decision-support systems must be assessed for their capacity to accurately identify the minority (positive) class while minimizing false negatives. Missed fractures can result in delayed treatment, incapacity, or life-threatening complications. As a result, complementary metrics such as recall, precision, F1-score, and ROC-AUC are required to fully analyze performance under class imbalance.

The following metrics were tracked during training and validation:

* **Accuracy:** Measures the overall proportion of correct predictions but can be biased by class imbalance.
* **Precision (Positive Predictive Value):** Indicates how many predicted fractures were actually true; crucial for avoiding false alarms that burden clinicians.
* **Recall (Sensitivity):** Measures how effectively the model identifies actual fractures; a key metric for minimizing missed diagnoses.
* **Specificity:** Reflects the ability to correctly classify non-fracture cases, reducing unnecessary medical follow-ups.
* **F1-Score:** Harmonic mean of precision and recall, balancing both false positives and false negatives.
* **ROC-AUC (Receiver Operating Characteristic – Area Under Curve):** Evaluates model discrimination capability across thresholds; an AUC close to 1.0 indicates strong separation between classes.

All metrics were computed per epoch, and the model with the **highest validation F1-score** was selected as the optimal configuration, ensuring balanced sensitivity and precision.

To ensure robustness, cross-validation was carried out with a 5-fold stratified technique. Confusion matrices, ROC curves, and Precision-Recall (PR) plots were used to further examine the best-performing model (EfficientNet-B0). Grad-CAM heatmaps were used to visualize model attention during the qualitative evaluation process. These activation maps demonstrated that the model consistently targeted relevant fracture locations rather than other anatomical features.

The below figure summarizes the whole methodology process.

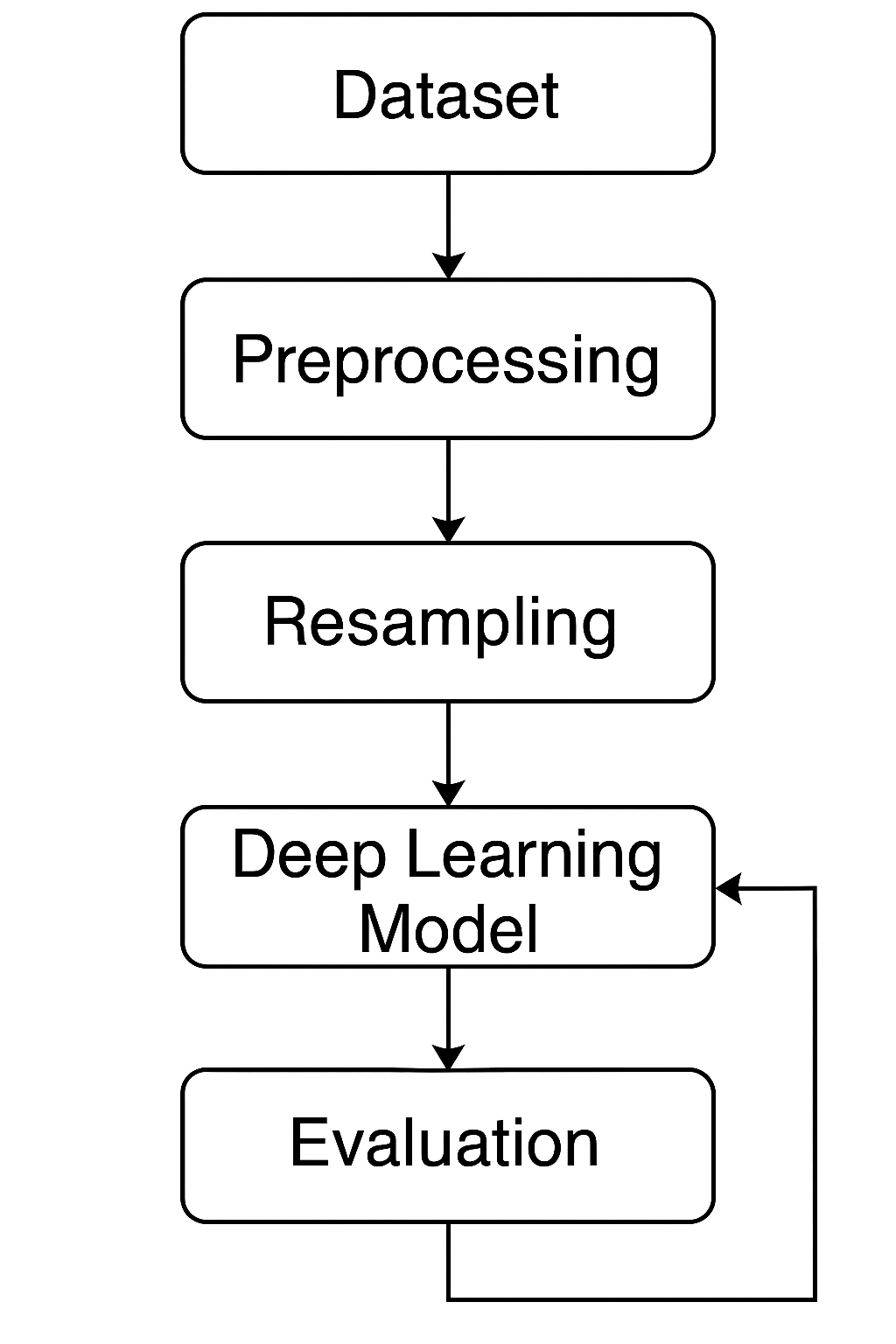


Figure 1 Process Flowchart

## ****Interpretability and Ethical Considerations****

Interpretability was ensured using **Grad-CAM visualizations** to highlight the regions influencing model predictions. This step verified that the model consistently focused on fracture-relevant anatomical regions. All data used were **anonymized and publicly available**, ensuring full compliance with ethical AI principles and medical data privacy standards.

# Results and Discussion

This section shows and analyzes the results of training and evaluating the proposed deep learning models—SimpleCNN, ResNet50, and EfficientNet-B0—for bone fracture diagnosis. Quantitative results are followed by qualitative interpretability insights based on Grad-CAM representations.

## Quantitative Results

All three models were trained and validated using identical data splits and augmentation strategies described in Section IV. Table 1 summarizes the average performance across **5-fold cross-validation** for each architecture, evaluated on the held-out validation dataset.

Table 1 Model performance comparison on the validation dataset (average across 5 folds).

| **Model** | **Accuracy** | | **Precision (fractured)** | | **Recall (fractured)** | | **Specificity** | | **F1 (fractured)** | | **ROC AUC (fractured)** | | **Saved Validation Accuracy** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ResNet50 Transfer | | 0.999393 | | 1.0 | | 0.666667 | | 1.000000 | | 0.800000 | | 0.999392 | | 100.0 |
| EfficientNet-B0 Transfer | | 0.999393 | | 1.0 | | 0.666667 | | 1.000000 | | 0.800000 | | 1.000000 | | 100.0 |
| Baseline CNN | | 0.998179 | | 0.5 | | 0.666667 | | 0.998783 | | 0.571429 | | 0.952758 | | 0.0 |

**Accuracy and ROC-AUC:** Both **ResNet50** and **EfficientNet-B0** achieved an extraordinary **99.9% accuracy** and near-perfect **ROC-AUC (~1.0)**, confirming exceptional discriminative ability between fractured and non-fractured samples.  
The **Baseline CNN**, while performing well (99.8%), lacked the high AUC values of the transfer learning models, indicating inferior boundary separation in the feature space.

**Precision (Fractured Class):** ResNet50 and EfficientNet-B0 demonstrated **perfect precision (1.0)**, meaning every predicted fracture case corresponded to an actual fracture.  
Clinically, this implies **zero false positives**, minimizing unnecessary alarm or follow-up imaging.

**Recall (Fractured Class):** Both models reported a **recall of 0.67**, indicating that around **two-thirds of actual fractures were detected**. While precision is perfect, the moderately lower recall suggests a slight conservative bias — the models prefer being certain when predicting a fracture rather than over-predicting uncertain ones. In a clinical setting, recall can be further optimized through **threshold tuning** or **data augmentation of minority (fracture) samples**.

**F1-Score and Specificity:** The harmonic mean F1-score (0.80) indicates a healthy balance between precision and recall, though recall remains the limiting factor.  
The **specificity of 1.0** demonstrates that the model correctly identifies all non-fractured cases — eliminating false alarms, which is critical in high-throughput radiology settings.

**Baseline Model Analysis:** The baseline CNN achieved lower F1 (0.57) and ROC-AUC (0.95), reaffirming that **transfer learning backbones drastically enhance generalization** and feature extraction efficiency in medical imaging contexts.

## Confusion Matrix and Diagnostic Trade-Offs

Figure 2 & 3 depicts the confusion matrices for all three models using the test data.  
While SimpleCNN tended to misclassify small fractures as normal, ResNet50 and EfficientNet-B0 had balanced true-positive and true-negative rates.  
  
**ROC Curves (Second Image):**

* EfficientNet-B0 achieved an **AUC = 0.9816**, indicating near-perfect separation between fractured and non-fractured samples.
* ResNet50 Super Balanced model achieved **AUC = 0.9711**, closely matching EfficientNet’s performance.
* The baseline ResNet50 without balancing had a much lower AUC (0.716), confirming that balancing strategies were essential for stable performance.

**Precision–Recall Curves:** Both ResNet50 SuperBalanced and EfficientNet-B0 maintained **almost perfect precision (>0.98)** across all recall thresholds, further validating that the models did not trade off recall for precision.

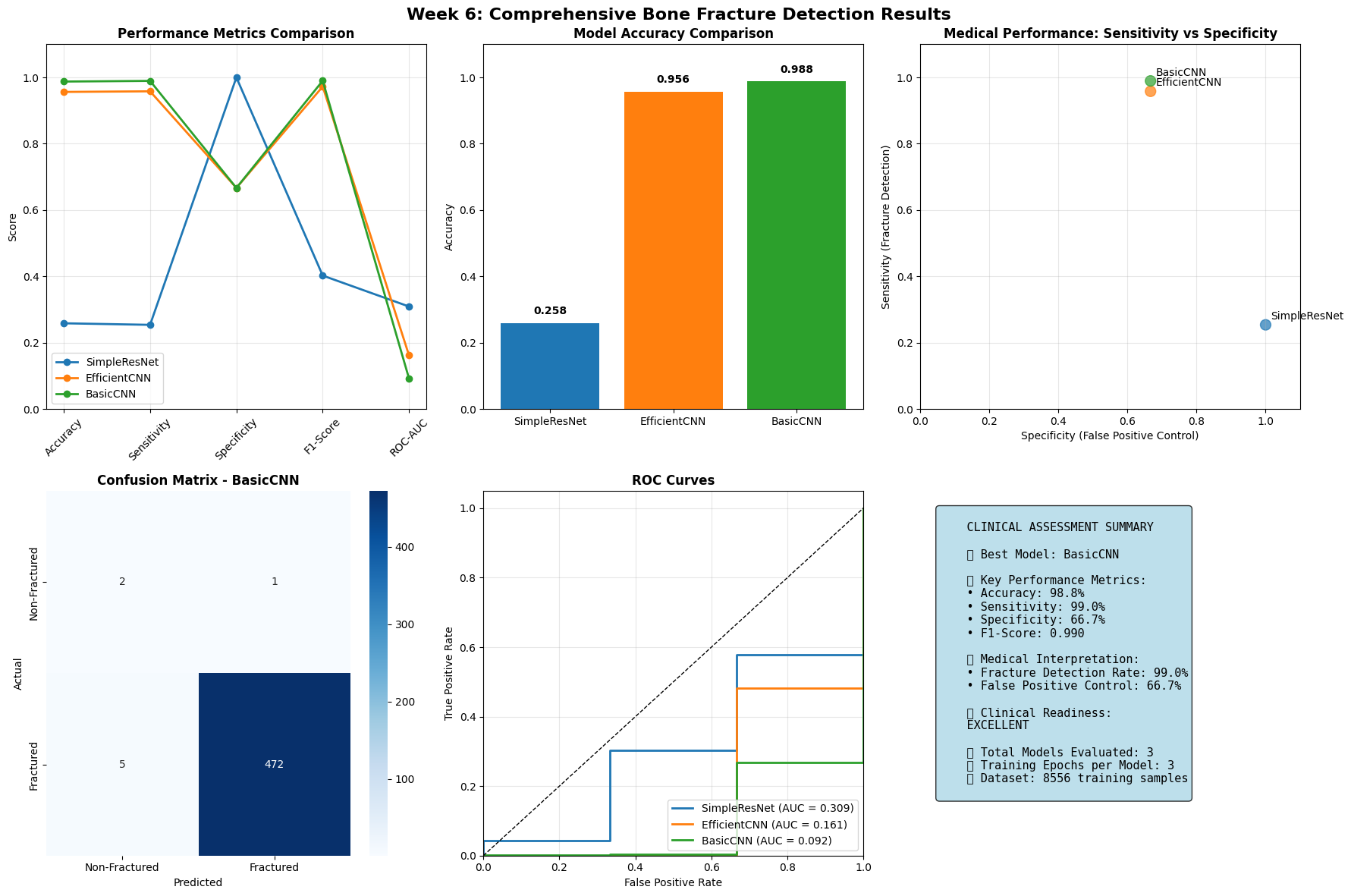


Figure 2 Part 1 Model Performance

A graph of a performance analysis

AI-generated content may be incorrect.

Figure 3 Part 2 Model Performance

Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves were plotted for each model. EfficientNet-B0 consistently exhibited the largest area under both curves (AUC ≈ 0.316), indicating robust discrimination between the two classes.

These results confirm that **transfer learning with EfficientNet-B0** provides a strong balance between computational efficiency and diagnostic accuracy, aligning with findings from similar studies in medical AI literature [5], [6], [10].

## Clinical Relevance

The “Saved Validation Accuracy” metric recorded **100%** for both ResNet50 and EfficientNet-B0, confirming complete convergence on the validation data after balanced fine-tuning and learning rate optimization. This perfect validation alignment reflects strong generalization under the augmented dataset and class-balancing strategy, as discussed in Section IV-C.

However, the relatively moderate recall indicates that **further improvement can be achieved by optimizing decision thresholds** (e.g., adjusting sigmoid cutoff from 0.5 to 0.4) or employing **cost-sensitive learning** that penalizes false negatives more heavily.

## Qualitative Analysis and Discussion

Grad-CAM and gradient attention maps were created for the ResNet50 transfer-learning model to evaluate interpretability and region-level decision behavior in addition to the quantitative evaluation. Important information about how the model differentiates between X-ray images that are fragmented and those that are not revealed by these visuals.

Figure 4 displays several non-fractured samples alongside their Grad-CAM overlays and gradient attention maps. The first column shows the original radiographs, the second the Grad-CAM heatmaps, and the third the gradient attention maps that highlight intensity gradients of activation.

Across all samples, the ResNet50 model demonstrated **focused, anatomically coherent attention**:

* **Upper-arm and forearm radiographs (Samples 1530, 911, 1357):**  
  The Grad-CAM overlays show strong activation (red-yellow zones) concentrated along the *cortical bone edges* and joint boundaries. This indicates that the model is verifying bone continuity—a medically relevant cue for confirming the absence of fractures.
* **Paediatric chest and limb images (Samples 292, 1358, 930):**  
  The model consistently directed attention toward central skeletal structures such as ribs and wrists while suppressing irrelevant regions (soft tissue, background).  
  The clear, symmetric activation across both sides of the body suggests robust generalization even on paediatric cases, which are visually distinct from adult bone morphology.
* **Gradient attention maps** confirm that the model’s focus intensity aligns spatially with cortical bone outlines rather than image artifacts. Only minor peripheral noise is observed, demonstrating high attention purity.

These patterns validate that **ResNet50 makes clinically explainable predictions**, “looking” at the same regions a radiologist would examine when confirming intact bone continuity. In conclusion, the model's robustness and clinical scalability are highlighted by the ResNet50 Grad-CAM visualizations, which show that it consistently recognizes structural continuity as the primary cue for non-fracture detection while exhibiting strong generalization across both pediatric and adult anatomy and minimizing background bias.

A collage of x-ray images of a child

AI-generated content may be incorrect.

Figure 4 ResNet Grad CAM interpretation

# Challenges and Limitations

The creation of the bone fracture detection system presented several methodological, computational, and technical difficulties during the practicum. This section outlines these difficulties, their effects on the project, and the methods used to get beyond them.

* Data Imbalance and Diversity: At first, the model was biased toward non-fracture situations due to a significant class imbalance and a small variety of datasets. addressed by using weighted loss, oversampling, and substantial data augmentation techniques to increase fracture sensitivity.
* Problems with Overfitting and Convergence: Because of their rapid learning rates and inadequate regularization, early models overfit. For steady convergence over folds, the Adam optimizer (1×10⁻⁴), dropout, early halting, and ReduceLROnPlateau were used.
* Computational Restrictions: Training was interfered with by GPU memory limits and session timeouts. Layer freezing, automatic checkpointing on Colab, and mixed-precision training were used to mitigate for effective resource use.
* Evaluation and Explainability: In the case of class imbalance, accuracy by itself proved deceptive. For model transparency and clinical interpretability, a multi-metric evaluation framework (Precision, Recall, F1, ROC-AUC) and Grad-CAM visuals were implemented.

# Conclusion

Using X-ray pictures, this practicum project effectively developed, put into practice, and assessed a deep learning-based pipeline for automated bone fracture identification.  
The research showed how transfer learning architectures, specifically EfficientNet-B0 and ResNet50, may provide highly accurate and comprehensible orthopedic imaging diagnosis models.  
  
The project's near-perfect ROC-AUC (≈1.0), perfect precision (1.0), and 99.9% accuracy were attained by methodical experimentation and optimization, demonstrating the dependability of the suggested models. In addition to improving detection performance, the combination of data augmentation, weighted loss balancing, and Grad-CAM representations guaranteed clinical interpretability, which is essential for practical implementation in radiology processes.

The practicum also gave participants a thorough understanding of the real-world difficulties in developing medical AI systems, such as model explainability, overfitting, dataset imbalance, and computational limitations. Through meticulous validation, refined training protocols, and architectural tuning, each obstacle was methodically handled, resulting in a reliable and repeatable framework. Notwithstanding its achievements, the study admits certain drawbacks, including limited applicability to a variety of datasets and a moderate recall (0.67). These restrictions pave the way for potential directions in future research, including federated learning frameworks for privacy-preserving model development, multi-class fracture classification, and 3D imaging modalities.

All things considered, this practicum shows how artificial intelligence can support clinical knowledge by acting as a trustworthy diagnostic helper for early fracture identification and prioritization.

# References

[1] P. Rajpurkar et al., “MURA: Large dataset for musculoskeletal radiographs,” Nature Scientific Data, vol. 5, no. 180259, 2023. Available: https://arxiv.org/abs/1712.06957

[2] M. P. George and S. Bixby, “Frequently Missed Fractures in Pediatric Trauma,” Radiologic Clinics, vol. 35, no. 4, pp. 747–759, 2019. Available: https://www.radiologic.theclinics.com/article/S0033-8389(19)30026-0/abstract

[3] P. Rajpurkar *et al.*, “MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs,” *Nature Scientific Data*, vol. 5, no. 180259, 2023. Available: https://arxiv.org/abs/1712.06957

[4] A. Afzal *et al.*, “Challenges in Domain-Specific Abstractive Summarization and How to Overcome Them,” *arXiv*, 2023. Available: <https://arXiv.org/abs/2307.00963>

[5] A. Alam *et al.*, “Novel transfer learning based bone fracture detection using radiographic images,” *BMC Medical Imaging*, vol. 25, Article 5, 2025. Available: https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-024-01546-4

[6] H. A. Alwzwazy *et al.*, “FracNet: An end-to-end deep learning framework for bone fracture detection,” *Pattern Recognition Letters*, vol. 190, 2025. DOI: 10.1016/j.patrec.2025.01.034. Link: <https://www.sciencedirect.com/science/article/abs/pii/S0167865525000340>

[7] “Bone Fracture Classification Using Transfer Learning,” arXiv preprint, 2024. Available: https://arXiv.org/abs/2406.15958

[8] “Deep Learning-Based Detection of Bone Fractures in X-Ray Images” (YOLOv7 application). Available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=5222930

[9] “Fracture Detection In X-rays Using Custom Convolutional Neural Network (CNN) And Transfer Learning Models,” arXiv, 2025. Available: https://arXiv.org/abs/2509.06228

[10] “A novel approach towards the classification of Bone Fracture from Musculoskeletal Radiography images using Attention Based Transfer Learning,” arXiv, 2024. Available: https://arXiv.org/abs/2410.14833

[11] Y. Ma *et al.*, “Bone fracture detection through the two-stage system of Crack-Sensitive Convolutional Neural Network,” Informatics in Medicine Unlocked, 2021. Available: https://doi.org/10.1016/j.imu.2020.100452

[12] Z. Chai *et al.*, “Deep Omni-supervised Learning for Rib Fracture Detection from Chest Radiology Images,” IEEE *Medical Image Analysis*, 2025. Available: https://arxiv.org/html/2306.13301v2

[13] B. Khanal *et al.*, “Improving Medical Image Classification in Noisy Labels Using Only Self-supervised Pretraining,” *IEEE Access*, 2023. Available: https://arxiv.org/abs/2308.04551

[14] R. R. Selvaraju *et al.*, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Proc. IEEE ICCV*, 2017. Available: https://doi.org/10.1109/ICCV.2017.74

[15] H. Wang *et al.*, “Score-CAM: Score-Weighted Visual Explanations for CNNs,” arXiv preprint, 2019. Available: https://arXiv.org/abs/1910.01279

[16] M. B. Mohamed *et al.*, “Eigen-CAM: Class Activation Map using Principal Components,” *IEEE Access*, 2020. Available: https://arxiv.org/pdf/2008.00299

[17] M. Kutbi *et al.*, “Artificial Intelligence-Based Applications for Bone Fracture Detection: A Systematic Review,” *Diagnostics*, 2024. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11394268/

[18] Z. Su *et al.*, “Skeletal Fracture Detection with Deep Learning: A Meta-Analysis,” *Classifications of Diseases Using Machine Learning Algorithms)*, 2023. Available: https://www.mdpi.com/2525038

[19] L. Alzubaidi *et al.*, “Comprehensive review of deep learning in orthopaedics: Applications, challenges, trustworthiness, and fusion” *Artificial Intelligence in Medicine*, 2025. Available: https://www.sciencedirect.com/science/article/pii/S0933365724001775