Deep Learning for Bone Fracture Classification from Medical Images

Abstract

Bone fractures are a common and serious health issue, and deep learning techniques have shown promise in assisting radiologists with fracture detection from radiological images. This project leverages a dataset of X-ray images of bone fractures to train image classification models for automated fracture diagnosis. Convolutional Neural Networks (CNNs) are implemented for this task, exploring advanced architectures such as EfficientNet and Inception. For example, EfficientNet is known for an excellent balance of accuracy and computational efficiency, while Inception has been widely used to classify abnormalities in medical imaging. The trained models aim to detect the presence of fractures and categorize them by type, providing a computer-aided tool for precise bone fracture identification from X-ray imagery.

1. Introduction

Projectional radiography produces 2-dimensional images through exposing objects to electromagnetic radiation. X-rays are a type of electromagnetic radiation, commonly used in medical imaging to assist medical diagnoses -such as fractures- by producing a 2-D grayscale image referred to as a Roentgenogram or more commonly, an X-ray image.

Other means to detect fractures such as Ultrasound, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) scans exist, and although they may produce higher quality or 3-D, coloured images, they have higher costs and risks associated with them.

Part of the Computer Assisted Diagnoses (CAD) field, this project considers the role of augmenting roentgenograms to mimic features present in other medical imaging methods to improve fracture classification. In particular, we consider whether adding Volume rendering, Colour and Edge detection to grayscale X-ray images allows deep learning models to better distinguish between the classes of fractures.

Although other learning methods and models were considered, those presented in this report include fine-tuning pre-trained models (such as EfficientNet and Inception) and our own CNN-based architecture to perform the classification task. These decisions were motivated by the context and performance of the models (more in **Section 6**).

2. Background: CAD in Medical Imaging

Computer assisted diagnosis began in the late 1960s when researchers first applied basic image processing algorithms to radiographs. As of 2025, CAD has evolved into a decision support platform that leverages machine learning methods, including deep convolutional neural networks, transformer architectures and large scale data integration.

Widespread adoption by healthcare institutions has led to improved diagnostic accuracy shortened reporting times and better patient outcomes while ongoing research explores federated learning and explainable artificial intelligence to ensure equitable and transparent deployment.

Modern systems assist in delivering real time risk assessment, anomaly detection and treatment recommendations, with key applications in medical imaging including lung nodule identification in chest CT, diabetic retinopathy screening, and automated fracture detection in musculoskeletal imaging.

In England, 47.2 million musculoskeletal image test were produced by the NHS for patients, of which the most common image tests performed were X-ray images, making up almost half of the share at 22.6 million (NHS England, 2024). X-ray images are most commonly produced for fracture detection, which is the process of identifying whether a partial or complete break in a bone is present.

Our subsequent analysis is focused on classifying fractures based on the pattern or type of break, with Table 1 briefly describing the different classes of fractures considered.

Despite their widespread use, radiologists typically perform worse at identifying fractures from X-ray images compared to other tests. Research finds this ability could be upto 6 percentage points worse when using X-ray images instead of CT images (Etli et al., 2020), and 23% worse than Ultrasound images under certain circumstances (Khan et al., 2023). Nevertheless, the utility and prevalence of X-ray imaging largely stems from its relative benefits in terms of lower cost of deployment and the limited levels of radiation exposure when producing the images (U.S. FDA, 2023).

Table 1. Fracture classification descriptions.

CLASS	DESCRIPTION
AVULSION FRACTURE	FRAGMENT OF BONE AT TENDON OR LIGAMENT SEPARATED FROM MAIN BONE.
COMMINUTED FRACTURE	BONE BROKEN IN AT LEAST TWO PLACES.
DISLOCATION FRACTURE	BONE MISALIGNED FROM NORMAL POSITION.
GREENSTICK FRACTURE	BEND IN THE BONE CAUSING CRACKS ON ONE SIDE WITHOUT A COMPLETE BREAK.
HAIRLINE FRACTURE	SMALL CRACKS OR SEVERE BRUISING IN THE BONE (ALSO KNOWN AS STRESS FRACTURES).
IMPACTED FRACTURE	SIMILAR TO COMMINUTED FRACTURE BUT WITH BROKEN ENDS COMPRESSED TOGETHER.
LONGITUDINAL FRACTURE	BREAK ALONG THE LENGTH OF THE BONE.
OBLIQUE FRACTURE	BONE BREAK AT AN ANGLE.
PATHOLOGICAL FRACTURE	WEAKENED BONE AREA NOT CAUSED BY TRAUMA, USUALLY DUE TO AN UNDERLYING CONDITION.
SPIRAL FRACTURE	BREAK THAT WRAPS OR TWISTS AROUND THE BONE.

3. Literature Review

3.1. FracAtlas: A Dataset for Fracture Classification

In 2023 (Abedeen et al., 2023) introduced the FracAtlas dataset, a curated collection of 4,083 musculoskeletal radiographs spanning hand, leg, hip, and shoulder regions, each manually reviewed by two radiologists and verified by an orthopedist for both fracture presence and anatomical region. Of these, 717 images contain fractures, yielding 922 total fracture instances annotated with both bounding boxes and pixel-level masks, while all scans carry global labels indicating view (frontal, lateral, oblique), region, hardware presence, and fracture count (Abedeen et al., 2023).

3.2. Kaggle Medical Imaging

In 2024, (Basak, 2023) released the Medical Imaging Bone Fracture Colorized Image Data on Kaggle, comprising pseudo colored X-ray and CT scans annotated for fracture presence and anatomical region. Basak's baseline "bonefac" notebook fine tunes ImageNet-pretrained EfficientNet B0, applying stratified five fold cross validation and the Adam optimizer while incorporating extensive data augmentation (random rotations, flips, brightness/contrast shifts) to discriminate fractured from non-fractured samples. Kaggle Community contributors have since adapted YOLO-based detection pipelines such as YOLO v8 implementations by Chowdhury Rituparna and YOLO v7 models by the mdciri team to first localize fracture regions and then classify them, achieving high real-time detection precision on lower-quality or unprocessed images.

3.3. Other Related Work

Over recent years, deep learning architectures have become increasingly sophisticated for medical imaging tasks, including automated fracture detection. This has led researchers to apply state-of-the-art CNNs, attention modules, and even transformer models to bone fracture detection and classification. A non-exhaustive list of recent approaches is as follows:

1. (Tahir et al., 2024) proposed an ensemble of multiple CNN backbones (MobileNetV2, VGG16, InceptionV3, and ResNet50) to perform binary fracture detection on X-ray images of the humerus. The ensemble leverages histogram-equalized preprocessing and global average pooling to aggregate features. On the public MURA-v1.1 humerus dataset, this model achieved high performance (92.96% accuracy, recall and F1 92%), outperforming the individual architectures.

2. (Oh et al., 2023) integrated novel feature-fusion and attention mechanisms into CNN classifiers for wrist fracture detection on X-rays. They augmented EfficientNet-B0 and DenseNet169 backbones with a HyperColumn approach and a Convolutional Block Attention Module (CBAM) to enhance localization of fracture features. This hybrid model improved detection performance: for example, the AUC of the DenseNet- based model rose from 87.78% to 91.45% when HyperColumn and CBAM were applied. Grad-CAM visualizations further showed that the network's attention aligned well with the actual fracture sites.

3. (Chen et al., 2024b) designed WCAY, an object-detection model based on YOLO for multi-site X-ray fracture localization. Their WCAY model replaces portions of the standard YOLO backbone with dynamic snake convolutions (DSConv) to better capture elongated bone structures, and incorporates a novel weighted channel attention (WCA) mechanism to fuse features across the network. On public fracture X-ray datasets (FracAtlas and GRAZPEDWRI-DX), WCAY outperformed baseline YOLO by 8.8% in mean average precision, and achieved 93.9% accuracy on the "fracture" class. This demonstrates that adding specialized attention modules can significantly boost CNN-based fracture detection.

4. (Guan et al., 2024) introduced an enhanced vision transformer (ViT) model for single-stage femur (thighbone) fracture detection on X-ray images. They modified a Pyramid Vision Transformer to include overlapping patch embed-

dings (preserving spatial continuity) and two new attention types (scale-aware and spatial-aware) to fuse multi-scale features. Evaluated on a curated dataset of 4,000 thigh-bone X-rays, the transformer-based detector achieved an AP (average precision) of 53.7% and an AP50 of 87.0%, surpassing previous CNN-based methods on this task. This study highlights the emerging role of transformer architectures in medical image fracture detection.

5. (Wang et al., 2022) applied CNNs to CT imaging for multi-region mandibular fracture detection. Their pipeline first used a 2D U-Net to segment the mandible bone on each spiral CT slice, and then applied a ResNet-based classifier to detect fractures in each of nine anatomical subregions of the mandible. In a large dataset of 686 patients (with 1,506 annotated fractures), this approach achieved over 90 per cent accuracy in each subregion (mean AUC 0.956). This demonstrates effective use of combined segmentation and classification CNNs for CT-based fracture analysis.

3.4. Our Contribution

In our work, we design and train custom convolutional neural networks for fracture classification by leveraging transfer learning with state-of-the-art backbones. We implement a flexible model-building function using a pre-trained InceptionV3 base (and analogously a MobileNetV2 backbone) with custom modifications: after the convolutional base we add a global average pooling layer and dropout, followed by a final Dense output layer with sigmoid (binary) or softmax (multiclass) activation . We freeze most of the pre-trained layers and selectively fine-tune only the top layers (except for batch-normalization layers) to adapt high-level features to our medical data . All models are compiled with the Adam optimizer and appropriate loss (binary cross-entropy for the two-class detector, sparse categorical cross-entropy for the ten-class classifier) and trained with early stopping and checkpointing. These architectural and training design choices enable strong predictive performance even on limited and noisy X-ray data.

Key innovations in our approach are:

1. Two stage cross dataset pipeline

We introduce a two stage classification pipeline that first trains a binary fracture detector on the high quality FracAtlas dataset to pre filter images, and then applies a multiclass classifier on the curated MFC dataset for subtype identification. While prior studies such as WCAY (Chen et al., 2024b) and (Tahir et al., 2024) focus on single stage models applied to one dataset, our approach uniquely integrates two public datasets in series, reducing label noise and class imbalance before multiclass training and thereby improving fracture type accuracy.

2. Sensitivity oriented thresholding and fine tuning

We employ an extremely low decision threshold (0.11, based on Youden's statistic) on the binary detector to reduce the false negatives during pre-filtering, preserving all true fracture instances for the multiclass stage. In parallel, we adopt a selective fine tuning strategy on pretrained EfficientNet and Inception backbones, freezing most convolutional layers except batch normalization and the final few blocks. This combination of recall prioritization and targeted adaptation to medical imagery offers a robust and efficient training scheme not previously reported in fracture classification research.

4. Data

We utilise different dataset for each stage of our pipeline.

Stage 1 is a binary classification task to detect whether an X-ray image contains a fracture. We used the FracAtlas dataset, which comprises 4,083 X-ray images (717 positive, 3,366 negative). The image quality is high, and this dataset has been used in prior studies (Chien et al., 2024; Galić et al., 2023).

Stage 2 addresses multiclass fracture classification using the MFC dataset (Multiclass Fracture Classification), which contains 1,132 fracture images annotated with ten fracture types (between 80-156 images per class) and is publicly available [1]

Before multiclass classification, all 1,132 MFC images are first passed through the binary model to verify the presence of a detectable fracture; this step is necessary because MFC images are of lower quality and largely unprocessed. By filtering out images without clear fracture features, we ensure that only samples with detectable fractures are included in the final analysis.

Hence, the second stage model is intended to assist clinicians by specifying the fracture type, for images where a fracture is detected.

5. Evaluation Metrics

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Table 2. Binary Classification Confusion Matrix

We now present metrics used for our classification problems and discuss when each is most appropriate [2].

[1] https://kaggle.com/datasets/pkdarabi/bone-break-classification-image-dataset

[2] refer to Appendix for different measures considered, including ROC, TPR, TNR

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Specificity} = \frac{TN}{TN + FP}$$

$$\textbf{Sensitivity} = \frac{TP}{TP + FN}$$

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{F}_1\textbf{-Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

$$\textbf{Youden's Index} = \text{Sensitivity} + \text{Specificity} - 1$$

For the FracAtlas dataset, measures such as accuracy may not reflect our model's performance well due the imbalance in classes.

Specifically, for our fracture detection model we try to minimise the Type **II** (false negatives) as this would result in the patient not getting treated despite having a fracture. Hence, focus on the Sensitivity measure would be appropriate to ensure fewer instances of fractures are incorrectly classified.

However, as a greater proportion of all instances are assigned to the positive class, Sensitivity approaches 1. Hence, we also report the Youden's Index score. This provides balance between the Specificity and Sensitivity measures and is particularly useful for binary classification medical contexts by balancing the trade-off between False Positives and False Negatives (Youden, 1950; Böhning et al., 2008).

For the MFC dataset -following our pipeline- all images would have a fracture present. Hence, our models should be able to identify which class a fracture belongs to guide treatment plans. This shifts the importance to the Precision measure, as it considers how well diagnoses are made for within the same class.

Finally, note that the F_1-Score is the harmonic mean of Precision and Sensitivity, making it a valuable metric for simultaneously evaluating both measures for both datasets.

6. Model Selection

6.1. Basic CNN Architecture

We implemented a basic Convolutional Neural Network (CNN) model to serve as a comparison baseline for multiclass image classification. The model consists of three convolutional blocks, followed by a fully connected (FC) layer with 512 neurons. This CNN architecture has approximately 51.6 million parameters, of which 1,920 are non-trainable. This is summarised in Table 3.

Each convolutional block comprises two convolutional layers with 3×3 kernels and 'ReLU' activation functions, followed by batch normalization, max pooling and dropout between each layer. The number of filters increases progressively from 32 in the first block, to 64 in the second, and 128 in the third. This design helps the model learn more detailed and complex features as it goes deeper.

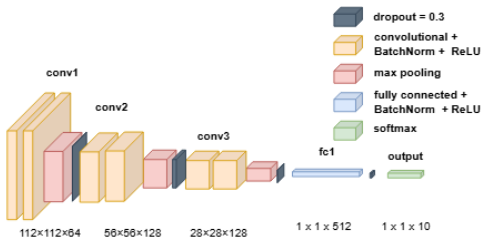

Figure 1. Basic CNN Architecture

After the convolutional blocks, the feature maps are flattened and passed through a fully connected dense layer with 512 units, again using batch normalization and dropout.

Note that for regularisation and to reduce overfitting, Dropout rate of 0.3 is applied after each convolutional and dense layer as well as with batch normalization being used to improve convergence.

Layer	Input Size	Parameters
Conv2D + BatchNorm	$224 \times 224 \times 3$	896
Conv2D + BatchNorm	$224 \times 224 \times 32$	9,472
MaxPooling2D	$224 \times 224 \times 32$	0
Dropout (0.3)	$112 \times 112 \times 32$	0
Conv2D + BatchNorm	$112 \times 112 \times 32$	18,688
Conv2D + BatchNorm	$112 \times 112 \times 64$	37,248
MaxPooling2D	$112 \times 112 \times 64$	0
Dropout (0.3)	$56 \times 56 \times 64$	0
Conv2D + BatchNorm	$56 \times 56 \times 64$	74,240
Conv2D + BatchNorm	$56 \times 56 \times 128$	147,584
MaxPooling2D	$56 \times 56 \times 128$	0
Dropout (0.3)	$28 \times 28 \times 128$	0
Flatten	$28 \times 28 \times 128$	0
Dense + BatchNorm	$1 \times 1 \times 100352$	51,379,200
Dropout (0.3)	$1 \times 1 \times 512$	0
Dense (Softmax)	$1 \times 1 \times 512$	5,130
Total	–	**51,682,458**

Table 3. Basic CNN architecture

6.2. InceptionV3

The original InceptionV1 architecture (Szegedy et al., 2015) -introduced as GoogleLeNet in 2014- approached architecture design by increasing width and depth of the model simultaneously.

The InceptionV3 CNN architecture from 2015 (Szegedy et al., 2016) iterates on this and is optimized eing the model wider rather than solely relying on multiple deep layers (seen in Figure 2). It has a total of 42 major building blocks (convolutional and pooling layers, ignoring activation functions and auxillary layers) and a lower error rate than the previous Inception versions.

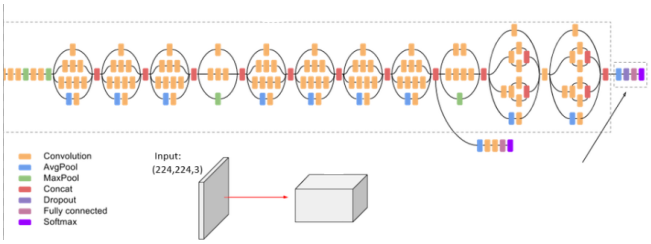

Figure 2. InceptionV3 Model architecture (inc)

The Inception V3 model has consistently demonstrated strong performance across a variety of medical imaging applications, including skin lesion detection (Chen et al., 2024a) where it was used to detect Monkeypox lesions, achieving an accuracy of 96.71%, other studies have applied the InceptionV3 for Breast Cancer Image classification (Ayana et al., 2022) and lung cancer classification (Kumaran S et al., 2024).

From fine-tuning the InceptionV3 architecture, we hoped to replicate the balance of computational efficiency and architectural depth to enable feature extraction from X-ray images. Our results find this model performing best for the classifying the both the FracAtlas and MFC dataset, hence, it is also used as the backbone for Stage 2 in our pipeline.

Fine-tuning InceptionV3:

We use the pretrained ImageNet weights - to leverage transfer learning- and exclude the original classification head. For fine-tuning we unfroze pre-trained model from the layers indexed 249 and onward , while adding a global average pooling layer, followed by a dropout layer (0.5) to reduce overfitting.

To ensure stable convergence, we set BatchNormalisation (bn_trainable=True) layers as trainable, during the finetuning process, this is generally good practice in transfer learning scenarios (Li et al., 2016).

Finally, based on which stage we use the InceptionV3 Model, we specify the size of the dense output layer, with a sigmoid activation function for the binary problem, and softmax activation function multiclass classification problem.

Table 4. Custom InceptionV3 Architecture (Modified for Multi-class Classification)

Layer	Details	Output Size
Input	–	$224 \times 224 \times 3$
Conv Stem	Varied	$\sim 112 \times 112 \times 32$
Conv + Pool	Varied	$\sim 35 \times 35 \times 192$
$3 \times$ Inception A	–	$35 \times 35 \times 288$
$5 \times$ Inception B	–	$17 \times 17 \times 768$
$2 \times$ Inception C	–	$8 \times 8 \times 1280$
Conv (Projection)	–	$8 \times 8 \times 2048$
Global AvgPool	–	$1 \times 1 \times 2048$
Dropout (0.5)	–	$1 \times 1 \times 2048$
Dense (Softmax)	3 classes	$1 \times 1 \times 3$

Our choice for freezing the earlier layers was motivated by the earlier layers in the InceptionV3 model learning general features like edges, corners and textures (which would be particularly important after augmentation) and we want to preserve any useful general pretrained knowledge.

We also opt for freezing layers to reduce the number of trainable parameters in our model. This would avoid overfitting to our FracAtlas dataset and allow the model to generalise to identify fractures from our multiclass image dataset, which includes lower quality images.

Table 4 summarises the InceptionV3 architecture we use after fine-tuning.

6.3. EfficientNet-B0

For the MFC dataset, we also use the EfficientNetB0 architecture to classify fractures. First presented by Tan & Le (2019), the model employs a compound scaling strategy (uniformly scales depth, width and resolution) and has been successfully used in past medical image classification problems, particularly demonstrating strong performance in categorizing Brain Tumor MRI images (Mahesh et al., 2024), and skin cancer classification (Kanchana et al., 2024).

Compound scaling doesn't effect the operations used within a layer of the network, only expanding the networks, width, depth and resolution, hence the need for a good baseline network.

The authors designed a baseline network in order to experiment with the scaling, the EfficientNet-B0, which we use, it's design optimises for accuracy and FLOPS (computational cost of running the model).

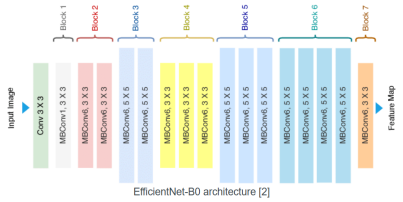

Figure 3. Efficient Net Architecture (Ahmed & Sabab, 2022)

The main building block of this architecture is the Mobile Inverted bottleneck (MBConv), inspired by the MobileNetV2 (Sandler et al., 2018). The MBConv is an inverted residual block with a combination of depthwise separable convolutions, with additional Squeeze and Excitation (SE) optimisation. Here rather than the pattern for a normal residual block, in terms of the number of channels (narrow then wide then narrow), the inverted residual block (wide then narrow then wide) is used. The MBConv layer starts with a depthwise convolution, then a pointwise (1x1) convolution is applied, which expands the number of channels, and finally another (1x1) convolution is applied to reduce the channels back to the original number, giving it that 'bottleneck' design, which allows for effective and efficient feature extraction. The SE block helps the model learn to focus on essential features and suppress less important ones, making it suitable for medical images like X-rays, where only certain parts of the image tells you the diagnosis, the rest is just noise. The SE block makes use of global average pooling to reduce spatial dimensions of the feature map to a single channel and then 2 FC layers are used, allowing the model to learn channel-wise feature dependencies and create attention weights, which are then multiplied to the original feature map.

We primarily use the EfficientNet-B0 model, although by varying the scaling coefficients: $\{depth, width, resolution\}$ there is a trade-off between model-size and accuracy to obtain EfficientNet variants (B1, B2, ... , B7)[3].

Fine-tuning EfficientNet-B0:

The EfficientNet-B0 is used as a backbone for the multiclass image classification. The model is also initialized, with ImageNet pretrained weights and configured without the top classification layer (include_top=False). Input is resized to (224x224x3) and feature representations extracted from the base network were passed through a global average pooling layer, followed by a dropout layer with a rate of 0.5, in order to mitigate overfitting. Finally, a dense output layer is used with 2 units (Binary Fractured/Not fractured problem)/ 10 units (multiclass all 10 fractures)/ 3 units (reduced class problem), and softmax activation is appended to perform classification across the classes.

To allow the method to adapt to the new task, while preserving the general feature representation, we froze the majority of the EfficientNet-B0 layers, except for the batch-norm layers, which remained trainable to assist with overfitting. Fine-tuning was enabled from layer index 200 onward allowing the last 37 layers of the base model to update during training. The model was compiled with the Adam Optimiser (learning rate = 1×10^{-4}) and trained using categorical cross entropy loss. We also employed early stopping and model checkpoint callbacks to preserve the best model and prevent overfitting.

Comment on architectures presented

In addition to the architectures reported above, we also conducted preliminary experiments with a Vision Transformer (ViT) and an ensemble method based on InceptionV3. For the multiclass fracture classification task, the ViT model achieved peak validation accuracy of 14% and demonstrated poor generalisation. We attribute this underperformance to the relatively small dataset size, as transformer-based models generally require large-scale data to outperform convolutional architectures (Dosovitskiy et al., 2021).

For the ensemble method, we trained individual InceptionV3 models on separately augmented subsets of the training data and aggregated predictions across learners. Although the strongest individual model attained 46% accuracy, ensemble aggregation decreased overall performance due to conflicting predictions and overfitting. Given these limitations, we focus our analysis on InceptionV3, which performed best ex-post, and EfficientNetB0, which we hypothesised ex-ante to offer competitive performance on smaller datasets due to its MobileNet-inspired efficiency.

7. Model Training

7.1. Evaluation

For training and evaluation of binary classifier on FracAtlas dataset, we minimize the binary cross entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\left[y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\right]$$

and use the evaluation metrics referenced in **Section 5**

As mentioned previously, these evaluation metric better contextualise our binary classification problem due to the class imbalance present in the FracAtlas dataset. Additionally, specific for the fracture detection problem we present the ROC and AUC metrics to capture the model's overall ranking capability regardless of threshold.

For fracture type classification of X-ray images, we aim to

[3]B0 selected to maximise accuracy while applying constraints to the available resources (memory and FLOPS)

minimize categorical cross entropy:

$$\mathcal{L}_{\text{CCE}} = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i)$$

and we focus accuracy achieved on the test set; for multi-class classification this is evaluated according as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

7.2. Data Augmentation

To improve model generalization and prevent overfitting on the relatively small FracAtlas dataset, we apply a series of data augmentation techniques during training. The transformations include random horizontal flipping to account for lateral symmetry in bone structures, random rotation (±10%) to simulate slight variations in image acquisition angles, and random zooming (±20%) to mimic variations in distance from the X-ray source. We also adjust contrast and brightness by ±20% to account for differences in imaging conditions, equipment, and illumination levels.

For the MCF dataset, a more extensive augmentation strategy was adopted to address the limited dataset size and encourage the model to generalize better to diverse imaging conditions. Each training image was augmented with thirteen distinct color transformations, including gamma correction, Gaussian blur, histogram equalization, contrast stretching, various color mappings, and more. This augmentation is initially done by the Kaggle user 'SHUVO KUMAR BASAK-4004.O' in his notebook 'bonefac' to highlight fractures or abnormal regions, making them easier for doctors and medical professionals to identify and analyze (Basak, 2025).

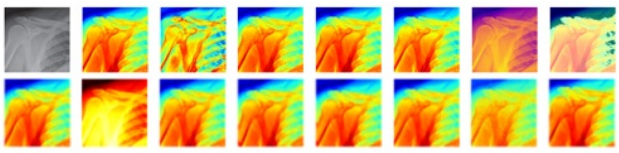

Figure 4. Fracture augmentations applied

In addition to these color-based augmentations, we applied five geometric and intensity-based transformations specifically designed for X-ray data (Chandola et al., 2024): minor rotations to simulate misalignment during capture, translations to account for positioning variability, brightness and contrast adjustments to reflect differences in exposure, Gaussian noise injection to mimic sensor noise, and zoom effects to simulate variable image resolutions. These augmentations are applied to each train image through the customized function 'multi_augment_generator' which takes images, specified batch size, and output image size as input, outputting batches of augmented images. In addition, all augmented images were color-mapped using the jet colormap to ensure visual consistency across transformations.

7.3. Common Methods

All models were trained using the same early stopping and model checkpointing callbacks for consistently comparing results. Early stopping terminates training when validation loss stops improving, thus mitigating overfitting. Model checkpointing was used to save weights corresponding to the lowest validation loss throughout training.

Moreover, for fine-tuning of pre-trained models, images were resized to match the 224×224 pixels input size required. The data augmentations techniques described were consistently applied to the training set for each models being compared and he Adam optimizer was used with categorical cross-entropy loss and learning rates carefully chosen at or below 1×10^{-4}.

Finally, due to limited training samples per class after filtering for images with fractures detected, our Stage 2 models perform multiclass classification to the three best-represented categories in the dataset.

7.4. InceptionV3 Training (FracAtlas Dataset)

First, the InceptionV3 model described in Section 6.2 was trained using the FracAtlas dataset (stage 1). Training employed the Adam optimizer with a learning rate of 1×10^{-2} and binary cross-entropy loss. Training parameters included a batch size of 32 and a maximum of 50 epochs, with early stopping set at a patience of 3 epochs.

7.5. CNN Training (MFC Dataset)

The CNN model was trained on the multiclass fracture (MFC) dataset (stage 2), specifically on three dominant classes: Fracture Dislocation, Comminuted Fracture, and Pathological Fracture. The model was trained using categorical cross-entropy loss and the Adam optimizer with a learning rate of 1×10^{-4} for up to 75 epochs.

7.6. InceptionV3 Training (MFC Dataset)

The InceptionV3 model was then trained on the MFC dataset and utilised similar methodologies to the CNN model, using categorical cross-entropy loss optimized with Adam optimizer (learning rate 1×10^{-2}). Again, training occurred over a maximum of 75 epochs with early stopping (patience set to 10) and model checkpointing based on validation loss employed. This was done due to higher training accuracy was observed early with validation accuracy fluctuated significantly between epochs, indicating potential overfitting.

7.7. EfficientNetB0 Training (MFC Dataset)

The EfficientNetB0 model was trained in two distinct phases to optimise transfer learning. Initially, pretrained convolutional layers were frozen, focusing training solely on dense classification layers for 20 epochs with a learning rate of 1×10^{-4}. In the second phase, upper convolutional layers were selectively unfrozen and fine-tuned for an additional 20 epochs using a reduced learning rate of 5×10^{-6}. Early stopping and learning rate reduction on plateau were employed to mitigate overfitting.

8. Results and Discussion

The InceptionV3 model on the FracAtlas dataset achieves an 87.58% accuracy, with an F_1-Score of 0.61 However, more importantly the Sensitivity and Specificity are 67.83% and 88.37% respectively, resulting a Youden score of 0.56. This suggest moderate - good discrimination between the classes from the binary classification model (Youden, 1950). Figure 7 in A.1.shows the ROC curve with from inceptionV3 classifier.

For the second stage, the evaluation demonstrated significant performance differences among the trained models. InceptionV3 trained on the reduced class MFC dataset exhibited the highest test accuracy at 67.39%, suggesting it better captured distinctions among fracture classes. Conversely, the CNN and EfficientNetB0 models showed lower performance, with test accuracies of 52.17% and 26.09%, respectively. The training and validation accuracies and losses are recorded during training and is presented in Figure 5 and 6 in A.1.

Model Architecture	Test Accuracy (%)
InceptionV3	**67.39**
CNN	52.17
EfficientNetB0	26.09

Table 5. Test accuracy scores for different architectures

Similarly, we compare our models ability to discriminate between the reduced classes using the Precision metric. Again, the EfficientNetB0 model performed worse across the reduced set of classes with Precision being in the range between 0 - 0.26. For the InceptionV3 model this is slightly better by being between 0.18-0.35 and the Basic-CNN architecture had the best range of precision between 0.33-0.36 for each class.

The lower accuracy from the basic CNN's is likely due to the simplicity in design preventing the model from going deep enough to capture specific differences between fracture classes; however, this is balanced by the large number of trainable parameters. While we believe the EfficientNetB0 model performed poorly due to the limited dataset size.

This project aimed to extend the previous research focusing on single region of interests and the binary fracture classification problem by utilising image augmentation techniques.

Our results present the difficulty of using a model trained over a limited sample to classify different types fractures across multiple regions of interests. However, when focusing on the smaller problem of limiting the number of fracture classes based on sample availability, we did observe improved accuracy, suggesting that for a sufficiently large database, Computer Assisted Diagnosis methods have scope for improvement.

This emphasises the importance of matching model complexity with available data, especially in healthcare settings where individual heterogeneity and privacy concerns results in limited and noisy datasets.

9. Conclusion

This study introduces a novel two-stage deep-learning pipeline for bone fracture classification from X-ray images, integrating two public datasets in sequence. In the first stage, a binary fracture detector is trained on the high-quality FracAtlas dataset to pre-filter images; an extremely low decision threshold (0.01) is applied to maximize sensitivity and ensure virtually no true fractures are missed. The filtered images are then fed into a second-stage multiclass classifier trained on the curated MFC dataset for detailed fracture-type identification. Three model architectures were evaluated in this pipeline: a custom CNN, EfficientNetB0, and InceptionV3, each with selective fine-tuning of pretrained layers. The InceptionV3-based model achieved the highest accuracy on the multiclass task, while the CNN and EfficientNetB0 models were less accurate and prone to overfitting on the limited data.

Further Work: Despite these advances, several limitations point to areas for future improvement. The multiclass fracture dataset is relatively small and imbalanced across classes, contributing to limited generalization and overfitting in complex models. Future work should explore stronger data augmentation or synthetic oversampling to balance the classes and mitigate this issue, as well as cost-sensitive or weighted loss functions to handle label imbalance. Incorporating model interpretability techniques (e.g. Grad-CAM) could help clinicians understand the model's predictions and increase trust. Integrating an explicit fracture localization step (for example, a segmentation or object detection model to pinpoint fracture regions) might further enhance accuracy. Finally, validating and extending this pipeline on larger, multi-center X-ray datasets (and potentially other imaging modalities) will be critical to ensure robustness and clinical utility in diverse real-world settings.

References

Advanced guide to inceptionv3. https://cloud.google.com/tpu/docs/inception-v3-advanced. Accessed: Dec. 12, 2021.

Abedeen, I., Rahman, M. A., Prottyasha, F. Z., Ahmed, T., Chowdhury, T. M., and Shatabda, S. Fracatlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs. *Scientific data*, 10(1): 521, 2023.

Ahmed, T. and Sabab, N. H. N. Classification and understanding of cloud structures via satellite images with efficientunet. *SN Computer Science*, 3(1):99, 2022.

Ayana, G., Park, J., Jeong, J.-W., and Choe, S.-w. A novel multistage transfer learning for ultrasound breast cancer image classification. *Diagnostics*, 12(1):135, 2022.

Basak, S. K. Medical imaging bone fracture colorized image data. https://www.kaggle.com/datasets/shuvokumarbasak2030/medical-imaging-bone-fracture-colorized-img-data, 2023. Accessed: 2025-05-02.

Basak, S. K. Medical imaging bone fracture colorized img data, 2025. URL https://www.kaggle.com/dsv/11124104.

Böhning, D., Böhning, W., and Holling, H. Revisiting youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. *Statistical Methods in Medical Research*, 17(6):543–554, 2008.

Chandola, Y., Uniyal, V., Bachheti, Y., Lakhera, N., and Rawat, R. Data augmentation techniques applied to medical images. 5:483–501, 07 2024.

Chen, J., Lu, Z., and Kang, S. Monkeypox disease recognition model based on improved se-inceptionv3. *Journal of Intelligent & Fuzzy Systems*, 46(4):8811–8828, 2024a.

Chen, P., Liu, S., Lu, W., Lu, F., and Ding, B. Wcay object detection of fractures for x-ray images of multiple sites. *Scientific Reports*, 14(1):26702, 2024b.

Chien, C.-T., Ju, R.-Y., Chou, K.-Y., and Chiang, J.-S. Yolov9 for fracture detection in pediatric wrist trauma x-ray images. *Electronics Letters*, 60(11):e13248, 2024.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Etli, I., Kozaci, N., Avci, M., and Karakoyun, O. F. Comparison of the diagnostic accuracy of x-ray and computed tomography in patients with wrist injury. *Injury*, 51(3): 651–655, 2020.

Galić, I., Habijan, M., Leventić, H., and Romić, K. Machine learning empowering personalized medicine: A comprehensive review of medical image analysis methods. *Electronics*, 12(21):4411, 2023.

Guan, B., Yao, J., and Zhang, G. An enhanced vision transformer with scale-aware and spatial-aware attention for thighbone fracture detection. *Neural Computing and Applications*, 36(19):11425–11438, 2024.

Kanchana, K., Kavitha, S., Anoop, K., and Chinthamani, B. Enhancing skin cancer classification using efficient net b0-b7 through convolutional neural networks and transfer learning with patient-specific data. *Asian Pacific Journal of Cancer Prevention: APJCP*, 25(5):1795, 2024.

Khan, A. A., Fatima, Z., Bacha, R., Rokhan, B., Akhtar, S., Iqbal, M., and Raheem, I. Comparing ultrasonography with plain radiography in the diagnosis of long bone fractures. *Journal of Diagnostic Medical Sonography*, 39(6): 575–587, 2023.

Kumaran S, Y., Jeya, J. J., Khan, S. B., Alzahrani, S., and Alojail, M. Explainable lung cancer classification with ensemble transfer learning of vgg16, resnet50 and inceptionv3 using grad-cam. *BMC medical imaging*, 24(1): 176, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y., Wang, N., Shi, J., Liu, J., and Hou, X. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Mahesh, T., Gupta, M., Anupama, T., Geman, O., et al. An xai-enhanced efficientnetb0 framework for precision brain tumor detection in mri imaging. *Journal of Neuroscience Methods*, 410:110227, 2024.

NHS England. Diagnostic imaging dataset annual statistical release 2023/24. https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/, April 2024. Accessed: 2025-05-02.

Oh, J., Hwang, S., and Lee, J. Enhancing x-ray-based wrist fracture diagnosis using hypercolumn-convolutional block attention module. *Diagnostics*, 13(18):2927, 2023.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tahir, A., Saadia, A., Khan, K., Gul, A., Qahmash, A., and Akram, R. Enhancing diagnosis: ensemble deep-learning model for fracture detection using x-ray images. *Clinical Radiology*, 79(11):e1394–e1402, 2024.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

U.S. FDA. Medical x-ray imaging, 2023. URL https://www.fda.gov/radiation-emitting-products/medical-imaging/medical-x-ray-imaging. Accessed: 2025-05-02.

Wang, X., Xu, Z., Tong, Y., Xia, L., Jie, B., Ding, P., Bai, H., Zhang, Y., and He, Y. Detection and classification of mandibular fracture on ct scan using deep convolutional neural network. *Clinical Oral Investigations*, 26(6):4593–4601, 2022.

Youden, W. J. Index for rating diagnostic tests. *Cancer*, 3 (1):32–35, 1950.

A. Appendix

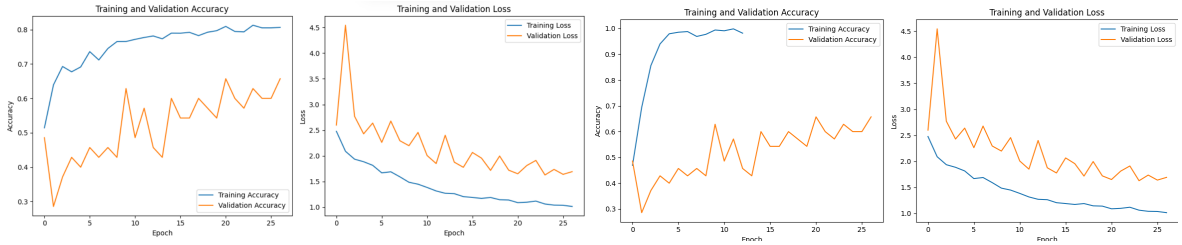

Figure 5. Training and validation accuracy for CNN(left) and InceptionV3(right)

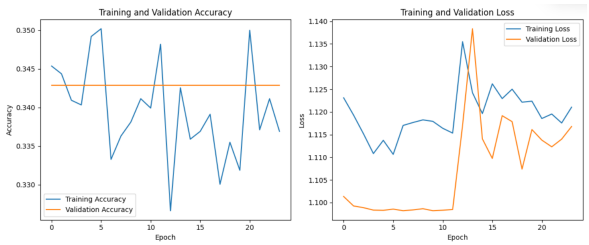

Figure 6. Training and validation accuracy for EfficientNet-B0

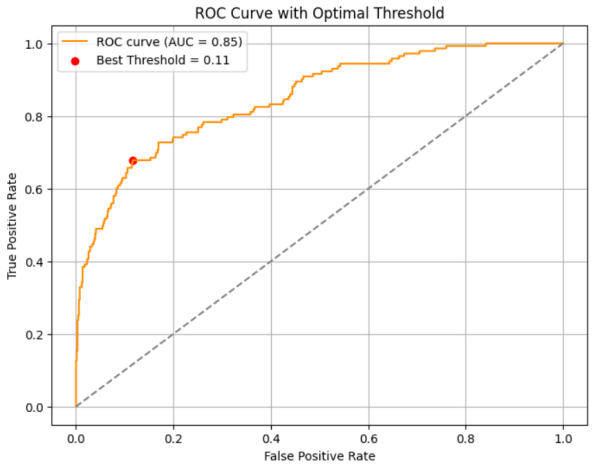

Figure 7. ROC curve from InceptionV3 FracAtlas Classifier

Contribution Statements: each 25%