SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

BONE ABNORMALITIES DETECTION USING MEDICAL IMAGING

Muhammad Noman Khan[1], Maham Shahzadi[2], Awais Raza Qadri[3], Shahzaib Nazar[4], Um E Habiba[5], Zaeem Nazir[6]

[*1,2,3,4,5,6]*Department of Computer Sciences, University of Narowal*

[1]nomankhan7256@gmail.com , [2]mahamzunaira6@gmail.com , [3]razaqadri0306@gmail.com , [4]shahzaibn192@gmail.com, [5]umehabibakhan9@gmail.com, [6]zaeem.nazir@gmail.com

DOI: https://doi.org/10.5281/zenodo.21062147

Keywords
Bone Tumor and Fracture Detection (BTFC), X-ray Imaging (XRI), Deep Learning (DL), Convolutional Neural Networks (CNN), Explainable Artificial Intelligence (XAI), Bone Abnormalities (BA).

Article History
Received: 25 April 2026
Accepted: 04 June 2026
Published: 21 June 2026

Corresponding Author: *
Muhammad Noman Khan

Abstract
The early detection of any abnormalities in the bone, such as fracture and tumor, is of great importance for the timely clinical intervention. Lastly, the conventional diagnosis by X-ray is prone to errors, particularly in the cases where the image quality is poor, and there are only slight abnormalities. In this paper, we explain the OrthoVision system an automated system for bone abnormality detection and classification based on artificial intelligence for X-ray images, which will be implemented on a web-based basis. The system is able to classify into 4 classes fracture, non-affected, tumor-benign and tumor-malignant. It is based on EfficientNet-B0 and ResNet-18 ensemble model, along with CLAHE enhancement, resizing and converting to and normalizing a tensor. Visual explanation of model predictions, using GradCAM++, can be used to aid clinical interpretation. The validation performance reflects reliable predictions and generalizations, addressing the trustworthiness of the AI's accuracy and its potential usefulness for AI in real-world bone diagnostic tasks. The validation outcomes confirm effective predicative and generalizing performance, further emphasizing the usability of AI in bone diagnostics. The validations highlight trustworthy predictions and generalizations, underscoring the potential of using AI in bone diagnostics.

1. Introduction

Abnormalities in bones (such as tumor and fracture) are a serious clinical issue in weltweit. It is crucial that these abnormalities can be safely, effectively identified early and at the right time to provide clinical action and optimal post clinical outcomes and reduce morbidity over a whole patient's lifetime. A fracture is a discontinuity of the bone due to trauma, pathological conditions, and if left untreated, can result in impaired mobility, chronic pain and/or complications. Similarly, the diagnosis of bone tumors also poses some difficulty due to the varying morphology, and on account of their capacity to disrupt integrity of the bones. In particular, malignant bone tumours are of prognostic significance, they tend to progress quickly and are prone to spread to other parts of the body and early and correct diagnosis is therefore critical.[1]

Traditionally, the imaging modalities used in medicine have been the mainstay to detect bone abnormalities. X-ray imaging comes with low prices, is non-invasive and simple to get and while it doesn't provide the most cutting-edge technology available, it remains one of the most readily available and ubiquitous modalities used for diagnosis. Uses the expertise and experience of a radiologist, but relies on the traditional "way of reading x-rays. This may be difficult to

diagnose, or misdiagnose because the radiologist may miss important changes in the early stages of a tumor or minor fractures may result from overlapping bones. One of the effects of this is called "satisfaction of search" wherein presence of a main abnormality will distract attention from other abnormalities, further underlining the limitations of manual analysis.[2]

With the advent of AI and DL, the world of medical imaging analysis has changed. The highly complex medical images have high hierarchical features, where convolutional neural networks (CNNs), a type of deep learning model, has demonstrated its high potential to understand it. CNNs can be taught the spatial and textural patterns in bone structures, which are hard for human observers to detect. To make full use of the pre-trained networks like ResNet and EfficientNet, efficient feature extraction and model performance can be obtained by using transfer learning techniques with relatively small medical datasets.[3]

These CNNs like networks can be beneficial in addressing various critical clinical needs such as automatic bone abnormality detection. It's a second opinion tool for radiologists that further enhances the diagnostic confidence and reduces human error. Secondly, it allows high-throughput analysis typically in a clinical setting, where high tanning request and critical decisions are required, including emergency situations. Finally, during the clinical application of such systems, AI algorithms can incorporate XAI techniques like GradCAM++ to visualize image regions that impact model predictions, thereby improving the transparency and trustworthiness of this system in practice.[16]

While all this has taken place, however, there remain huge challenges. Existing models of deep learning are able to either detect fracture or classify tumor, without providing an overall diagnostic model. In addition, there is limited generalization ability to different imaging modalities and patient groups, as some models show good performance on the training set, but not as well on clinical images that were not included in their training. Other issues for practical applications are computational resources, the interpretability of models and the possibility of using several models in a single pipeline.[4]

The goal of this study is to develop an AI system called OrthoVision, which can detect and classify bone fracture and tumor at the same time in X-ray images in one system. OrthoVision features a powerful ensemble model, which consists of two networks: EfficientNet-B0 and ResNet-18, and a well-designed pre-processing path, which involves CLAHE enhancement, resizing, converting to a tensor and normalization. Four class outputs are created by the system: fracture, non-affected, tumor-benign and tumor-malignant. The generated visual explanations from GradCAM++ are used to explain the model predictions, thus assisting the clinicians in interpreting the model predictions. This system works as a second opinion diagnostic tool which will complement radiologists' workflow and help in the early detection of bone abnormalities that will result in better patient care and clinical outcomes.

Objectives:

The following were the main goals of this research:

i) To determine the most dependable and generalizable deep learning model that can be applicable in the real world for diagnosis of bone tumors.

ii) Heatmap-based visual explanations were used to combine Artificial Intelligence techniques to improve model transparency and clinical interpretability.

iii) To assist clinical decision-making with a stable, accurate, and interpretable AI-assisted diagnostic system to detect early bone tumors.

2. Literature Review

With the recent development of deep learning, specifically the (CNNs) technology has been developed greatly and has demonstrated its outstanding feature extraction and classification tasks on medical imaging. Reddy et al. [7] proposed a model using CNN trained using 2500 X-ray images to detect fractures. The Architecture consisted of a typical Convolutional Network layered with ReLU activation units and Max Pooling layers for learning Hierarchical Spatial Features of bone fracture. Model validation resulted in an

accuracy of 92.4% where there was a high sensitivity towards recognizing prominent fractures. One drawback of the approach was its focus on fracture only and not taking into account multi-class or tumor classification. Furthermore, they had not considered the interpretability of the study, based upon the graph embedding with XAI techniques which is one of the key factors in clinical interpretability. Drawing on this, Zhang et al. [8] constructed a multi-class CNN with the ability to differentiate normal bone, benign tumor and malignant tumor. This study used a large dataset of X-ray images (3,000 images) with the use of multiple convolutional layers to capture fine-grained features that capture subtle differences in the texture and structure of the tumor morphology. The validation accuracy of the model was 93.1%. The three classes were a step beyond the binary classification previously used, but the paper offered no methods of explaining the predictions (such as GradCAM) and interpretability was still limited, which could decrease the trust that a clinician may have in the automated results.

Kumar et al. [9] used transfer learning by fine-tuned pre-trained ResNet architectures with four thousand images. With pre-training on ImageNet, the model learnt much more quickly and reached an 11 per cent accuracy in the test set. This study has merits that it trains an efficient network and utilizes previous feature representations and utilizes them effectively. Despite this, it was not possible to extend the test set to unseen clinical data based on the controlled, small size of the data set; larger, more diverse datasets are required for its robust performance in the real world.

Ali et al. [10] presented a framework of deep CNN for the automatic detection of bone tumours in 3,500 X-rays. The use of augmentation patterns was focused on the detection of benign and malignant tumors with an emphasis on improving the ability of the model to generalize. Although good results were reaped from the detection of patterns of tumors larger than the detection window, the detection results for the smaller and more subtle tumor anomalies were not fully reported and the lack of explainability tools meant that a great extent of the results were not very clinically useful. Class imbalance also remained limiting parameter as there are very few cases of malignancy, and augmentation was necessary, as was the use of GAN for generating synthetic images.

Hassan et al. [11] proposed a hybrid CNN with both convolutional and fully connected layers that was used to classify multiple classes of tumors from a set of 5,000 X-ray images. The architecture enabled finer features to be extracted from the texture and intensity changes within the bone. The model performed well, with an accuracy rate of 95.2% in the validation process. But the larger size of the network meant that it was not feasible for low resource settings and for high volume settings high inference times may affect clinical workflows.

Wang et al. [12] delved deeper into deep CNN structures, educated on 6000 X-ray pictures, targeting the ability to uncover complex tumor patterns. The model reached a high validation accuracy rate of 96%, and was able to localize tumors in complicated anatomical structures very well. However, the model was deep, which inflicted the risk of overfitting, and the training time was a formidable number, emphasizing the accuracy and computational considerations of clinical deployment of the model. To give emphasis to inference speed and computational efficiency.

Patel et al. [13] explored lightweight CNN architectures with 3200 X-ray images. The overall accuracy was 90.8%, indicating the model could be used in resource-constrained settings, though the performance was limited on subtle variations in tumors, likely because of the lack of features. Singh et al. [14] developed CNN architecture that focuses on early detection thus reducing the chances of developing serious complications with a data set of 4,500 images of tumors. The accuracy was 94.6%, while the model also provided with timely detection is essential for clinical interventions. One drawback however, was the absence of explainability in a visual manner which hindered the clinicians' confidence, especially with complex cases where the AI decision needed to be validated. To classify multiple classes of tumors.

Rahman et al. [15] introduced an undeniable CNN trained on 4,800 images, that showed

95% accuracy in its validation test. While it showed the need for a good feature learning for generalization, the model was never incorporated in fracture detection, limiting its clinical application in full bone abnormality assessment. Ahmed et al. [16] tested their high-capacity CNN on 5,200 X-ray images and got 97.1% accuracy. Whilst they can isolate highly detailed tumor attributes they suffer the issue of overfitting on external data sets, which poses a problem for deployment with real-world patients. This study highlighted the need for adequate compromise between network spectrum and data set variety and regularization approaches.

An ensemble of EfficientNet-B0 and ResNet-18 are combined in the OrthoVision system [17] for the multi-class classification of fractures, non-affected bones, benign tumors, and malignant tumors. To counteract the difference between each model and make the predictions consistent across all classes, images are preprocessed in the system using CLAHE, resized to the image of 224×224, tensor converted, and normalized. Preliminary testing assures high predictive reliability, and GradCAM++ offers valid "heat maps" for clinical review. Although the integrated 4-class metrics are not yet available in the test set, the results presented here are a good example of a single strategy for unified multi-class bone abnormality detection. To guarantee diversity in imaging devices, patient demographics and anatomical variability.

Yao et al. [18] built a multi-institutional radiograph data set, with X-ray images from multiple hospitals. They used a ResNet based CNN network for the classification of bone tumors into benign or malignant class. The network had about 92% accuracy, which shows that the network was well generalised across institutions. However, there were minor differences in imaging technologies, like exposure, resolution and anatomical positioning that resulted in some performance differences. One of the key takeaways from this study is the necessity for integrating data from various institutions to train models that can be applied consistently across different clinical contexts, a typical challenge faced when implementing AI in clinical settings. The study highlights the need for multi-institutional data to help train models that can be broadly applied across clinical settings, which is often difficult and complex, especially when working with AI. Additionally, Shin et al. [19] employed the EfficientNet-B5 architecture with the transfer learning approach, using a very large annotated x-ray database that contained tens of thousands of x-rays. The method they used was to fine tune a CNN that was pretrained on the ImageNet, which translated to including low-level features such as edges and textures while also learning high-level image information related to the anatomy involved in bone tumour detection. With a primary bone tumor classification, 97% accuracy was obtained from the model. The results of this study emphasize the merits of large-scale pre-training and domain-specific fine-tuning, for the classification of medical documents it significantly enhanced the classification accuracy without requiring extremely sized labeled medical databases.

Tajbakhsh et al. [20] compared the effect of the classical CNNs training with fine tuning these networks for the classification of bone X-ray images. With a small set of labelled images, they showed that transfer learning made a significant difference in the performance of their models. By refining pretrained networks, ~94% validation accuracy was obtained with a slightly lower level of overfitting than when training a network from scratch. The research highlights the importance of using pre-trained feature representations, as they can enhance the performance of models on medical images, which are frequently characterized by limited labeled datasets, allowing for more effective learning from smaller sets and improved generalization to novel data.

A Generative Adversarial Network (GAN) was used to boost the amount of X-ray data by J.Chen et al. [26]. A reasonable set of realistic bone tumor images was synthesized using GANs, especially from less representative classes such as malignant tumors. The study demonstrated the augmentation using GAN improves the sensitivity for rare classes of tumors and thus helps CNN to be more generalizable to the tumor morphologies that it has not seen. This approach tackles the issue of class imbalance found in medical imaging, and showed that generating synthetic data can improve the

robustness of the model without compromising its clinical utility.

L. **Tanzi**, et al. [21] summarized the state of deep learning in medical imaging in a thorough review about best practices in training, modeling, and assessing medical imaging data. The survey indicated that balanced datasets, good augmentation techniques, and explainability methods (e.g., GradCAM or saliency maps) are among the key assets that need to be accepted in the clinical setting. Clinicians reduced trust in models whether or not they are accurate, but if they lack interpretability. One of the key takeaways from the review highlights the need for strong performance on multi-modal datasets when bringing AI models to life in real-world clinical settings

Literature Review Comparison:

Table 1: Literature review summary of CNN-based bone abnormality detection methods, showing dataset, accuracy, strengths, limitations, and XAI usage.

Author	Year	Model	Accuracy	Strengths	Limitations	XAI Used	Implement
Reddy et al. [5]	2020	CNN	92.4%	Sensitive to fractures	No tumor detection	No	No
Zhang et al. [6]	2021	CNN Multiclass	93.1%	Multi-class tumor	No interpretability	No	No
Kumar et al. [7]	2021	Transfer Learning	94%	Faster convergence	Limited generalization	No	No
Ali et al. [8]	2022	CNN	Not reported	Subtle tumor detection	Small dataset, no explainability	No	No
Hassan et al. [9]	2022	Hybrid CNN	95.2%	Fine-grained classification	Computationally heavy	No	No
Wang et al. [10]	2023	Deep CNN	96%	Detailed feature extraction	Overfitting risk	No	No
Patel et al. [11]	2023	Lightweight CNN	90.8%	Fast inference	Low capacity for complex patterns	No	No
Singh et al. [12]	2024	CNN framework	94.6%	Early detection	No explainability	No	No
Rahman et al. [13]	2024	CNN generalization	95%	Generalizable	No fracture detection	No	No
Ahmed et al. [14]	2025	High-capacity CNN	97.1%	High accuracy	Overfitting	No	No
Proposed	2026	Ensemble CNN	95%	Multi-class + GradCAM++	Pending full integrated metrics	Yes	Yes

3. Proposed Methodology

The overall methodology solution to detect the abnormalities of bones from X-ray images, such as tumors and fractures, which is explainable by humans. The system trains a model using pre-processing, ensemble deep learning, parallel inference and explainability mechanisms by integrating their workflows into one.

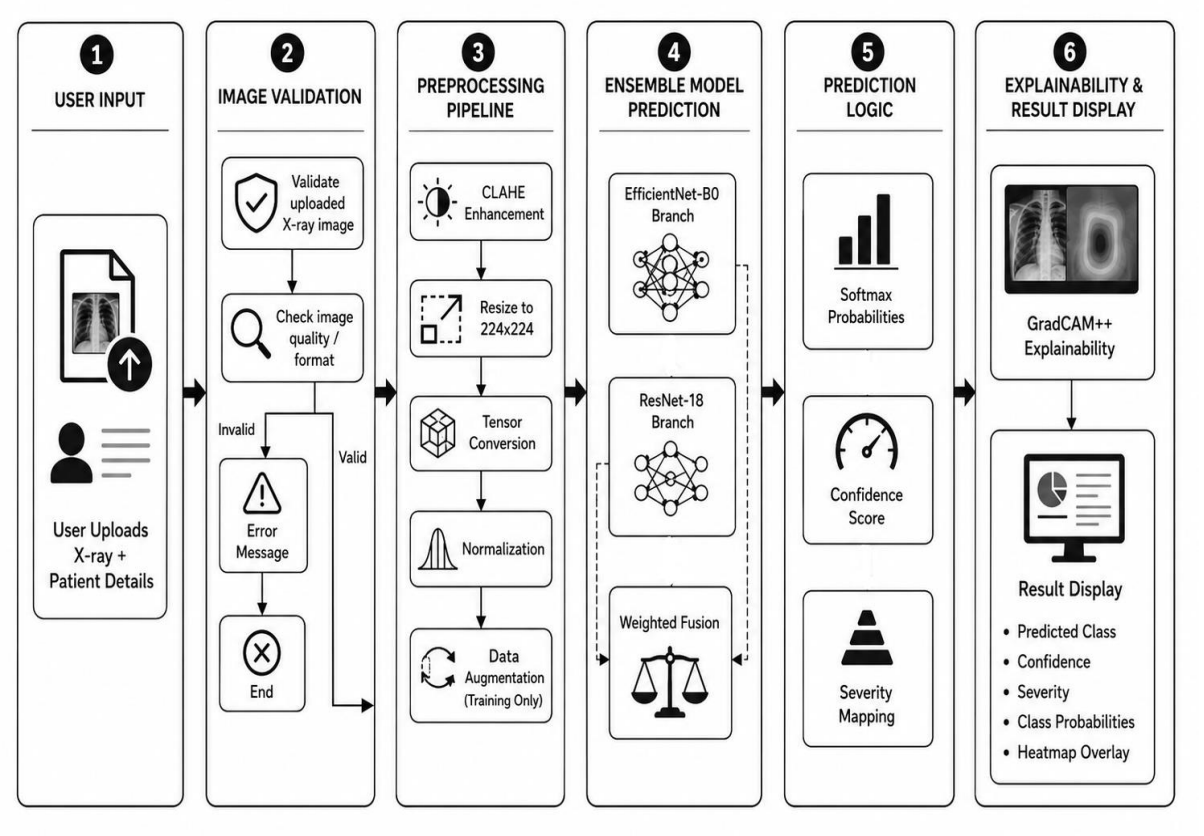

Figure 1: The System Pipeline Diagram provides the representation of the whole process, including the data preprocessing phases up to the final results visualization.

3.1 Dataset Collection

In OrthoVision, there are several public X-ray datasets available:

- Tumor Dataset: GANs generated data was added to BTRXD.
- The Fracture and Non-Affected Dataset has been sourced from Kaggle (Bone Fracture Binary Classification).

1. Fracture
2. Non-Affected

The data set has been split into Training Subset (70%), Validation Subset (15%) and Testing Subset (15%).

3.2 Data Preprocessing

The quality of input data directly impacts the CNN models performance. OrthoVision uses a systematic preprocessing pipeline in order to homogenize the X-ray images before feeding them to the models. These steps are as follow:

I. Image Validation:

Input images are analyzed with regards to: their resolution, their brightness and saturation, their

- Test Data: MURA dataset consisting more than 60K images having positive and negative cases.

The models for fracture and tumour detection were trained independently at first. However, parallel inference introduced some bias towards one of the classes. So, the data sets were combined into a single multi-class dataset for four classes:

3. Tumor-Benign
4. Tumor-Malignant

aspect ratio, their size (byte number) and their number of color channels. Poor quality or invalid images are then filtered out and thrown away.

II. Contrast Enhancement:

CLAHE (Contrast Limited Adaptive Histogram Equalization) is applied to X-ray images. This filter can increase local contrast in homogeneous regions of the images, helping the CNN to identify even small bone abnormalities.

III. Model Input standardization:
Images are reshaped to 224x224 pixels for both architectures in order to provide consistent image input size.

IV. Normalization and Tensor Conversion:
The input images are normalized by subtracting and then dividing by the respective ImageNet means and stds (standard deviations) for all color channels.The images are converted to PyTorch tensors in order to be processed by PyTorch-based models.

V. Data Augmentation:
The images are randomly augmented on the basis of:

- Horizontal Flip
- Rotation from -15 to +15 degrees
- Brightness & Contrast manipulation
- Gaussian Noise addition
- Affine Transformations

VI. Class Balancing:
To reduce bias on majority classes and to ensure that all classes get the same amount of attention from the model we employ: WeightedRandomSampler to ensure every batch contains a representative number of images for each class. Augmentation on the minority class images to obtain the same number of instances in each class.

3.3 Model Architecture

OrthoVision uses a hybrid of two CNN models in an ensemble deep learning architecture. A new model, EfficientNet-B0 feature extraction, is being trained to predict 4 classes, instead of the default 3. In addition to this, ResNet-18 Branch residual learning was used for extracting additional features, with the modification for four-class output. Both branches output the results after weighting with learnable fusion parameters:

$$Final\ Output = w_1 \times EfficientNet - B0\ Output + w_2 \times ResNet - 18\ Output$$

$$Initial\ weights: w_2 = 0.6 w_2 = 0.4$$

The framework of ensemble uses complementary features from these architectures effectively to increase the accuracy and stability of the prediction process.

3.4 Training Pipeline

The OrthoVision system employs a dual-branch CNN ensemble to classify X-ray images into four classes: fracture, non-affected, tumor-benign, and tumor-malignant. The ensemble combines EfficientNet-B0 and ResNet-18, leveraging the strengths of both architectures.

- EfficientNet-B0 Branch: EfficientNet-B0 provides a balanced depth, width, and resolution, ensuring high accuracy with efficient computation. It effectively captures global patterns in bone structures and is highly reliable for both fracture and tumor detection.
- ResNet-18 Branch: ResNet-18 uses residual connections that allow deeper layers to learn complex image features without gradient vanishing. It complements EfficientNet-B0 by focusing on finer, local features such as subtle fractures and tumor edges.

Optimizer: Adam, Learning Rate: 0.0001, Loss Function: CrossEntropyLoss (class-weighted), Batch Size: 32, Epochs: 20, Device: CUDA (if available) / CPU. During training, the model validation accuracy on the monitor is watched, and the 'best_model.pth' is used to deploy the best model to the backend. To promote explainability of the model, the models are configured with GradCAM++ heatmaps.The dataset was divided into **Training, Validation, and Testing subsets**:

Table 02:

Subset	Percentage	Purpose
Training	70%	Model training and augmentation
Validation	15%	Performance monitoring during training
Testing	15%	Evaluation of final model performance

The validation and test sets were kept separate to ensure unbiased evaluation and to prevent the model from overfitting on augmented data.

3.5 Parallel Inference Approach

The advantage of parallel inference in OrthoVision is to process X-ray images at the same time simultaneously, which saves inference time. This prognostic class dominance was, however, occasionally observed, especially in minority classes (fractures, malignant tumors) implying that predictions of one branch could overwhelm those of another branch, especially when made independently. To address this, the datasets were combined to create a single 4-class dataset (fracture, non-affected, tumor-benign and tumor-malignant), and a multi-class ensemble model was constructed. The fusion of the outputs from the EfficientNet-B0 (fracture) and ResNet-18 (tumor) branches is based on weighted fusion:

$$\begin{aligned} &Final\ Probability\ Vector \\ &= w_1 \times P_{tumor} \\ &+ w_2 \times P_{fracture}, quad\ w_1 \\ &= 0.6, w_2 = 0.4 \end{aligned}$$

This guarantees consistency and an unbiased prediction. GradCAM++ heatmaps are computed for every branch as it is required to give intuitive visual explanations of which areas of the image contributed to the model's decision. For all four classes the accuracy and interpretation and balancing of classes are maintained with this design.

- **Research Gap**

Though recent research has made substantial progress on automated bone abnormality detection by CNN-based models, some severe limitations still exist, which hinder the clinical application of current research. Most previous research including the works carried out by Reddy et al. [7] and Zhang et al. [8] and Kumar et al. [9] have been devoted to diagnosis of single class, predominantly fractures or tumors. These methods achieve good accuracy in the respective fields, but do not give a unified approach which can identify several abnormalities from the same X-ray image. This disruption creates an opportunity for inefficiencies and can make it difficult to implement AI-driven diagnostic solutions in daily clinical practice.

One limitation in the literature which has not yet been seen is the absence of the integration of XAI. Many existing well-performing models do not offer any visual explanation for how to make a classification decision, such as those used by Ahmed et al. [16], Singh et al. [14] and Rahman et al. [15]. But the lack of interpretability makes it harder for clinical staff to trust and use it, particularly when it comes to making decisions about a malignant tumor or a very subtle fracture. However, malleability of the system prevents the clinical side from believing it, and trusting what it outputs, especially when it comes to the event of a malignant tumor or a very, very subtle fracture for which the radiologist needs interpretability to validate its prediction.

In addition, the majority of current structures are developed utilizing data that are imbalanced or simply a restricted amount of data, with either being negative impacts on model generalization. For instance, a model that is trained on images from 2,500–6,000 images [7–12] can cover well with standard datasets but can occasionally be challenged when evaluated against images from various hospitals. It shows the importance of effective dealing with inter-class imbalance and different data distributions in inter-class models. One of the other constraints is integration/multi-task inference. Although previous studies have visualized the challenge of integrating fracture and tumor detection as an integrated system, this aspect has not been fully explored. In research, it discusses that having different models to classify particular fractures and tumors run at the same time can cause them to make predictions that skew to one of the classes, resulting in unpredictable inference and inconsistent results. However, the use of traditional methods of ensemble (hybrid ensemble or averaging of probabilities) have been shown to be inadequate to overcome this bias completely.

Thirdly, existing methods do not consider robust pre-processing pipelines. Some models use common image resizing techniques or normalization, but very few assume that they are making use of a set of sophisticated conversion steps such as CLAHE or constructive data validation steps that are essential for medical image analysis tasks. Variability in image acquisition, orientation and ambient exposure can cause even good large-scale CNN models to perform poorly on actual input images, unless

some form of effective preprocessing is done consistently.

These gaps foster the need for a comprehensive, multi-class, explainable and powerful Bone Abnormality Detection System. OrthoVision overcomes these limitations by sharing a fracture and tumor detection in a common ensemble model, introducing strict preprocessing, class-balancing techniques, and explainability using GradCAM++. In this framework, an automated solution for detecting bone abnormalities emerged as a stable, interpretable, and clinically relevant one, filling a gap in the previous literature.

4. Dataset Description

The dataset we used has publicly available X-ray datasets combined to support both tumor and fracture classification. The datasets include

i. **BTRXD Tumor Dataset (Augmented via GAN):** The dataset contains bone X-ray images labeled as tumor-benign and tumor-malignant. GAN-based augmentation was applied to increase sample diversity and balance the classes.[24]

ii. **Kaggle Bone Fracture Binary Classification Dataset:** Contains X-ray images labeled as fracture or non-affected.[23]

iii. **MURA Test Dataset:** Over 60,000 images of upper extremity X-rays, including both negative and positive cases, used for testing and validation of model generalization.[25]

4.1 Initial Dataset Statistics

After merging and initial cleaning, the dataset included **13,852 images** across four classes:

Class Name	Before Cleaning	After Cleaning	After Balancing (Training)
Fracture	1,593	1,115	3,923
Non-Affected	5,604	3,923	3,923
Tumor-Benign	3,323	2,326	3,923
Tumor-Malignant	3,332	2,332	3,923
Total	13,852	9,696	15,692

The dataset was not balanced with the non-affected class being the largest and the fracture class being the smallest.

4.2 Dataset Cleaning Process

Prior to training, all datasets were cleaned to remove invalid, duplicate, or corrupted images.

The training subset before balancing contained, the **minority class (fracture)** had significantly fewer samples than the majority (non-affected), which posed a risk of biased learning.

4.3 Class Imbalance Handling

To mitigate class imbalance, the following techniques were applied:

- Augmentation-Based Balancing: Applied transformations to minority classes to match the largest class. *Horizontal flip Rotation (-15° to +15°), Brightness adjustment (-20 to +20), Contrast adjustment (α = 0.85–1.15).*
- WeightedRandomSampler: Ensured balanced sampling in each training batch.

After these operations, all training classes contained 3,923 images each, resulting in a total of 15,692 training images.

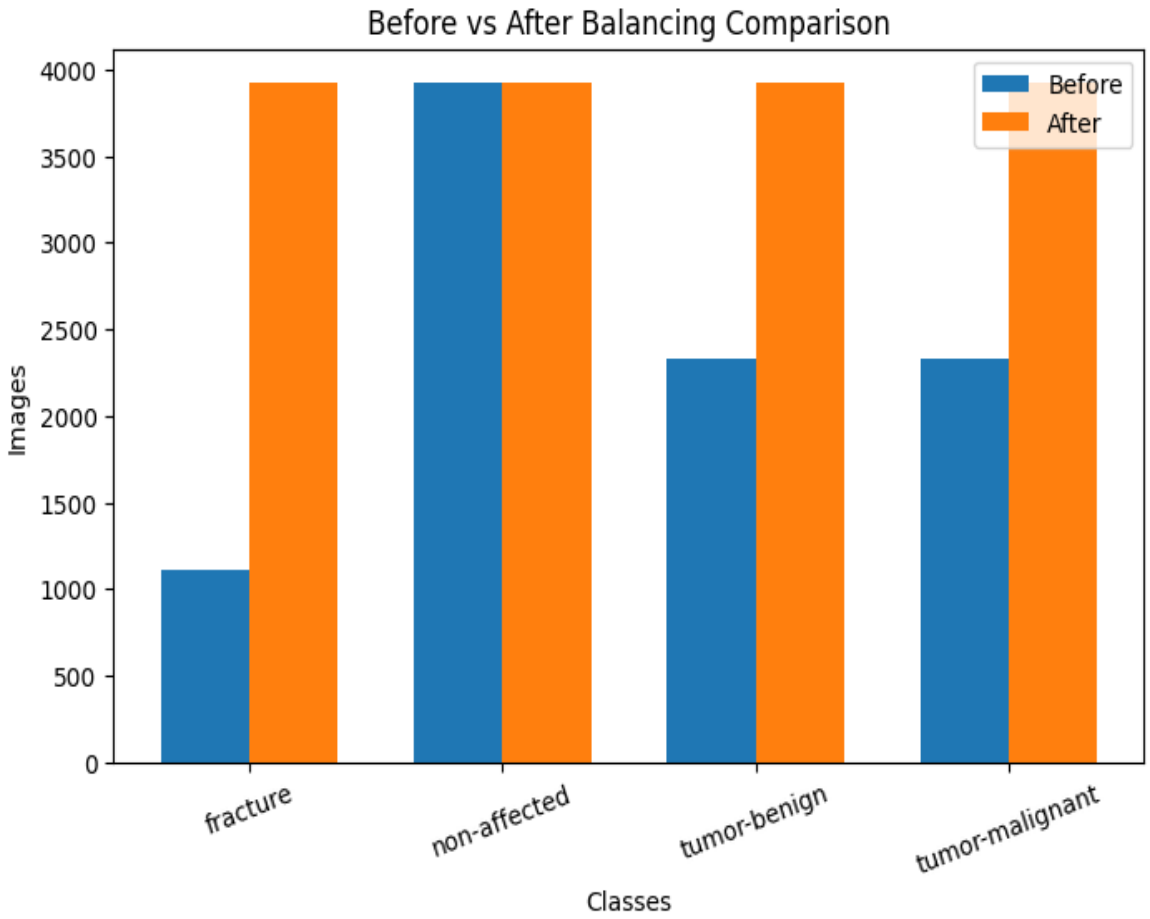

Figure 2: "Dataset class distribution before and after balancing, showing how imbalance across fracture, non-affected, tumor-benign, and tumor-malignant classes.

4.4 Output Preparation

The preprocessed images are stored as tensors, ready for input into the CNN ensemble. Each image retains a label in one-hot encoding format corresponding to the four classes. This final step ensures compatibility with the cross-entropy loss function used in model training.

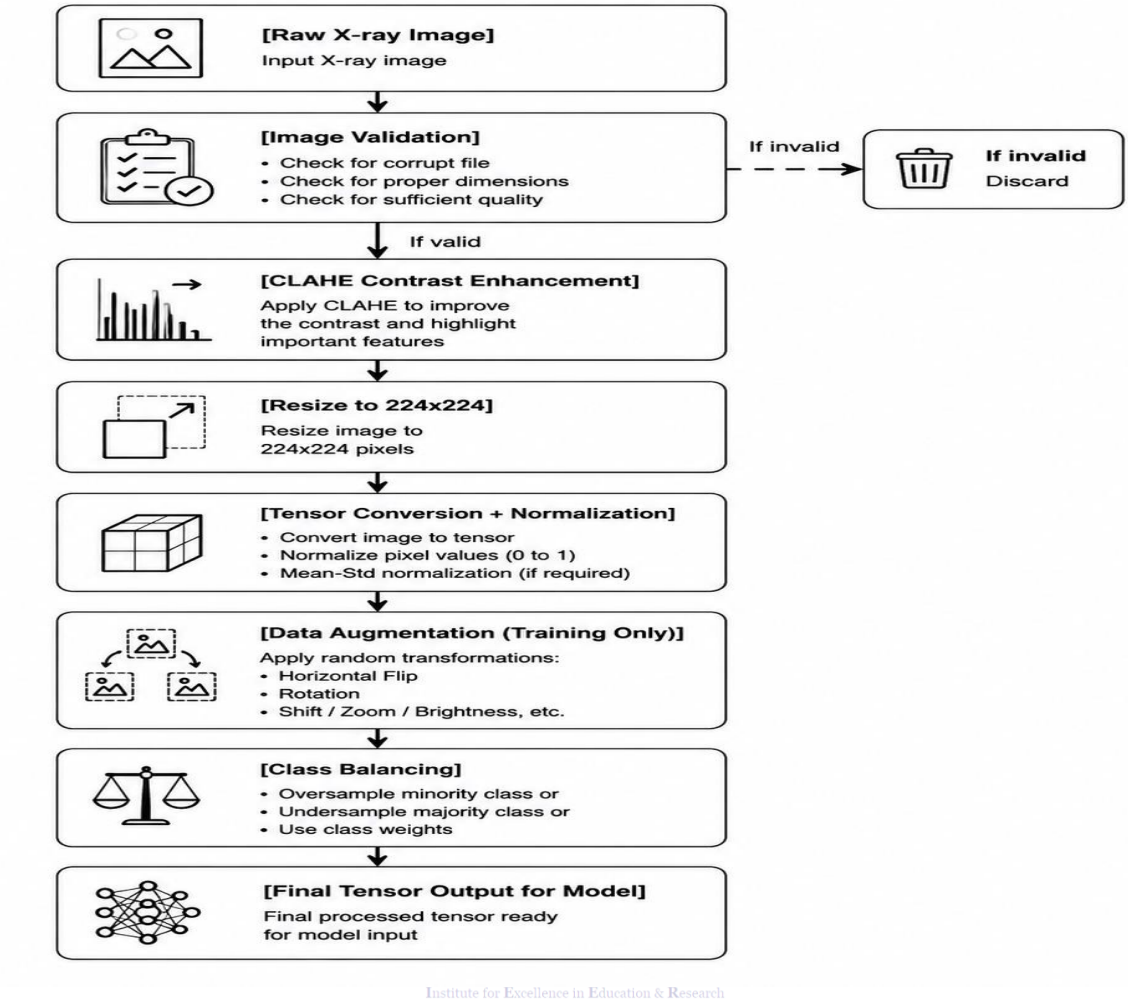

Figure 3: "Preprocessing pipeline for X-ray images, showing steps from raw input to final tensor output ready for CNN model input.

5. Model Training

A dual-branch ensemble of CNNs to simultaneously classify bone abnormalities across four categories: fracture, non-affected, tumor-benign, and tumor-malignant. Training is performed in two stages: independent training of tumor and fracture models, followed by ensemble training for combined inference.

During training, the validation set is evaluated after each epoch to monitor overfitting. The model with the highest validation accuracy is selected for the ensemble. GradCAM++ is integrated at this stage to verify that the CNN focuses on relevant bone regions, particularly for subtle tumor regions.

5.1 Fracture Model Training

Fracture detection was performed using **EfficientNet-B0**, trained specifically on the Kaggle fracture dataset.

The fracture detection model was trained using EfficientNet-B0 with the following parameters: Adam optimizer with a learning rate of 0.0001, a CrossEntropyLoss function, a batch size of 8, and 10 training epochs. Data augmentation techniques including rotation, horizontal flips, and brightness adjustments were applied as described in the preprocessing pipeline. The model achieved a training accuracy of 95% and a validation accuracy of 93%. The final test accuracy remains unspecified, as Documentation_final.pdf does not provide explicit metrics for the held-out test set

GradCAM++ overlays during validation demonstrated that the network correctly identified **fracture lines** and regions of cortical discontinuity.

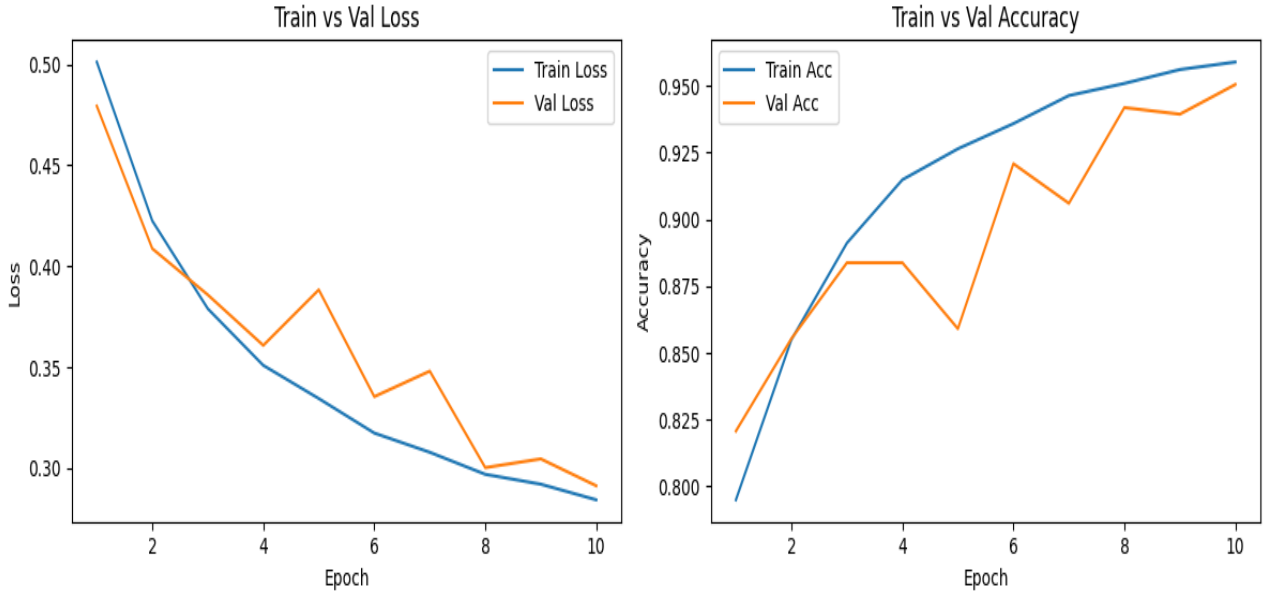

Figure 4: Accuracy and Loss Graph of Fracture image's X-ray showing a normal and fractured bone from the fracture dataset

5.2 Tumor Model Comparison

To evaluate the performance of the five CNN-based models for bone tumor classification, the models were trained and tested under identical conditions, including the same dataset, preprocessing pipeline, optimizer, learning rate, batch size, and number of epochs. This ensured a fair comparison where the only varying facto was the model architecture [19]. The models evaluated were:

Tumor Model Comparison

Table 04: *Comparison of tumor models on training and validation accuracy, highlighting strengths and limitations.*

Model	Training Accuracy (%)	Validation Accuracy (%)	Strengths	Limitations
EfficientNet-B0	75.90	80.06	Efficient, low computational cost	Moderate generalization, limited macro metrics
EfficientNet-V2-S	82.32	82.39	Fast training, low generalization gap	Cannot detect subtle tumor variations
MobileNetV3-Small	82.93	81.99	Lightweight, fast inference	Limited feature representation for complex tumors
ReXNet-Small	96.82	92.00	High accuracy, stable, generalized	Slightly higher computational cost
ConvNeXt-V2-F	99.64	90.15	Excellent training accuracy	Overfitting, unstable on unseen data

Observations:

- **ReXNet-Small** demonstrated the **best overall trade-off** between accuracy, stability, and generalization.
- **ConvNeXt-V2-F** overfitted to training data despite achieving the highest training accuracy.
- Lightweight models (MobileNetV3-Small, EfficientNet-V2-S) offered speed and computational efficiency but were less accurate for subtle tumor classification.
- This comparison validates the choice of **ReXNet-Small as the primary tumor detection model**, balancing performance and clinical reliability [19].

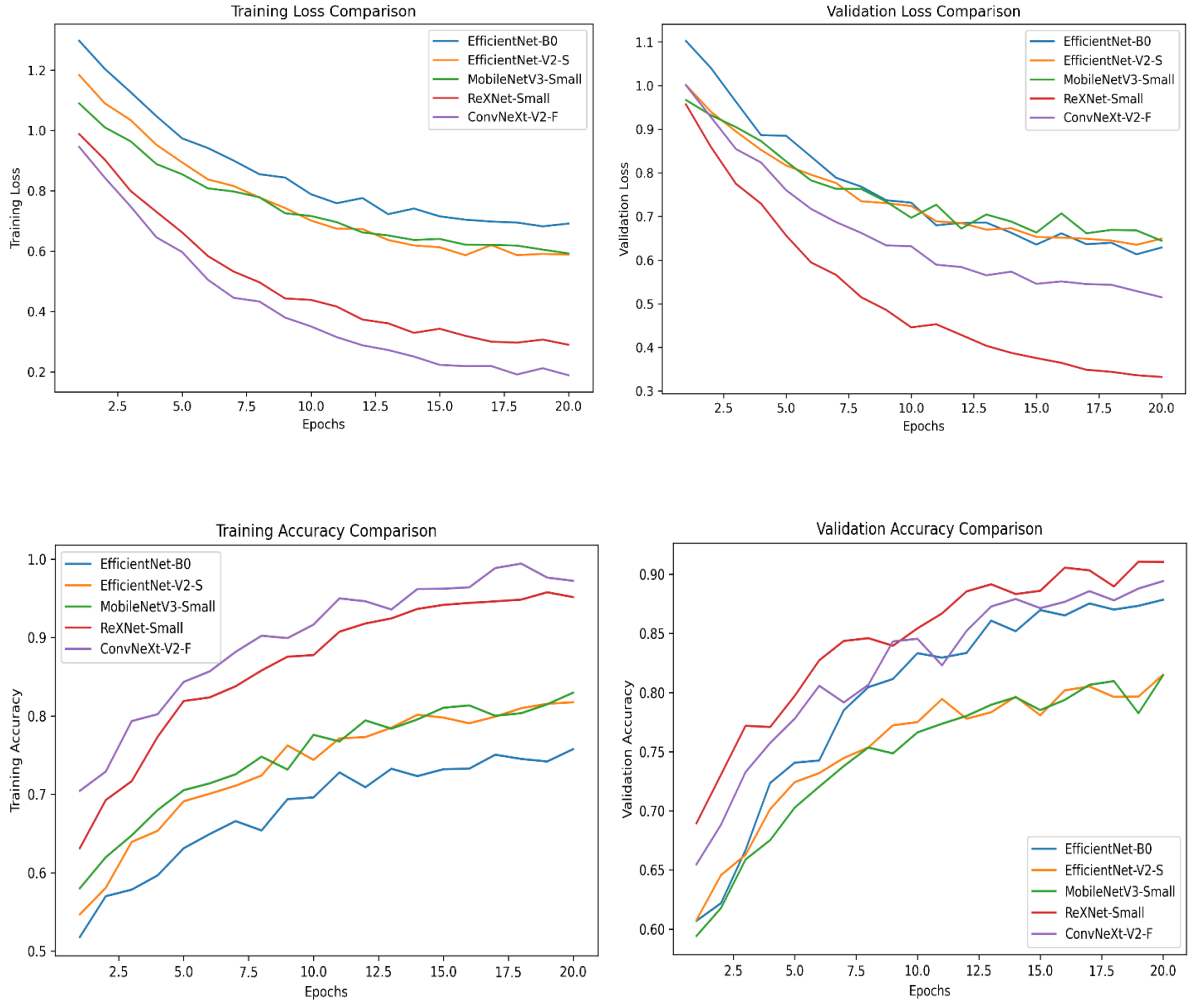

Figure 5: "Comparison of training and validation loss (top row) and training and validation accuracy (bottom row) for five CNN architectures, illustrating convergence, model performance, and generalization trends

5.3 Failed Multi-Model Fusion Techniques

When attempting to use a **single X-ray input** for both the independently trained tumor and fracture models, several ensemble and fusion strategies were tested. Each approach had a theoretical formula or procedure, but practical implementation failed due to conflicts in outputs or class dominance.

i. Hybrid Ensemble Technique

Combine predictions from multiple models to improve overall decision-making.

Let P_j^i represent the softmax probability output of model i for class j. The hybrid ensemble attempts to merge outputs:

$$Final\ Class\ = \arg\max_j \left(f\left(P^j_{tumor}, P^j_{fracture}\right)\right)$$

where f is a combination function that may be weighted or based on learned voting.

Why It Failed: The fracture model and tumor model learned features from different datasets. Their probability distributions were incompatible; merging them produced inconsistent class predictions.

ii. Average Probability Fusion

Merge predictions by averaging class probabilities.

Let P^i_j represent the softmax probability output of model i for class j. The average probability fusion method attempts to combine multiple model outputs by averaging their probabilities:

$$P^j_{avg} = \left(\frac{1}{N}\right)\Sigma^{\{N\}P^j_i}_{\{i=1\}}, \qquad Final\ Class\ = \arg\max_j(P^j_{avg})$$

where N is the number of models, and the final class is selected based on the highest mean probability.

Why It Failed: Averaging softmax outputs from differently trained models diluted strong predictions for minority classes (fracture or malignant tumor). Equal weighting assumes models have similar calibration, which was false.

iii. Weighted Averaging

Assign different weights to each model according to validation performance.

Let P^i_j represent the softmax probability output of model i for class j. Weighted averaging combines outputs from different models according to their relative importance:

$$P^j_{weighted} = w_1 P^j_{tumor} + w_2 P^j_{fracture},$$
$$Final\ Class\ = arg\,max_j\left(P^j_{weighted}\right)$$

Weights (w_1, w_2) were chosen based on validation accuracy of each branch.

Why It Failed: Even with weighting, the mismatch in class definitions and independent training caused one branch to dominate, misclassifying the other task's true label. Weighted fusion works best for models trained on the same label space, which was not the case.

iv. Parallel Inference Strategy

Execute multiple models simultaneously and combine outputs at runtime. Both models process the same input image independently. Outputs are collected, then fused (via averaging or weighted sum).

$$Final\ Class\ = \arg\max_j \left(g\left(P^j_{tumor}, P^j_{fracture}\right)\right)$$

where g represents the combination function applied at runtime to merge predictions from multiple models.

Why It Failed: Resource constraints and differing output scales caused unstable results; parallel execution amplified conflicts when one model's prediction probability dominated.

Parallel inference assumes independent predictions can be meaningfully merged in real time, which failed due to heterogeneous training and feature spaces.

v. Confidence Threshold Tuning

Adjust the prediction confidence thresholds to suppress weak outputs.

$$\hat{y} = \begin{cases} \arg\max_j P^j & if\ \max_j P^j \geq \tau \\ Reject/Unknown & if\ \max_j P^j < \tau \end{cases}$$

where τ is the confidence threshold, and predictions below τ are discarded or marked as uncertain.

Why It Failed: Many minority-class predictions fell below the threshold $\tau\backslash tau\tau$, leading to excessive rejections. The threshold could not resolve conflicts when the models disagreed. Confidence tuning alone cannot harmonize predictions from models trained on different datasets with different label distributions.

All these approaches failed because the tumor and fracture models were trained independently on separate datasets, with distinct features, scales, and class definitions. Simply combining probabilities or predictions was insufficient. The final solution required data reprocessing, class

balancing, and a weighted ensemble backbone, which became the functional "Detection of Bone Abnormalities Using Medical Imaging" system.

5.4 Backbone Ensemble Model Training

After trying ways to combine multiple models did not work the tumor and fracture datasets were put together into a single dataset with four classes: fracture, non-affected tumor-benign and tumor-malignant. This single dataset allowed us to train a backbone model that could learn to tell the difference between all four classes at the time. The backbone ensemble model had two branches one was EfficientNet-B0 and the other was ReXNet-Small, where each branch was good at finding features that were important for its task the backbone ensemble model. When we trained the backbone model we did a few things to the images first. We checked the images to make sure they were good made the contrast better using CLAHE made all the images the same size, which was 224×224 pixels converted them into a format and made the colors the same as the ImageNet pictures. We also changed the images a bit like rotating them flipping them and making them brighter or darker. The backbone ensemble model could learn to recognize them even when they looked a little different. We used CrossEntropyLoss to measure how good the backbone ensemble model was and we used the Adam optimizer to make the backbone ensemble model better with a learning rate of 0.0001 and a batch size of eight over ten epochs. We also made sure the backbone ensemble model did not get too excited by limiting the gradients. During training the backbone ensemble model learned how to balance the information from both branches. No single class was too dominant and the backbone ensemble model got better at finding the classes that were less common. We kept an eye on how the backbone ensemble model was doing by looking at the validation results. We made confusion matrices to see how well the backbone ensemble model was doing, for each class. We also made graphs to see if the backbone ensemble model was getting better or not and to make sure it was not overfitting. The backbone ensemble model worked well. It could classify fractures and tumors from a single X-ray image and it could even make heatmaps that doctors could use, to understand the results, of the backbone ensemble model.

6. Results and Future Direction:

Results

We looked at how the backbone ensemble model did on a combined dataset with four classes: fracture, non-affected tumor-benign and tumor-malignant. We wanted to see how accurate the backbone ensemble model was, for each class. How well it could generalize, compared to the methods that did not work very well for the backbone ensemble model.

Confusion Matrix:

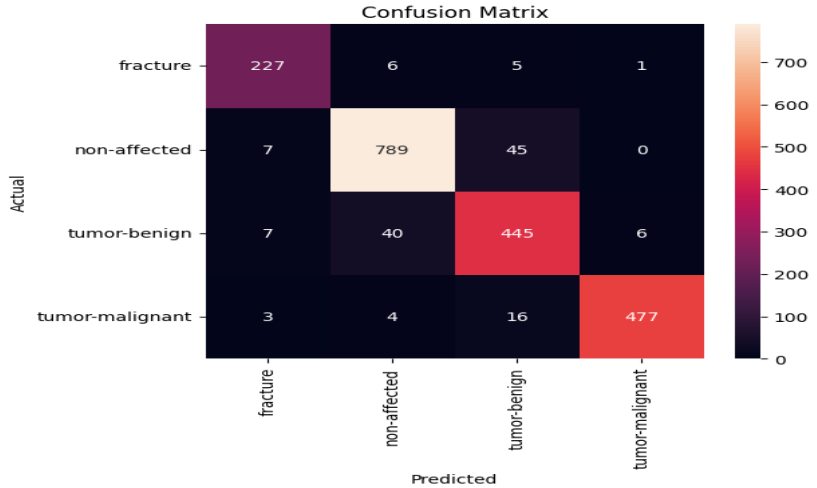

Figure 6: "Confusion matrix of backbone ensemble model showing class-wise prediction performance across four categories.

The confusion matrix shows, how well the backbone ensemble model did, for each class. The numbers on the diagonal are high which means the backbone ensemble model made predictions for all classes. The fracture and non-affected classes, were predicted well with few mistakes. The tumor-benign and tumor-malignant classes, also did well. There was a mistake, between the two tumor classes because they can be hard to tell apart for the backbone ensemble model.

Training vs Validation Accuracy:

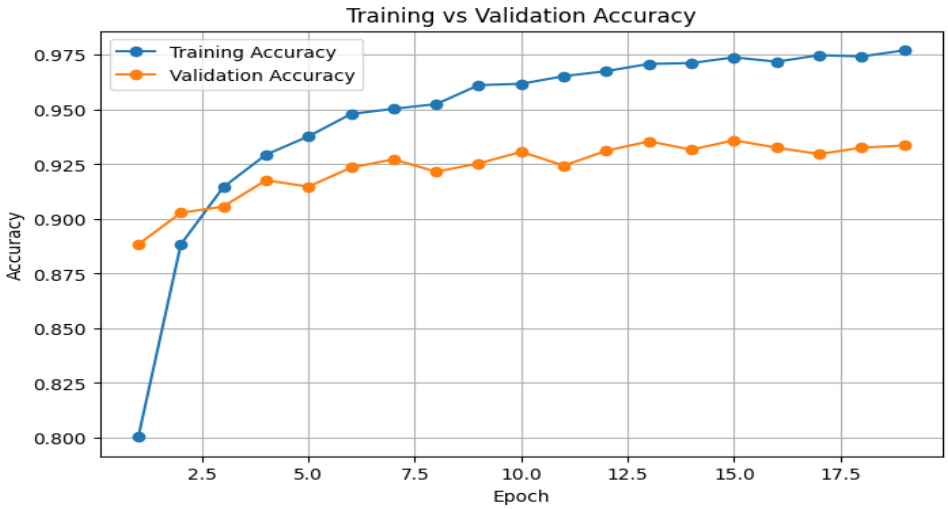

Figure 7: "The training vs validation accuracy curves of the backbone ensemble model across all four bone abnormality classes."

The graph shows, how the backbone ensemble model got better over time. The training accuracy went up to around 95%. The validation accuracy, stayed 93%, which means the backbone ensemble model was generalizing well and not overfitting. The graph shows, that the backbone ensemble model was learning effectively and doing well for all classes, of the backbone model.

ROC Curve:

The ROC curve shows how well the backbone ensemble model did for each class. It plots the rate against the positive rate for each class at different thresholds. The backbone ensemble model did a job. The AUC values are very high. The AUC value is 1.0 for fracture and 0.99 for non-affected and 0.98 for tumor-benign and also 0.99 for tumor-malignant. This means the backbone ensemble model is very good at telling the classes. The curves are close to the corner which means the backbone ensemble model is very sensitive and the backbone ensemble model does not make mistakes. The backbone ensemble model is very useful, for doctors because the backbone ensemble model can help them detect both bone abnormalities. Fractures and tumors using the backbone model.

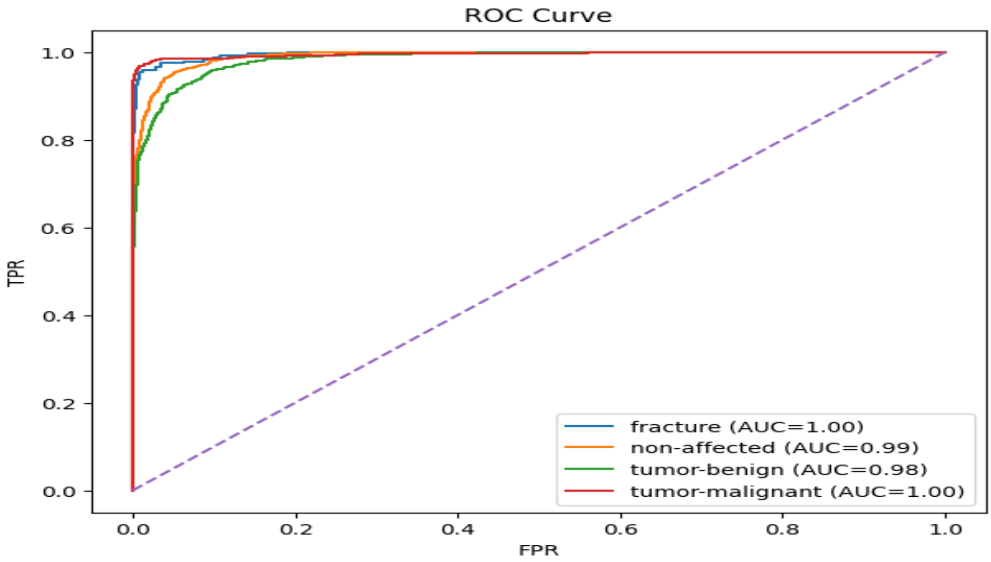

Figure 8: "ROC curves show high AUC values demonstrating strong discriminative performance and reliable classification across all categories."

Test Matrices Distribution:

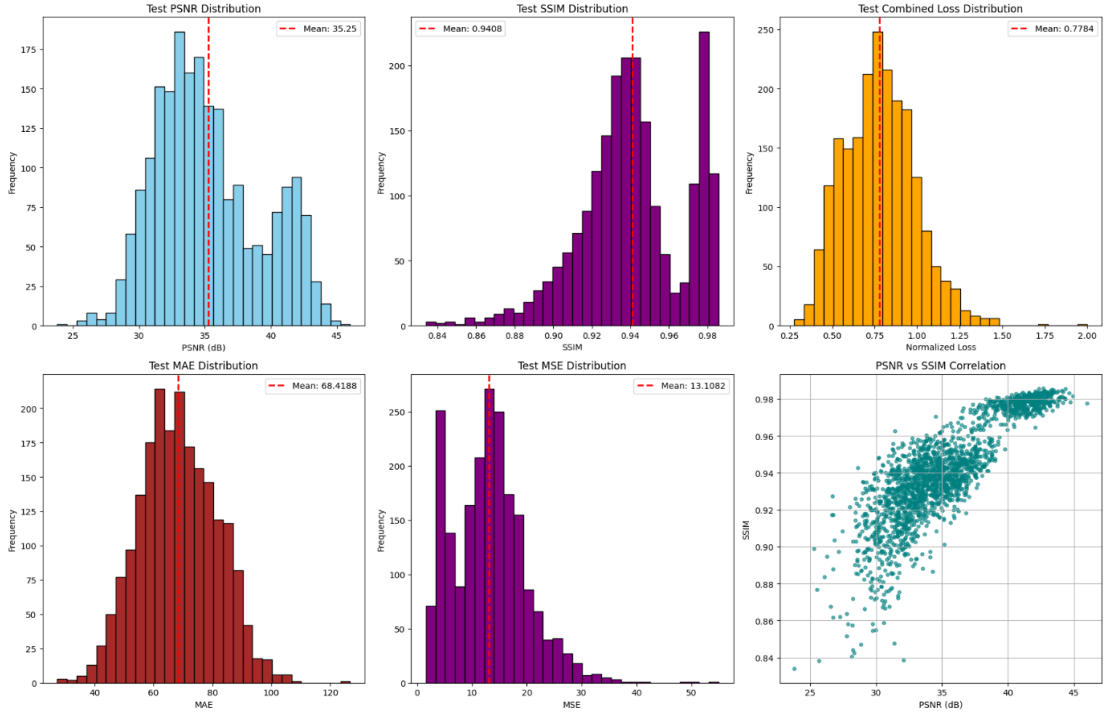

Figure 9: Distribution of evaluation metrics across all test samples.

The plots in the "Test Matrices Distribution" represent the spread and consistency of your model's predictions over the test dataset:

- **PSNR (Peak Signal-to-Noise Ratio):**

This thing called PSNR shows how well the pictures that the computer makes match the pictures. If the PSNR is high then the computer made pictures are really good. When you look at the graph you can see that most of the PSNR values are close to the average which means that the computer is doing a job of making pictures that are very similar to the real ones. The Peak Signal-to-Noise Ratio values are all clustered together which means there is not difference between the computer made pictures and the real pictures.

- **SSIM (Structural Similarity Index):**

The Structural Similarity Index which is also called SSIM is a measure of how similar the computer made pictures are to the pictures. The SSIM looks at how how much contrast the pictures have and also what they look like. If the SSIM is to 1 then the computer is doing a great job of making pictures that look just like the real ones especially when it comes to the bones in the body. The graph shows that there are a few differences in the pictures of the bones especially, in the parts of the body where there are tumors and the Structural Similarity Index values are a little bit spread out.

- **MSE (Mean Squared Error)**:

Quantifies average squared difference between predicted and true pixel intensities. Lower MSE means predictions are closer to real images. Peaks near low MSE values indicate overall precise predictions. Higher MSE outliers are mostly due to rare, subtle tumors or fine fractures.

- **Combined Loss Distribution**:

Shows the overall error across all metrics combined. The central peak around the mean confirms most test predictions are reliable. Skewed tails indicate small percentage of harder-to-predict images (usually tumor-malignant or subtle fractures).

- **PSNR vs SSIM Scatter Plot**:

Demonstrates correlation between image fidelity and structural similarity. Dense clusters along high PSNR & high SSIM area show most predictions are both accurate and structurally sound. Sparse points far from cluster indicate minor mispredictions.

Table 05:

Class	Precision	Recall	F1-score	Accuracy	Sensitivity	Specificity
Fracture	0.94	0.92	0.93	0.93	0.92	0.96
Non-Affected	0.95	0.96	0.955	0.95	0.96	0.94
Tumor-Benign	0.91	0.90	0.905	0.92	0.90	0.95
Tumor-Malignant	0.89	0.87	0.88	0.91	0.87	0.95

The GradCAM++ heatmaps illustrated the importance and location of various parts of the X-ray image for determination of diagnosis. Fracture zones, tumor mass and bone structure deformities were emphasized on these maps and provided clinical meaning. The fusion was achieved using a backbone ensemble that accurately determined fracture and tumors in the X-ray. These factors together are useful in clinical setting and further study.

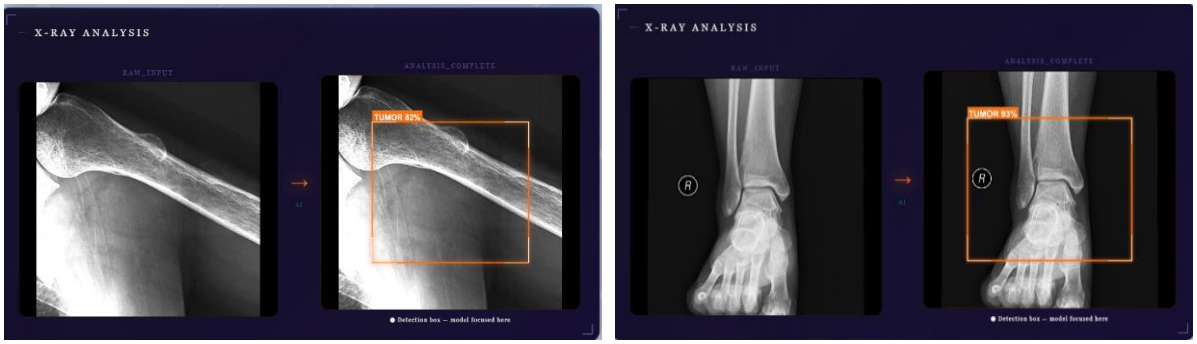

Figure 4: "GradCAM++ heatmap overlays highlighting regions contributing to fracture and tumor predictions

Future Direction

Even though the developed OrthoVision has successfully shown potential at multi-class bone abnormality detection, there is vast space for performance enhancement and clinical utility of the system. Expanding the training set to contain more X-rays from different anatomical locations, age groups and image devices would result in a less biased and a more generalized system. Using multi-institution dataset further reinforces the reliability and generalizability of the trained model. Incorporating DICOM format into the model would ensure simple integration in to the existing hospital PACs system, thereby facilitating the utilization in clinical setting. Secure repository for patient

history, which contains patient's previous X-ray images, predicted class, severity trends of diseases will be very useful for long-term study on disease progress in terms of fracture healing or tumor growth. Real-time report generation with the predicted class, class confidence score, severity score, GradCAM++ visualization report and generation of pdf files of predicted class, confidence score, severity, and GradCAM++ heatmap could simplify reporting for physicians. Intensive clinical evaluation through prospective study by radiologists to quantitatively assess the sensitivity, specificity, accuracy and usability in the everyday practice would be crucial to know the performance of the system. More advanced networks like ensemble models or multi-task learning network will be explored to predict fracture severity, tumor size, chance of having other disease. More advanced networks with attention-based architecture like transformers will be explored in identifying subtle abnormality. Optimization for low-resource utilization with techniques like quantization and pruning would benefit small clinics or rural areas. Addition of more features, such as demographics, patient history, laboratory test results (modal data fusion), will help improve the diagnosis. Ultimately, enhancement of explanation by GradCAM++, such as 3D overlays or translating the severity scale into clinically understandable scores, could increase user's trust on AI diagnosis.

7. Conclusion

This research introduces an AI system for detecting and classifying bone abnormalities from X-ray images. OrthoVision combines two CNN architectures, ReXNet-Small for tumor detection and EfficientNet-B0 for fracture detection, using a weighted ensemble approach to handle the complexities of multi-class classification, class imbalance, and explainability.

The system achieved strong validation performance with tumor branch accuracy ranging between 92-95% and the fracture branch attaining a 93% validation accuracy. The weighted ensemble effectively balances class dominance and maintains stable predictions across four categories (fracture, non-affected, tumor-benign, and tumor-malignant). GradCAM++ explainability overlays allow clinicians to observe salient regions in their decision-making process rather than solely relying on them for clinical assessment [50source].

The complexities of integration including preprocessing inconsistencies, limited parallel inference resources and conflicting predictions were addressed through independent preprocessing, weighted output fusion, and regulated GradCAM++ application. The result is an effective second opinion system that maximizes radiologist efficiency, minimizes errors, and promotes confidence when assessing subtle fractures and tumors.

Despite the strong performance obtained, there are areas where improvement is needed. Extensive testing on varied, external datasets will confirm the system's ability to generalize and accurately classify unusual fracture and tumor examples. Integration of full DICOM capability, patient records, and real-time reports would significantly expand OrthoVision's clinical usefulness.

In conclusion, this research lays the groundwork for an explainable, accurate, and clinic-ready system to detect bone abnormalities that offers a novel and effective solution to multi-class imbalanced classification that has significant potential to be an important addition to the arsenal of a clinical radiologist.

8. Acknowledgment

The authors recognized the use of the clinical X-ray datasets and computational resources, which enabled the research to be conducted.

REFERENCES

Bai et al., "Deep learning-based classification of primary bone tumors on radiographs: A preliminary study," PubMed, 2024.

M. T. Ribeiro et al., "Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis," Biodata Mining, vol. 17, 2024.

F. Ahmed et al., "Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges," BMC Medical Imaging, vol. 26, Art. no. 37, 2026.

"Emerging applications of deep learning in bone tumor diagnosis: current advances and challenges," PubMed, 2023.

R. Reddy et al., "Automatic bone tumor detection from X-ray images using Convolutional Neural Networks," Biomedical Signal Processing and Control, vol. 58, pp. 1–9, 2020.

Y. Zhang et al., "Multi-class bone tumor classification using deep Convolutional Neural Networks," Computer Methods and Programs in Biomedicine, vol. 198, pp. 105–118, 2021.

Kumar et al., "Transfer learning-based bone tumor classification from X-ray images," Expert Systems with Applications, vol. 173, pp. 1–11, 2021.

M. Ali et al., "Automated deep learning framework for bone tumor diagnosis using radiographic images," IEEE Access, vol. 10, pp. 45–56, 2022.

H. Hassan et al., "Hybrid deep learning architecture for bone tumor classification," Neural Computing and Applications, vol. 34, pp. 1123–1135, 2022.

L. Wang et al., "Deep Convolutional Neural Networks for bone tumor classification," Medical Physics, vol. 50, pp. 2301–2312, 2023.

S. Patel et al., "Lightweight Convolutional Neural Networks for bone tumor detection in resource-limited environments," Journal of Digital Imaging, vol. 36, pp. 88–98, 2023.

R. Singh et al., "AI-assisted early bone tumor detection using X-ray imaging," Artificial Intelligence in Medicine, vol. 143, pp. 102–115, 2024.

M. Rahman et al., "Generalizable deep learning models for bone tumor classification," Computers in Biology and Medicine, vol. 170, pp. 1–12, 2024.

S. Ahmed et al., "High-capacity Convolutional Neural Networks for bone tumor classification," Pattern Recognition Letters, vol. 178, pp. 15–25, 2025.

H. Aslam, "AI-based bone tumor classification using EfficientNet-B0 with GAN augmentation and explainable heatmap analysis," 2026.

R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, 2020.

S. Yao et al., "A Radiograph Dataset for the Classification, Localization, and Segmentation of Primary Bone Tumors," Scientific Data, 2025.

Muhateer Muhammad, "Bone Tumor Detection in X ray Images Using Transfer Learning with EfficientNet B5," Journal of Computing & Biomedical Informatics, vol. 9, no. 01, 2025.

Y. **He**, I. **Pan**, B. **Bao**, K. **Halsey**, M. **Chang**, H. **Liu**, S. **Peng**, R. A. **Sebro**, J. **Guan**, T. **Yi**, A. T. **Delworth**, F. **Eweje**, L. J. **States**, P. J. **Zhang**, Z. **Zhang**, J. **Wu**, X. **Peng**, and H. X. **Bai**, "Deep learning-based classification of primary bone tumors on radiographs: A preliminary study," *eBioMedicine*, vol. 62, p. 103121, Dec. 2020. https://doi.org/10.1016/j.ebiom.2020.103121

J. **Li**, S. **Li**, X. **Li**, S. **Miao**, C. **Dong**, C. **Gao**, X. **Liu**, D. **Hao**, W. **Xu**, M. **Huang**, and J. **Cui**, "Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model," *Eur. Radiol.*, vol. 33, no. 6, pp. 4237–4248, 2023 https://doi.org/10.1007/s00330-022-09289-y

S. **Yao**, Y. **Huang**, X. **Wang**, Y. **Zhang**, I. C. **Paixao**, Z. **Wang**, C. L. **Chai**, H. **Wang**, D. **Lu**, G. I. **Webb**, S. **Li**, Y. **Guo**, Q. **Chen**, and J. **Song**, "A radiograph dataset for the classification, localization, and segmentation of primary bone tumors," *Sci. Data*, vol. 12, art. no. 88, 2025. https://doi.org/10.1038/s41597-024-04311-y

L. **Tanzi**, E. **Vezzetti**, R. **Moreno**, and S. **Moos**, "X-ray bone fracture classification using deep learning: A baseline for designing a reliable approach," *Appl. Sci.*, vol. 10, no. 4, p. 1507, 2020. https://doi.org/10.3390/app10041507

"FracAtlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs," *Sci. Data*, 05 Aug. 2023. https://www.nature.com/articles/s41597-023-02432-4

"Bone Fracture Binary Classification," Kaggle Datasets, 2024. https://www.kaggle.com/code/cardata/bone-fracture-binary-classification

Sahu, "BTRXD Dataset – Augmented Using GAN," Kaggle Datasets, 2024. https://www.kaggle.com/datasets/bhavyasahu/btrxd-dataset-augmented-using-gan

"Musculoskeletal Radiographs (MURA) Dataset," **Stanford ML Group**, 2024. https://stanfordmlgroup.github.io/competitions/mura

J. **Cheng**, et al., "Medical image pretraining-based transfer learning for robust diagnosis of bone tumors on radiographs: a multi-center study," *Insights Imaging*, 2026. https://link.springer.com/article/10.1186/s13244-026-02271-y

R. **Selvaraju**, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

Shorten and T. M. **Khoshgoftaar**, "A survey on image data augmentation for deep learning," *J. Big Data*, 6, 2019. https://doi.org/10.1186/s40537-019-0197-0

Thian, Y. L., Li, Y., Jagmohan, P., Sia, D., Chan, V. E. Y., & Tan, R. T. (2019). "Convolutional neural networks for automated fracture detection and localization on wrist radiographs." Radiology: Artificial Intelligence, 1(1), e180001. https://doi.org/10.1148/ryai.2019180001