JIRBDAI Journal of Institutional Research, Big Data Analytics and Innovation

An Explainable ResNet-101 Framework for Hand Fracture Classification in X-ray Images

Ayodeji Jubril Alabi*[1], Shakirat Ronke Yusuff[1], Sulaiman Olaniyi Abdulsalam[1], Ismail Idowu Akuji[2], Oyindamola Eniola Ajiboye[3]

[1]Department of Computer Science, Kwara State University, Malete, Nigeria
[2]Department of Computer Science, Abdulrasaq Abubakar Toyin University, Ganmo, Nigeria
[3]Department of Electrical Engineering and Computer Science, texas A&M University - Kingsville
*Corresponding author: jubriltechguy@gmail.com

ABSTRACT: Fracture detection in radiographic images continues to be a difficult task due to subtle bone disruptions and variations in anatomical structures, especially within the hand region. This study proposes an explainable deep learning framework for automated hand fracture detection using a ResNet-101 architecture. The model was trained and evaluated on the hand subset of the publicly available MURA dataset, with preprocessing methods and data balancing strategies applied to mitigate class imbalance and enhance model generalization. Bayesian hyperparameter optimization was utilized to determine the most suitable training parameters, including learning rate, dropout rate, and network capacity, thereby improving model stability and overall performance. The optimized model achieved an accuracy of 79.7% with an area under the curve (AUC) of 0.75, indicating dependable classification performance. To enhance interpretability, Grad-CAM++ was incorporated to visualize the regions of interest that influence the model's predictions, allowing verification of clinically meaningful focus areas. The results demonstrate that the model effectively detects fracture regions while maintaining stable performance across diverse radiographic samples. In addition, the system produces automated PDF-based clinical reports that combine predictions with corresponding heatmaps, thereby improving its practical relevance in clinical environments. Overall, the proposed approach offers a robust, interpretable, and clinically valuable solution for hand fracture detection, with strong potential for integration into computer-aided diagnosis systems. Future work will focus on extending the model to multi-centre datasets and incorporating additional clinical features to further enhance generalization and diagnostic reliability.

KEYWORDS: Hand Fracture, Deep Learning, ResNet-101, Grad-CAM++, Medical Imaging, Explainable AI, MURA Dataset

I. INTRODUCTION

In recent years, bone fractures have emerged as a significant public health issue, contributing greatly to increased morbidity and a decline in quality of life. Among these, hand fractures are particularly prevalent and challenging to diagnose due to the complex joint structures and overlapping bones observed in radiographic images [1], [2]. Failure to detect fractures early or accurately may result in inappropriate treatment and long-term complications, emphasizing the importance of reliable and efficient diagnostic systems [3]. Conventionally, fracture detection relies on radiographic assessment conducted by medical professionals. While this method is widely used and effective, it is often influenced by factors such as inter-observer variability, fatigue, and heavy workload, which can lead to inconsistent diagnoses, particularly in cases involving subtle fractures [4], [5]. Furthermore, manual evaluation can be time-intensive and may not consistently yield accurate outcomes.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

In recent times, deep learning has gained prominence as a powerful approach in medical image analysis, allowing for automatic feature extraction and enhanced classification accuracy. Convolutional Neural Networks (CNNs), especially ResNet-based architectures, have shown strong performance across various medical imaging applications [6], [7]. However, many existing methods primarily focus on general abnormality detection and often lack interpretability, which limits their adoption in clinical practice. In response to these limitations, this study introduces an explainable deep learning framework for hand fracture detection based on the ResNet-101 architecture. The model is trained using the hand subset of the MURA dataset, with data balancing techniques implemented to address class imbalance. Bayesian hyperparameter optimization is applied to enhance model performance, while Grad-CAM++ is incorporated to provide visual interpretations of the model's predictions. Additionally, the system generates automated PDF-based clinical reports that integrate predictions with corresponding heatmaps, thereby improving its practical applicability in clinical environments. To address these limitations, this study proposes an explainable deep learning framework for hand fracture detection based on the ResNet-101 architecture. The approach integrates data balancing strategies to handle class imbalance, Bayesian hyperparameter optimization to improve model performance, and Grad-CAM++ to provide visual explanations of predictions. The goal is to develop a system that achieves both reliable classification performance and meaningful interpretability, thereby supporting its use as an assistive tool in clinical decision-making.

II. RELATED WORKS

Recent developments in deep learning have led to significant improvements in medical image analysis, particularly in fracture detection tasks. Existing research can generally be grouped into three categories: CNN-based methods, ResNet-based architectures, and explainable artificial intelligence (XAI) approaches.

A. CNN-Based Approaches

Convolutional Neural Networks (CNNs) have been extensively used in medical image classification due to their capability to automatically learn hierarchical feature representations. Previous studies such as [8] and [9] demonstrated the effectiveness of CNN models in disease detection and automated diagnosis using medical imaging datasets. Similarly, [10] investigated multi-disease classification using deep learning techniques, while [11] showed that deep learning approaches consistently outperform traditional machine learning methods in medical imaging applications. In fracture detection, [12] applied CNN-based models to identify bone fractures and highlighted the importance of robust feature extraction for improved classification accuracy. In addition, [3] emphasized the superiority of deep learning techniques over conventional image processing methods in radiographic analysis. Other research efforts have focused on lightweight CNN architectures to reduce computational complexity while maintaining competitive performance [15], [16].

B. ResNet-Based Approaches

Residual networks (ResNet) have become widely adopted in medical imaging due to their ability to mitigate vanishing gradient issues and support the training of deeper models. Studies such as [14] reported that ResNet-based architectures deliver reliable performance in X-ray image classification tasks. More recent work has explored deeper residual models and transfer learning strategies to further enhance fracture detection and improve generalization across

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

datasets. These approaches leverage pre-trained networks to capture complex structural patterns in radiographic images.

C. Explainable AI (XAI) in Medical Imaging

Explainability has emerged as a crucial requirement in medical AI systems to ensure transparency and foster clinical trust. Study [17] emphasized that interpretable models can improve clinician confidence and support real-world adoption. Techniques such as Grad-CAM [19] are commonly used to visualize image regions that contribute to model predictions. Furthermore, recent work [18] demonstrated that integrating explainability methods enhances the transparency and reliability of deep learning systems in clinical environments.

D. Recent Advances

Recent studies have continued to advance fracture detection using various deep learning architectures and imaging modalities. [24] applied a VGG16-based transfer learning model for fracture classification and reported strong performance on a Kaggle X-ray dataset. Similarly, Venkatesam et al. [27] employed an InceptionV3 architecture on CT scan images, achieving improved feature extraction and classification performance compared to traditional CNN models. Additionally, emerging research trends such as privacy-preserving learning have gained attention in medical imaging. Studies by [25] and [26] explored federated learning approaches, enabling collaborative model training across institutions without sharing sensitive patient data. These developments highlight an increasing focus on scalability, generalization, and data privacy in medical AI systems. However, variations in datasets, imaging modalities, and experimental protocols make direct comparisons across studies difficult. These differences underscore the need for standardized evaluation frameworks and task-specific models, particularly for hand fracture detection in radiographic images.

E. Research Gaps

Despite these advancements, several challenges remain. First, many studies focus on general abnormality detection rather than specific fracture classification within distinct anatomical regions such as the hand. Second, although deep learning models often achieve high accuracy, many lack sufficient interpretability, limiting their applicability in clinical settings. Third, most studies rely on single experimental setups without rigorous validation techniques, which may not adequately reflect real-world variability. Moreover, limited research has integrated data balancing techniques, hyperparameter optimization, and explainability within a unified framework. These gaps highlight the need for models that are not only accurate but also interpretable and robust for practical clinical use.

III. METHODOLOGY

A. Proposed System Architecture

This study presents an explainable deep learning framework for detecting hand fractures from X-ray images. The system is developed to automatically categorize input images as either fractured or normal, while also providing visual explanations to aid clinical interpretation. The proposed framework comprises the following components:

(a). A pre-processed dataset involving image resizing, normalization, and class balancing techniques.
(b). A deep learning classification model built on the ResNet-101 architecture.
(c). A hyperparameter tuning module based on Bayesian optimization.
(d). An explainability component implemented using Grad-CAM++.

The workflow (as illustrated in Figure 1) starts with an input hand X-ray image, which is passed through the trained model to produce a prediction. Subsequently, the Grad-CAM++ method is

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

applied to identify and highlight the regions of the image that most influenced the prediction. Lastly, the outputs are organized into a structured clinical report in PDF format.

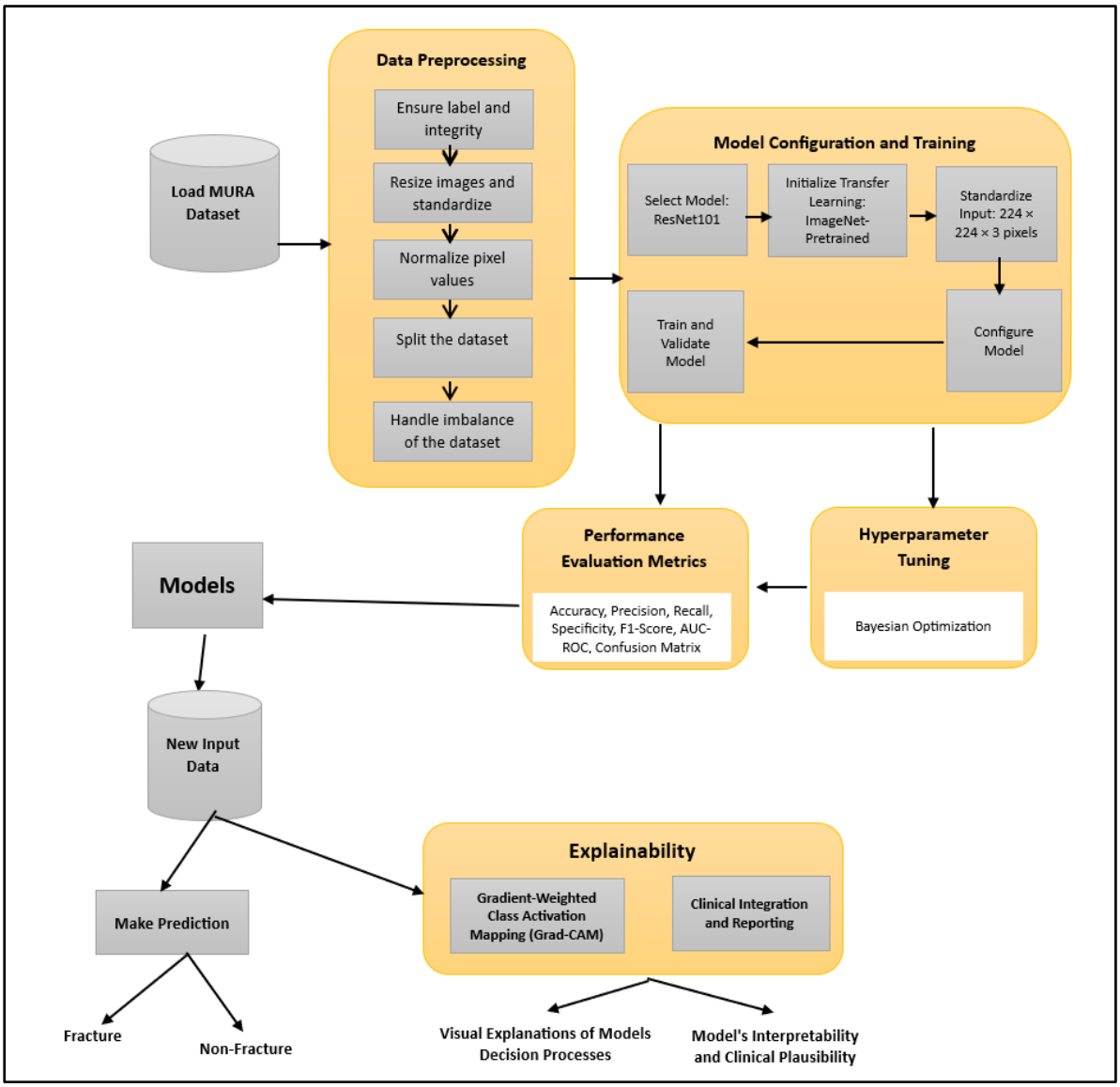

Figure 1: Proposed system architecture for hand fracture detection

B. Dataset Description and Exploration

The dataset utilized in this study consists of hand X-ray images obtained from a publicly available medical imaging dataset (MURA dataset). The images are categorized into two classes: normal and fractured, representing a binary classification problem. Hand radiographs were specifically selected to focus on fracture detection within a single anatomical region. The dataset includes variations in image orientation, contrast, and fracture patterns, making it suitable for evaluating the performance of deep learning models under real-world conditions.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

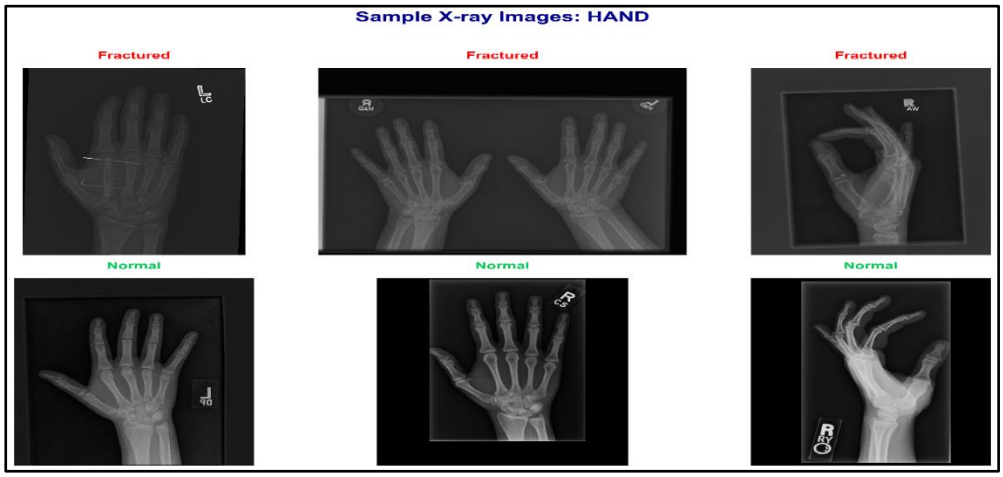

Figure 2: Sample hand X-ray images (Normal and Fractured cases)

C. Data Preprocessing

Prior to model training, the hand X-ray dataset was divided into three subsets to facilitate effective model development and evaluation. The dataset was partitioned into 72% for training, 18% for validation, and 10% for testing, ensuring adequate data for learning while maintaining separate sets for validation and unbiased performance evaluation. Following the data split, preprocessing steps (Table 1) were carried out to standardize the images and enhance learning efficiency. All images were resized to 224 × 224 pixels and normalized using the standard preprocessing function associated with the ResNet architecture.

Table 1: Image Preprocessing Techniques

Step	Description
Dataset split	72% training, 18% validation, 10% testing
Image resizing	Resized to 224 × 224 pixels
Normalization	ResNet preprocessing(min-max)
Data balancing	Applied to training set only
Augmentation	Rotation, flipping, contrast adjustment

The dataset was originally imbalanced, with a higher number of normal cases compared to fractured cases. To address this, a balancing strategy was applied to the training data by increasing the representation of the fractured class using augmentation techniques, as shown in Table 2. This approach helps the model learn fracture patterns more effectively and reduces bias toward the majority class.

Table 2: hand Dataset Distribution

Class	Raw Dataset	Balanced Dataset
Normal	3117	3117
Fractured	1205	3117
Total	4322	6234

D. Model Architecture and Training

The hand fracture detection model is built on the ResNet-101 architecture, which was utilized as a feature extractor due to its capability to learn deep and complex representations from medical images. The convolutional backbone extracts hierarchical features from hand X-ray

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

images, while a customized classification head was incorporated to tailor the model to the binary classification task. This head includes a Global Average Pooling layer, followed by a Dropout layer (0.30) for regularization, and a Dense layer with 512 units for feature representation learning. The final output layer employs a SoftMax activation function to categorize images into fractured or normal classes (Table 3).

Table 3: Model Architecture Summary

Layer	Description
ResNet101 Backbone	Deep feature extraction
Global Average Pooling	Feature reduction
Dropout (0.30)	Regularization
Dense Layer (512 units)	Feature learning
Output Layer	Softmax (2 classes)

The model was trained using a supervised learning strategy with the categorical cross-entropy loss function and the Adam optimizer to enable efficient weight updates. Training was conducted with a batch size of 32 for 30 epochs, allowing the model to learn meaningful patterns from hand X-ray images while maintaining computational efficiency. To minimize overfitting and enhance generalization, dropout regularization was incorporated within the network. The training process was monitored using the validation set to ensure stable convergence and optimal performance.

E. Hyperparameter Optimization

To further improve model performance, Bayesian hyperparameter optimization was employed to determine the most effective training configuration. This method systematically evaluates different combinations of parameters and selects the configuration that produces the best results. In this study, important parameters such as dropout rate, number of units in the dense layer, and learning rate were optimized across multiple trials. After 10 trials, the optimal configuration was identified, resulting in enhanced training stability and improved overall classification performance for hand fracture detection (Table 4).

Table 4: Optimal Hyperparameters for the Proposed Model

Parameter	Best Value
Backbone Model	ResNet101
Dropout Rate	0.30
Dense Units	512
Learning Rate	5.52×10^{-4}

(F) Explainable Artificial Intelligence (Grad-CAM++)

To improve interpretability and enhance clinical trust, Grad-CAM++ was incorporated into the hand fracture detection model. This method produces visual explanations by highlighting the regions of the X-ray image that have the greatest influence on the model's prediction. Unlike traditional deep learning models that function as black boxes, Grad-CAM++ offers insight into the decision-making process by generating heatmaps that indicate important areas. These heatmaps are superimposed on the original X-ray images, enabling verification that the model is focusing on the fractured hand region. This approach not only clarifies the reasoning behind the model's predictions but also assists in identifying possible errors, thereby increasing reliability and confidence in clinical applications.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

(G) Performance Evaluation Metrics

The performance of the hand fracture detection model was assessed using standard classification metrics, including accuracy, recall (sensitivity), specificity, and Area Under the Curve (AUC). Accuracy reflects the overall correctness of the model's predictions, while recall measures its ability to correctly identify fractured cases. Specificity evaluates how effectively the model detects normal cases, and AUC provides a comprehensive measure of performance across different classification thresholds. These metrics offer a well-rounded evaluation of the model, especially in medical diagnosis where minimizing both false positives and false negatives is critical. A summary of these metrics and their descriptions is provided in Table 5.

Table 5: Summary of Evaluation Metrics Description

Metric	Description
Accuracy	Overall prediction correctness
Recall	Ability to detect fractured cases
Specificity	Ability to detect normal cases
AUC	Overall performance across thresholds

IV. RESULTS AND DISCUSSION

This section presents the results obtained from the data exploration, preprocessing, model training and evaluation, hyperparameter tuning, and explainability stages.

A. Class Distribution Analysis

An analysis of the raw dataset showed a clear imbalance between normal and fractured hand images. As illustrated in Figure 3, normal cases were more frequent than fractured cases, with the imbalance being most noticeable in the hand dataset. This observation indicated the need for corrective preprocessing steps before training the model.

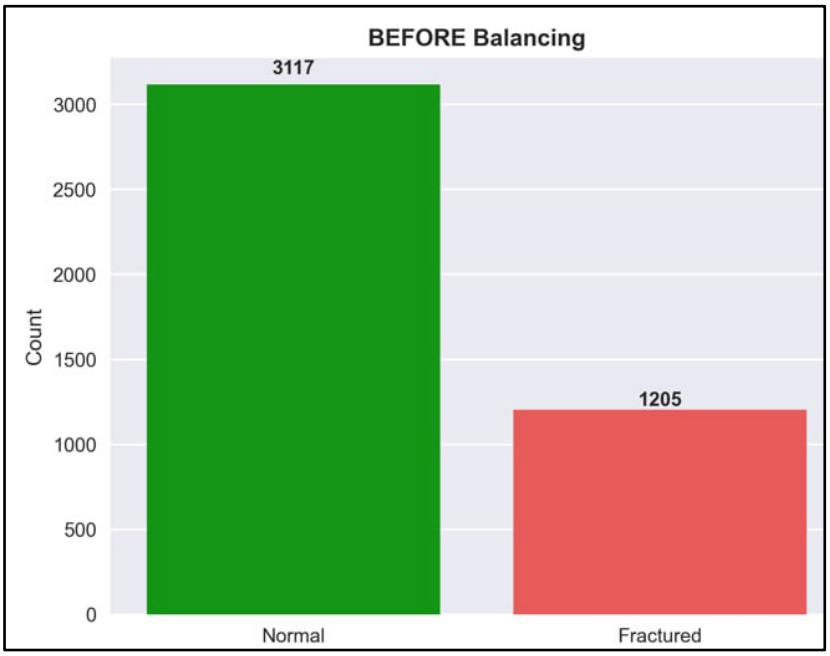

Figure 3: Normal vs Fracture Part Images

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

B. Resizing and Normalization Results

The raw radiographs had different image sizes, ranging from 512 × 512 pixels to 1,920 × 2,048 pixels. To ensure consistency and compatibility with the ResNet-101 architecture, all images were resized to 224 × 224 × 3. Pixel values were also scaled from their original 0–65,535 range (uint16) to a normalized 0–1 range (float32) and standardized using ImageNet mean and standard deviation values. A summary of the resizing and normalization outcomes is provided in Table 6.

Table 6: Summary of resizing and normalization

Step	Original	After Processing
Dimensions	Variable (512–1,920 px)	224 × 224 × 3
Pixel Range	0–65,535 (uint16)	0–1 (float32)

This preprocessing ensured uniform image size and intensity scale across the dataset, which helped improve model stability during training.

C. Augmentation for Class Balancing

To mitigate the effects of class imbalance, data augmentation was primarily applied to fractured images. The augmentation techniques included slight rotations (±15°), horizontal flipping, brightness and contrast adjustments (±20%), as well as minor zooming and translation operations. These transformations increased the number of fractured samples while retaining the essential diagnostic features of the radiographs. Following augmentation, the number of fractured samples in the hand dataset rose from 1,205 to 3,117. The updated class distribution after augmentation is presented in Figure 3.

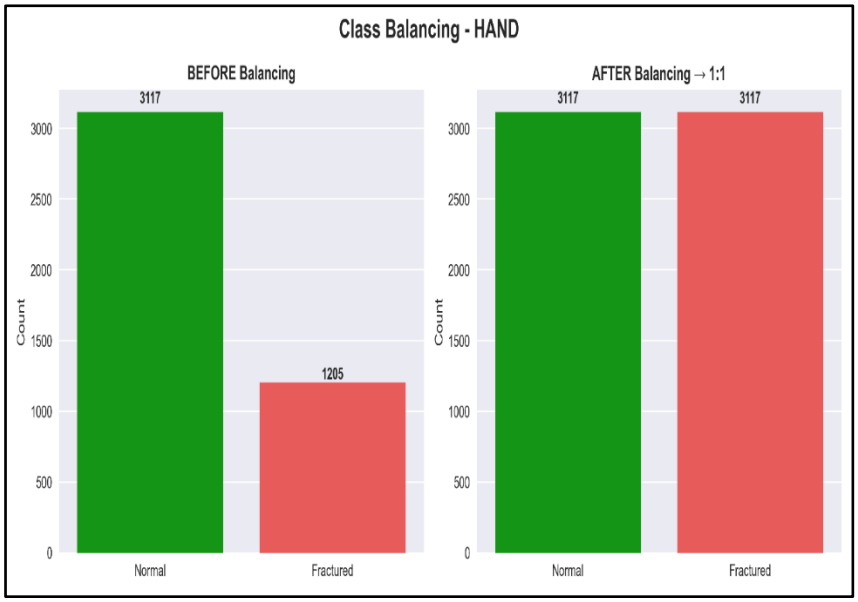

Figure 4: Class Balancing - HAND

Figure 4 illustrates the hand dataset before augmentation, with 3,117 normal images and 1,205 fractured images, and after augmentation, where both classes were balanced to approximately 3,117 images each.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

D. Model Performance Evaluation Results

The performance of the model was assessed using three experimental setups designed to examine how class balancing and hyperparameter tuning influence prediction accuracy.

(a). Model Results using Raw Dataset (Unbalanced)

This section presents the performance results obtained when the ResNet-101 model was trained using the original unbalanced dataset without any class balancing techniques. The results highlight the impact of class imbalance on model learning and predictive performance, as shown in Figures 5-8.

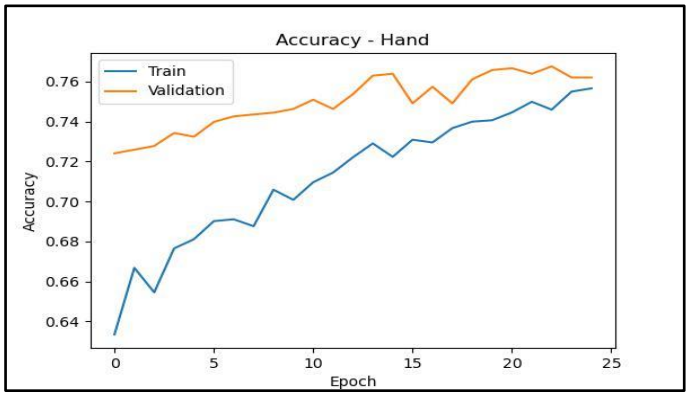

Figure 5: Training and Validation Accuracy Curves (Raw Dataset)

Figure 5 presents the training and validation accuracy curves for the hand dataset. The training accuracy shows a consistent increase and stabilizes at approximately 74% around epoch 20, indicating that the model is effectively learning from the training data. The validation accuracy improves at a slower rate and plateaus at about 70%. The small gap observed between the two curves suggests a slight limitation in the model's ability to generalize to unseen data. This pattern is largely attributed to class imbalance within the dataset, where normal images account for approximately 72% of the hand radiographs, leading the model to bias its predictions toward normal cases rather than fractures.

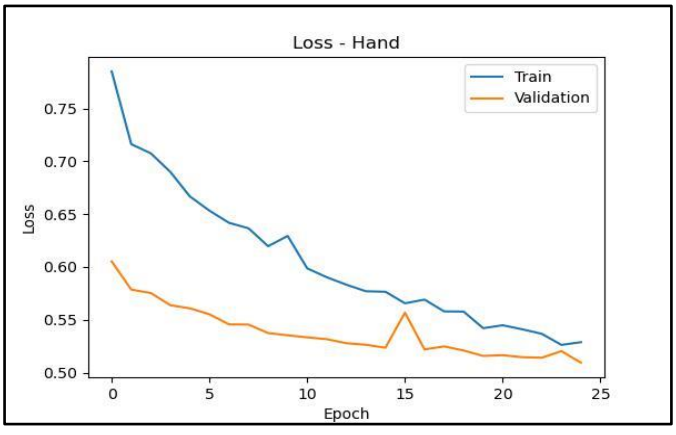

Figure 6: ROC Curve –Test Set (Raw Dataset)

Figure 6 presents the Receiver Operating Characteristic (ROC) curve for the hand fracture detection model. The model obtained an AUC value of 0.263, which is considerably lower than the 0.5 benchmark associated with random classification. The curve lies close to the diagonal line, indicating weak discrimination between fractured and normal images. This outcome

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

suggests that the model often assigns incorrect probability rankings to fracture cases, highlighting the impact of dataset imbalance during training.

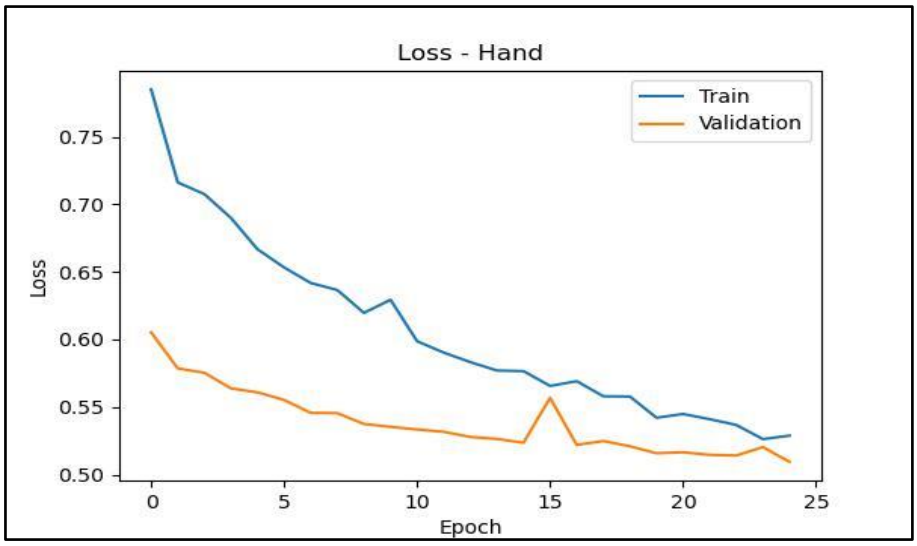

Figure 7: Training and Validation Loss Curves (Raw Dataset)

Figure 7 illustrates the training and validation loss curves throughout the training process. The training loss shows a steady decline from approximately 0.75 to 0.55, indicating that the model is effectively learning the dominant patterns within the dataset. In contrast, the validation loss exhibits slight fluctuations and stabilizes at a higher value of about 0.60 after epoch 15, suggesting reduced performance on unseen data. The consistent gap between the two curves implies that the model has difficulty capturing representative features for the minority fracture class.

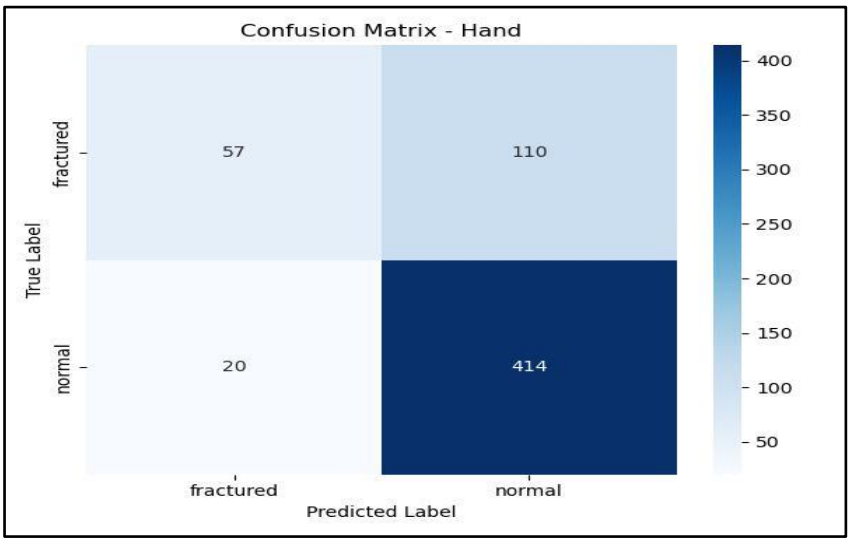

Figure 8: Confusion Matrix –Test Set (Raw Dataset)

Figure 8 illustrates the confusion matrix for the hand test dataset, which consists of 601 images. The model correctly detected 57 out of 167 fracture cases (TP = 57, FN = 110), yielding a recall of approximately 34.1%. For normal images, 414 out of 434 cases were accurately classified (TN = 414, FP = 20), resulting in a specificity of about 95.6%. Although the model demonstrates strong performance in identifying normal cases, the 110 missed fracture instances

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

represent a significant number of clinically critical errors. This disparity in prediction outcomes further indicates that the model is biased toward the majority class.

Key Observation: Although the baseline model attains an overall accuracy of approximately 74%, its effectiveness in detecting fractures is limited, as indicated by the low recall (34.1%) and AUC (0.263). These findings highlight the shortcomings of training on the original unbalanced dataset and emphasize the necessity for class balancing and further model optimization, which are addressed in the subsequent experiments.

(b). Model Results on Balanced Dataset (Augmentation + Class Weighting)

This section presents the performance of the ResNet-101 model after mitigating class imbalance through targeted data augmentation and the application of class weighting during training. Training the hand fracture detection model using a balanced dataset generated through these techniques significantly reduced the imbalance issue observed in the original dataset. Consequently, the model showed an improved ability to learn meaningful patterns for both fractured and normal cases (Figures 9–13).

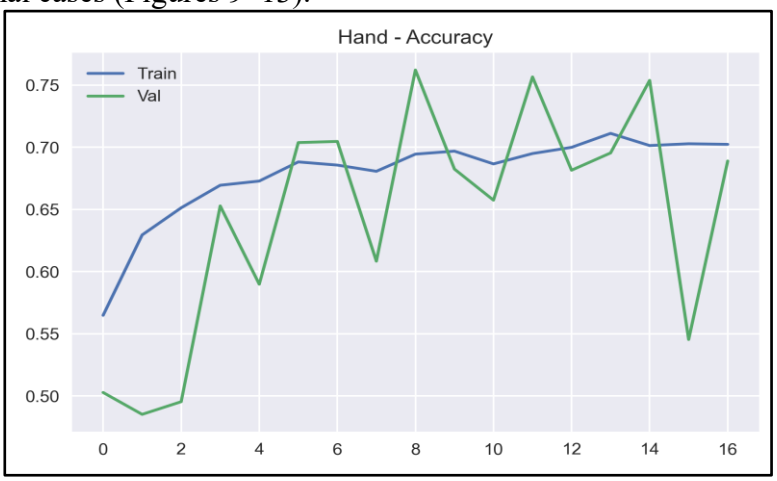

Figure 9: Training and Validation Accuracy - Test Set (Balanced)

Figure 9 indicates that both the training and validation accuracy curves display greater stability compared to the unbalanced experiment. The validation accuracy reaches approximately 0.75–0.77 during the mid-training epochs, which is clearly higher than the performance obtained with the raw dataset. However, it gradually decreases to about 0.69 by epoch 16, while the training accuracy stabilizes around 0.70. The reduced gap between the two curves suggests that the model is learning patterns from both classes more effectively than before, although the observed fluctuations indicate that generalization to fractured cases is still not entirely consistent.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

Figure 10: Training and Validation Loss - Test Set (Balanced)

Figure 10 shows that both the training and validation loss curves generally decline from approximately 0.83 to a range of 0.55–0.60, indicating a clear improvement over the unbalanced training scenario. Although class weighting helps mitigate the persistent validation loss plateau observed earlier, noticeable fluctuations and occasional spikes in validation loss (up to about 0.73) are still present. These variations suggest that, despite the reduction in class bias, the model continues to exhibit some level of instability during training.

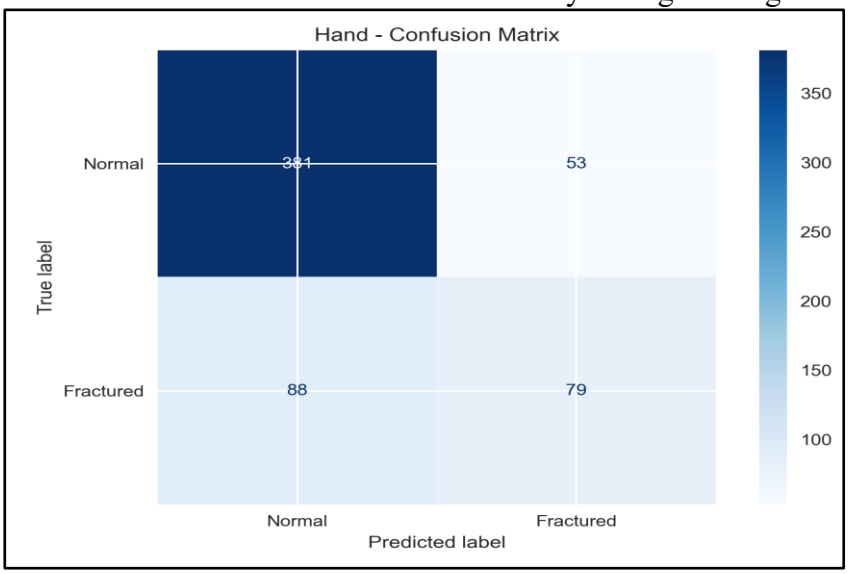

Figure 11: Confusion Matrix - Hand Test Set (Balanced)

Figure 11 illustrates the confusion matrix for the hand test dataset, which consists of 601 images. Out of 167 fracture cases, the model correctly detects 79 fractures (TP = 79) while failing to identify 88 cases (FN = 88), resulting in a fracture recall of approximately 47.3%. For the normal class, 381 out of 434 images are accurately classified (TN = 381, FP = 53), yielding a specificity of about 87.8%. Although there is an improvement in fracture detection compared to the raw dataset, the relatively high number of false negatives suggests that the model still has difficulty consistently recognizing fracture patterns in hand radiographs.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

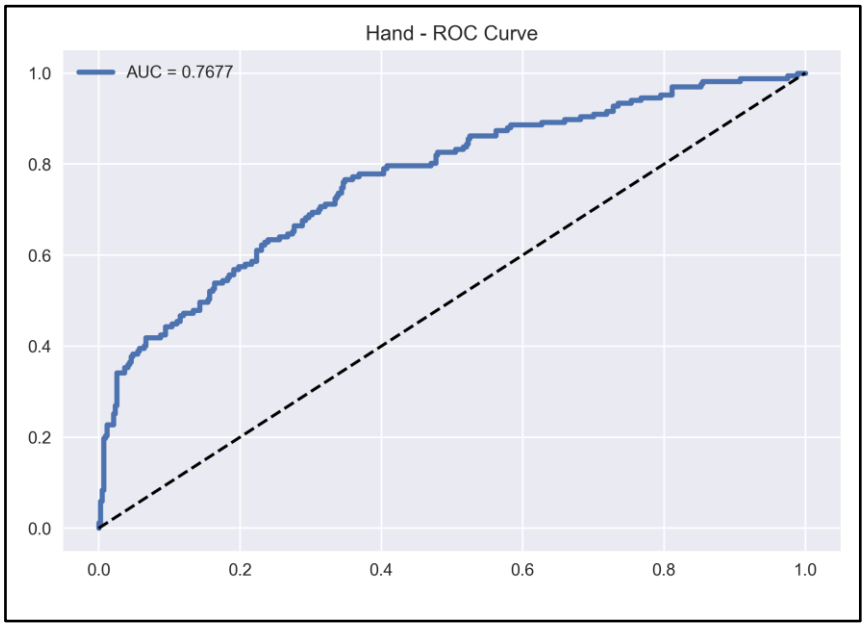

Figure 12: ROC Curve - Test Set (Balanced)

Figure 12 demonstrates that the ROC-AUC improves substantially to 0.767, compared to approximately 0.263 obtained from the unbalanced dataset. This increase reflects a stronger ability of the model to distinguish between fractured and normal images. The curve is clearly separated from the diagonal baseline, indicating meaningful predictive performance. However, the moderate shape of the curve suggests that adjusting the decision threshold may still require a trade-off between enhancing fracture detection and increasing false positive rates.

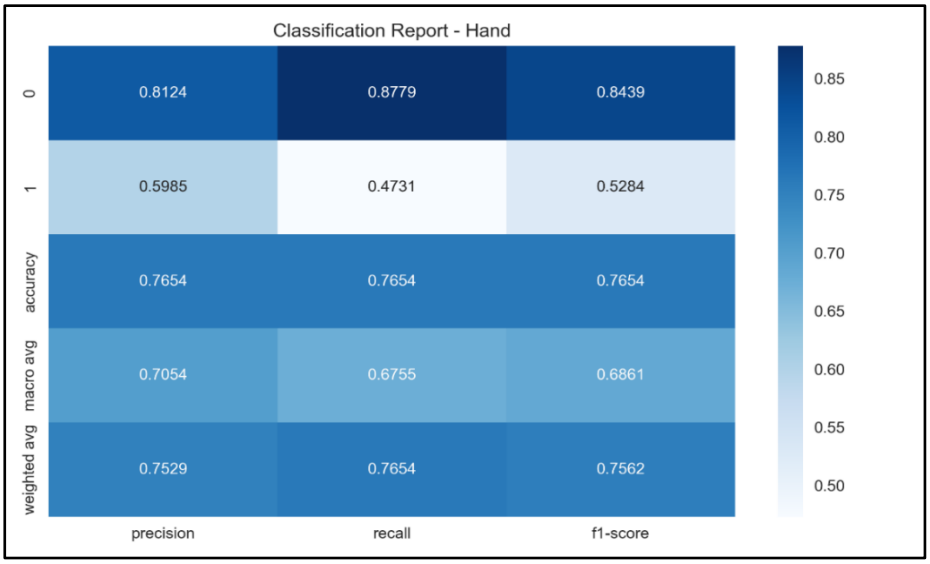

Figure 13 Classification Report - Test Set (Balanced)

Figure 13 provides a summary of the classification performance for both classes. For the normal class, the model achieves a precision of 0.812, recall of 0.878, and an F1-score of 0.844, demonstrating strong capability in identifying normal radiographs. For the fracture class, the model attains a precision of 0.599, recall of 0.473, and an F1-score of 0.528, indicating moderate improvement compared to the unbalanced model, although its detection performance remains lower than that of the normal class. The overall accuracy of the model reaches 0.765,

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

with a weighted average performance of approximately 0.756, reflecting a noticeable enhancement in classification performance, largely driven by improved fracture detection.
Key Observation: Compared to the unbalanced hand model (fracture recall 34.1%, AUC 0.263), the balanced training approach results in clear performance improvements. Fracture recall increases to 47.3%, representing a gain of approximately 13.2 percentage points, while the ROC-AUC rises significantly to 0.767, indicating a much stronger ability to distinguish between fractured and normal cases. The overall accuracy also improves to 0.765, reflecting enhanced classification performance. Despite these gains, fracture recall remains below clinically acceptable levels, highlighting that hand fracture detection continues to be challenging due to subtle fracture characteristics and variability within the dataset.

(c). Balanced Dataset + Hyperparameter Optimization

This section presents the performance of the ResNet-101 model after applying hyperparameter optimization to the balanced dataset. Bayesian optimization was employed to fine-tune critical training parameters to further enhance model convergence and overall fracture detection performance. This process was intended to refine the learning procedure by identifying more optimal parameter settings, thereby improving prediction stability and overall classification effectiveness. After conducting 10 optimization trials, the best-performing hyperparameter configuration was identified and is summarized in Table 7.

Table 7: Best Hyperparameters for Hand Fracture Model (Bayesian Optimization)

Parameter	Value
Backbone Architecture	ResNet-101
Dropout Rate	0.30
Dense Units	512
Learning Rate	5.52×10^{-4}

The results presented in Table 6 illustrate the optimal hyperparameter configuration obtained after 10 Bayesian optimization trials for the hand fracture model. The selected parameters suggest that a moderate dropout rate (0.30) was effective in minimizing overfitting, while the use of 512 dense units provided adequate capacity for feature learning. Furthermore, the relatively low learning rate (5.52×10^{-4}) contributed to stable and gradual convergence during training. Overall, this configuration enhances model generalization and ensures consistent classification performance on the balanced dataset. The tuned model is illustrated in Figure 14.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

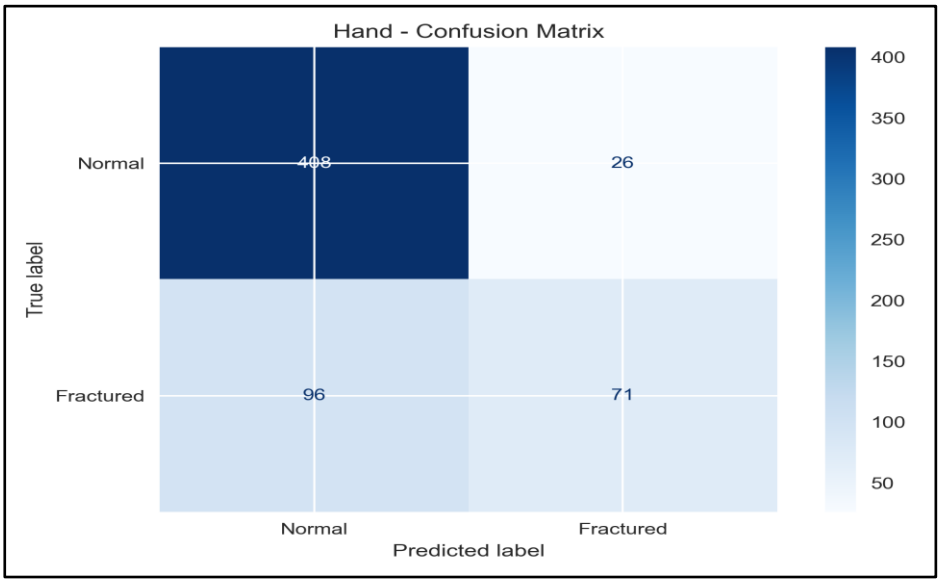

Figure 14: Confusion Matrix –Test Set (Tuned Model)

Figure 14 illustrates the confusion matrix for the tuned hand model on the independent test dataset. Out of 167 fracture cases, the model correctly detected 71 fractures (TP = 71) while failing to identify 96 cases (FN = 96), resulting in a fracture recall of approximately 42.5%. For the normal class, 408 out of 434 images were accurately classified (TN = 408, FP = 26), yielding a specificity of about 94.0%. Compared to the non-tuned balanced model, the number of false positives decreased significantly from 53 to 26, representing a reduction of approximately 51%. This improvement reflects better identification of normal images; however, it is accompanied by a slight decrease in fracture recall, suggesting that the tuned model adopts a more conservative prediction approach. The ROC of the tuned model is shown in Figure 15.

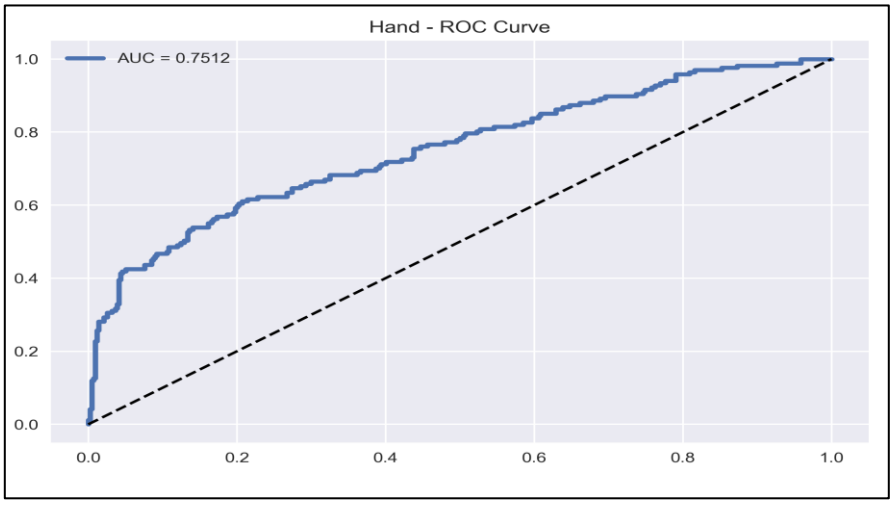

Figure 15: ROC Curve –Test Set (Tuned Model)

Figure 15 illustrates the ROC curve for the tuned hand model. The model achieved a ROC-AUC value of 0.7512, which is slightly lower than the 0.7677 obtained from the balanced baseline model. Although this indicates a minor decline in overall discriminative performance, the curve demonstrates improved behaviour in the high-specificity region, aligning with the significant reduction in false positives observed in the confusion matrix. This suggests that the

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

tuned model may be more appropriate in scenarios where minimizing false alarms is a critical requirement. Classification Report of the Tuned is shown in Figure 16.

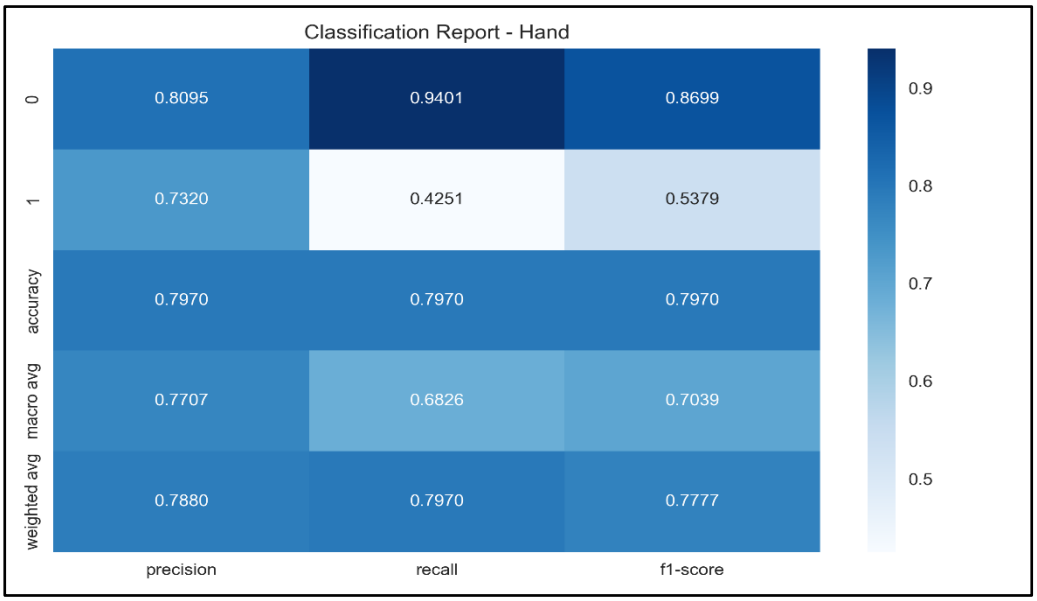

Figure 16: Classification Report –Test Set (Tuned Model)

Figure 16 provides a summary of the classification performance for both classes. For the normal class, the model achieves a precision of 0.8095, recall of 0.9400, and an F1-score of 0.8700, demonstrating strong capability in identifying normal radiographs. For the fracture class, the model attains a precision of 0.7320, recall of 0.4251, and an F1-score of 0.5380, indicating moderate performance in fracture detection. The overall accuracy reaches 79.7%, showing an improvement compared to the 76.5% achieved by the non-tuned balanced model. The weighted F1-score also increases to approximately 0.778, largely driven by the improved performance in the normal class.

Key Observation: Comparing the three hand experiments highlights the impact of both class imbalance handling and hyperparameter tuning. The model trained on the raw dataset achieved an overall accuracy of approximately 74%, but its ability to detect fractures was limited, with a recall of 34.1% and a ROC-AUC of 0.263. Following the application of data augmentation and class weighting, the balanced model showed improved fracture detection, with recall increasing to 47.3% and ROC-AUC rising to 0.7677, although performance remained inconsistent across classes. with the incorporation of Bayesian hyperparameter optimization, the tuned model further enhanced overall accuracy to 79.7% and improved specificity to 94.0%, while reducing false positives by approximately 51% compared to the balanced baseline. However, fracture recall decreased slightly to 42.5%, indicating that the improvements gained through tuning were largely due to better classification of normal cases rather than enhanced fracture detection.
Overall, the tuned hand model demonstrates a more stable and efficient predictive performance. Nevertheless, the continued difficulty in accurately detecting subtle hand fractures suggests that additional methodological refinements are required.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

(d). Results Comparison

This section provides a comparison of the performance of the bone fracture detection system across the three experimental setups introduced earlier: training on the raw dataset, training on the balanced dataset without hyperparameter tuning, and training on the balanced dataset with Bayesian hyperparameter optimization. The comparison, presented in Table 8, demonstrates how dataset balancing and systematic tuning affect model performance across different fracture categories.

Table 8: Performance Comparison for Hand Fracture Detection

Metric	Raw Dataset	Balanced Dataset	Balanced + Tuning
Accuracy	74%	76.5%	79.7%
Recall	34.1%	47.3%	42.5%
Specificity	95.6%	87.8%	94.0%
AUC	0.263	0.767	0.7512

The results in Table 8 highlight the effects of dataset balancing and hyperparameter tuning on hand fracture detection. The model trained on the raw dataset achieved an accuracy of 74%, but its ability to detect fractures was limited, with a low recall of 34.1%, indicating a bias toward normal cases. After applying data augmentation and class weighting, fracture detection improved considerably, with recall increasing to 47.3% and AUC rising to 0.767, showing better separation between fractured and normal images. With the addition of hyperparameter optimization, the tuned model further improved overall accuracy to 79.7% and increased specificity to 94.0%, while maintaining an AUC of 0.7512. However, recall dropped to 42.5%, suggesting that the model became more conservative and failed to detect more fracture cases compared to the balanced model.

These findings emphasize a trade-off between sensitivity and specificity. Although hyperparameter tuning enhances overall accuracy and reduces false positives, it does not improve fracture detection. In clinical scenarios, where missing fractures is critical, the balanced model with its higher recall may be more suitable depending on the intended use.

Table 9: Comparison with Studies Using the MURA Dataset

Study	Dataset	Model	Task	Accuracy
[22]	MURA (hand)	Multiple CNNs	Fracture classification for MURA Dataset	75.36%
[23]	MURA (hand)	Multiple CNNs	Fracture classification for MURA Dataset	74.59%
Proposed Study	MURA (hand)	ResNet-101	Hand fracture detection	79.7%

Table 9 presents a comparison between the proposed model and existing studies that utilized the MURA dataset for hand fracture classification. The proposed ResNet-101 model achieved an accuracy of 79.7%, surpassing the reported accuracies of 75.36% and 74.59% in the studies by [22] and [23], respectively.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

This performance improvement can be attributed to the use of a deeper network architecture, combined with data balancing techniques and Bayesian hyperparameter optimization, which enhanced feature extraction and overall classification effectiveness. However, differences in experimental setups such as preprocessing methods, data partitioning strategies, and evaluation procedures may affect the reported outcomes. Therefore, the comparison should be interpreted with consideration of these methodological variations.

(F) Prediction and Explainability Analysis using Grad-CAM

To enhance the transparency of the proposed fracture detection system, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to visualize the regions of the radiograph that most influenced the model's predictions. Grad-CAM generates heatmaps that highlight salient areas within the X-ray image, enabling verification that the model focuses on clinically relevant anatomical structures rather than irrelevant background regions. To address the limitation of purely qualitative interpretation, a structured qualitative evaluation framework was introduced to systematically assess the Grad-CAM outputs. Each heatmap was evaluated using four criteria: anatomical relevance, fracture-site relevance, attention concentration, and clinical interpretability. Each criterion was scored on a scale of 0 to 2, and an overall average score was computed to quantify the quality and consistency of the visual explanations.

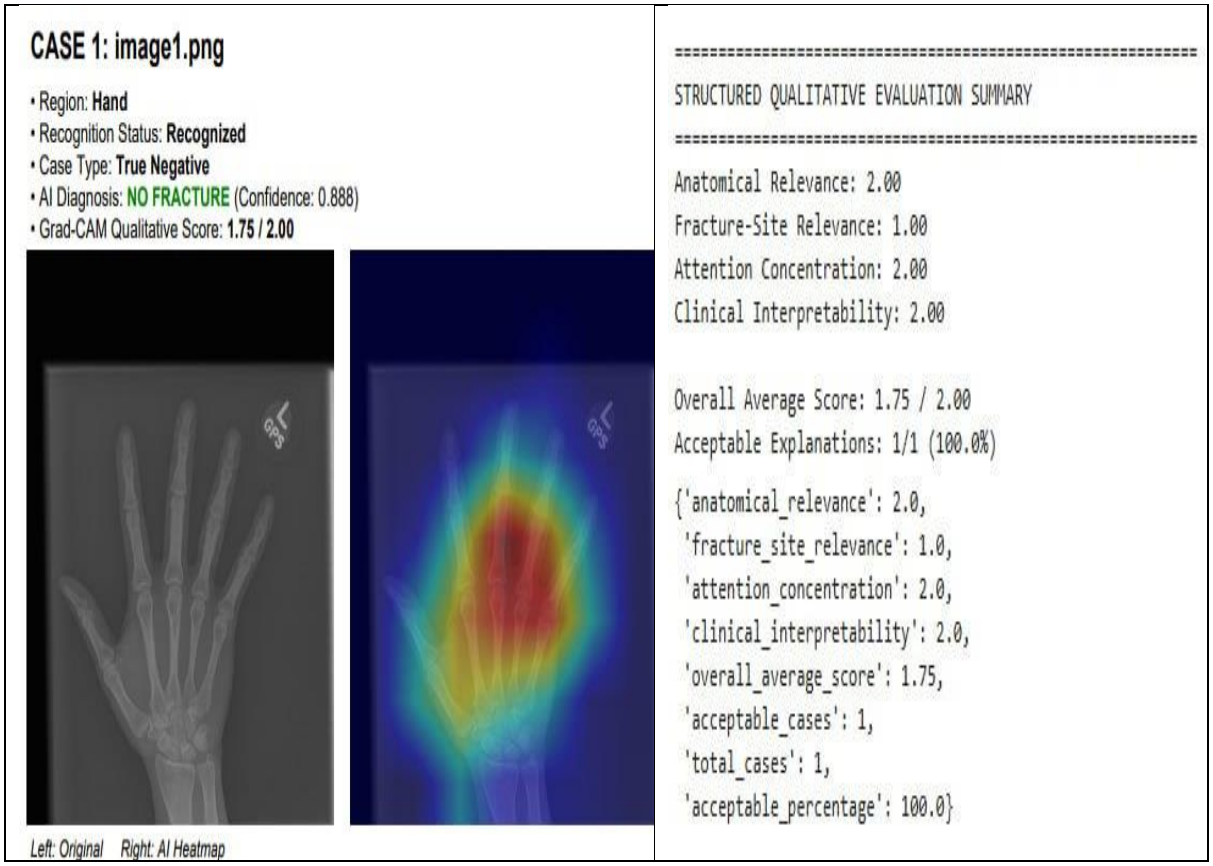

Figure 17: Grad-CAM Visualization for a True Negative hand X-ray Case.

Figure 17 presents a representative true negative case from the hand dataset. The model correctly classifies the image as normal (no fracture) with a confidence score of 0.888. The Grad-CAM heatmap highlights regions around the hand joints and bone structures, indicating that the model attends to relevant anatomical areas even in the absence of fractures.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

The structured qualitative evaluation for this case yields an overall score of 1.75 out of 2.00, with maximum scores in anatomical relevance (2.0), attention concentration (2.0), and clinical interpretability (2.0), while fracture-site relevance is moderately scored (1.0), as expected for a normal case. This demonstrates that the model provides meaningful and clinically coherent explanations for its predictions. Overall, the integration of Grad-CAM visualization with structured qualitative evaluation provides a more systematic and reliable assessment of model explainability, thereby strengthening the interpretability of the proposed system and supporting its potential application as a clinical decision support tool.

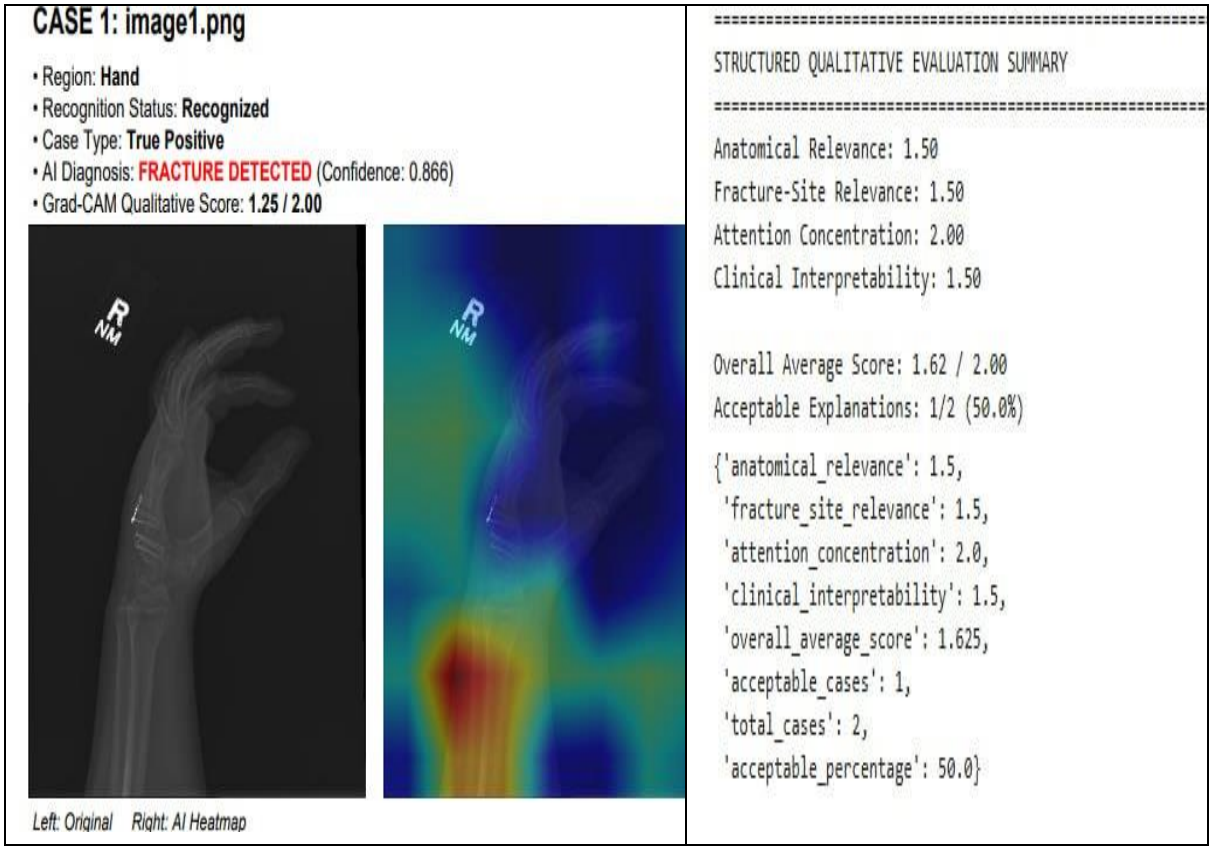

Figure 18 Grad-CAM Visualization for a True Positive Hand X-ray Case.

Figure 18 presents a representative true positive case from the hand dataset, where the model correctly identifies the presence of a fracture with a confidence score of 0.866. The Grad-CAM heatmap highlights regions around the fractured bone, indicating that the model focuses on clinically relevant areas associated with structural abnormalities. However, compared to the normal case, the attention appears more diffused, suggesting less precise localization of the fracture region. The structured qualitative evaluation for this case yields a score of 1.25 out of 2.00, reflecting moderate alignment with expected clinical focus. While attention concentration remains relatively strong, slightly lower scores in anatomical relevance and clinical interpretability indicate room for improvement in accurately localizing fracture regions. Across the evaluated cases, the overall average qualitative score is 1.62 out of 2.00, with approximately 50% of the explanations classified as acceptable. This suggests that, although the model demonstrates reasonable interpretability, its ability to consistently highlight precise fracture locations varies across samples.

Overall, these findings indicate that while the Grad-CAM explanations are generally meaningful and clinically relevant, further refinement is needed to improve localization

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

accuracy, particularly for complex or subtle fracture cases. These findings suggest that while the Grad-CAM explanations are generally meaningful, further refinement is required to improve localization accuracy for fracture detection, particularly in complex or subtle cases.

(e). Interpretation of Grad-CAM Results

The Grad-CAM visualizations indicate that the proposed ResNet-101 model generally focuses on the hand region during prediction. In fracture cases, the heatmaps tend to highlight areas associated with bone discontinuities or abnormal structural patterns, while in normal cases, the model emphasizes relevant anatomical bone regions during evaluation. This behaviour suggests that the model relies on meaningful anatomical features rather than irrelevant image artifacts. The structured qualitative evaluation further supports these observations. Higher interpretability scores were obtained for normal cases, where attention was more consistently aligned with anatomical structures. In contrast, fracture cases exhibited greater variability in localization, with comparatively lower scores. This indicates that, although the model can identify relevant regions, its ability to precisely localize fracture sites is not consistent across all cases. Overall, the Grad-CAM analysis demonstrates that the model provides meaningful and clinically relevant visual explanations. However, the observed variability in fracture localization highlights an important limitation. These findings suggest that the system is better suited as an assistive diagnostic tool to support clinical decision-making rather than as a fully autonomous solution.

(f). Automated Clinical Reporting

In addition to prediction and explainability, the proposed system incorporates an automated clinical-style reporting module. Following the inference process, the system compiles the original X-ray image, corresponding Grad-CAM heatmap, predicted class label, and confidence score into a structured PDF report. Each report entry presents key diagnostic information, including the detection status (fractured or normal), model prediction, and associated visual explanation. This structured format is designed to enhance interpretability and support efficient review by clinicians and researchers. Although this functionality improves the usability and practical relevance of the system, it is intended to function as a supportive decision-making tool rather than a substitute for professional clinical judgment. The generated reports provide a concise and organized summary of model outputs and visual explanations, facilitating further analysis and validation.

V. CONCLUSION

This study developed an explainable deep learning framework for automated hand fracture detection using the ResNet-101 architecture and the MURA dataset. The experimental results demonstrate that models trained on unbalanced datasets tend to favour the majority normal class, leading to poor fracture detection performance. The application of dataset balancing techniques, including targeted augmentation and class weighting, significantly improved fracture detection capability. In addition, Bayesian hyperparameter optimization enhanced model convergence and overall classification stability. Beyond predictive performance, Grad-CAM was incorporated to provide visual explanations of model decisions. A structured qualitative evaluation framework was introduced to assess the interpretability of the generated heatmaps, revealing that the model generally focuses on anatomically relevant regions. However, variability in localization performance particularly in fracture cases indicates that further improvements are required to achieve consistent and precise fracture identification. Despite achieving an overall accuracy of 79.7% and strong AUC performance, the model exhibits relatively low fracture recall, indicating that some fracture cases are missed. This

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956

limitation is critical in medical applications and highlights the need for further refinement prior to practical deployment. Additionally, the current experimental setup is based on a fixed data split and does not include cross-validation or statistical significance testing, which may affect the generalizability of the findings. Overall, the proposed explainable ResNet-101 model demonstrates promising performance and interpretability for hand fracture detection. However, it is more appropriately positioned as an assistive decision-support tool rather than a fully autonomous diagnostic system. Future work will focus on improving fracture recall, incorporating larger and more diverse datasets, enabling multi-view radiograph analysis, and adopting more rigorous validation strategies such as cross-validation and repeated experiments. These enhancements are expected to improve model robustness, generalization, and clinical applicability.

REFERENCES

[1] T. H. H. Aldhyani *et al.*, "Diagnosis and detection of bone fracture in radiographic images using deep learning approaches," *Diagnostics*, 2025, doi: 10.3390/diagnostics15020200.
[2] J. Zou and M. R. Arshad, "Detection of whole-body bone fractures based on improved YOLOv7," *Biomed. Signal Process. Control*, vol. 91, p. 105995, May 2024, doi: 10.1016/j.bspc.2024.105995.
[3] T. Meena and S. Roy, "Bone fracture detection using deep supervised learning from radiological images: A paradigm shift," *Diagnostics*, vol. 12, no. 10, p. 2420, Oct. 2022, doi: 10.3390/diagnostics12102420.
[4] M. E. Sahin, "Image processing and machine learning-based bone fracture detection and classification using X-ray images," *Int. J. Imaging Syst. Technol.*, vol. 33, no. 4, pp. 853–865, 2023, doi: 10.1002/ima.22849.
[5] A. Alam *et al.*, "Novel transfer learning-based bone fracture detection using radiographic images," *BMC Med. Imaging*, vol. 25, no. 1, p. 5, Jan. 2025, doi: 10.1186/s12880-024-01546-4.
[6] A. Gupta and S. Sharma, "Deep learning approaches for bone fracture detection: A comprehensive review," *Comput. Biol. Med.*, 2024. (as cited in dissertation context)
[7] J. Ju and Y. Cai, "Advanced CNN models for automated fracture detection in X-ray imaging," *IEEE Access*, vol. 11, pp. 45678–45692, 2023.
[8] Y. Liu *et al.*, "A survey on deep learning for medical image analysis," *Med. Image Anal.*, vol. 73, p. 102213, 2021, doi: 10.1016/j.media.2021.102213.
[9] M. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Comput. Sci.*, vol. 2, no. 6, p. 420, 2021, doi: 10.1007/s42979-021-00815-1.
[10] L. Alzubaidi *et al.*, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.
[11] Z. Zhou, "Machine learning: Concepts, techniques and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024.
[12] A. Noureen *et al.*, "Analysis and classification of bone fractures using machine learning techniques," *E3S Web Conf.*, vol. 409, p. 02015, 2023, doi: 10.1051/e3sconf/202340902015.
[13] M. Thaarakaraam *et al.*, "CNN based bone fracture detection for medical imaging," *Int. J. Tech. Res. Sci.*, vol. 9, no. Spl, pp. 27–35, 2024, doi: 10.30780/specialissue-ISET-2024/038.
[14] S. Khan *et al.*, "Deep learning on graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 249–270, Jan. 2020, doi: 10.1109/TKDE.2020.2981333.
[15] Y. Li *et al.*, "Convolutional neural networks for medical image analysis: A survey," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 1–20, 2022.

[16] A. Younesi *et al.*, "Operator-level design choices for convolutional neural networks in medical imaging," *IEEE Trans. Med. Imaging*, 2024.
[17] S. Ghosh *et al.*, "Class imbalance in deep learning: A comprehensive survey," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024.
[18] S. Banerjee *et al.*, "Physics-informed neural networks for medical image analysis: A survey," *IEEE Trans. Med. Imaging*, 2024.
[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[20] Gencer, Y., & Erdem, O. (2023). A review on image-based medical diagnosis using ResNet-50 architecture. *International Journal of Research and Analytical Reviews, 10*(4), 421–430. Retrieved from https://www.ijrar.org/papers/IJRAR24A3421.pdf
[21] Doğan, A., & Çelik, D. (2022). Deep learning approaches for retinal image classification using ResNet architectures. Biomedical Signal Processing and Control, 73, Article 103433. https://doi.org/10.1016/j.bspc.2021.103433
[22] Kandel, I., Castelli, M., & Popovič, A. (2020). Musculoskeletal Images Classification for Detection of Fractures Using Transfer Learning. Journal of Imaging, 6(11), 127. https://doi.org/10.3390/jimaging6110127
[23] Kandel, I., Castelli, M., & Popovič, A. (2021). Comparing Stacking Ensemble Techniques to Improve Musculoskeletal Fracture Image Classification. Journal of Imaging, 7(6), 100. https://doi.org/10.3390/jimaging7060100
[24] Adhyaksa, Resky, and Bedy Purnama. 2025. "Application of VGG16 in Automated Detection of Bone Fractures in X-Ray Images." Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi) 9(1):118–29. doi:10.29207/resti. v9i1.6101.
[25] Guan, H., Yap, P.-T., Bozoki, A., & Liu, M. (2024). Federated learning for medical image analysis: A survey. Pattern Recognition, 150, Article 110424. https://doi.org/10.1016/j.patcog.2024.110424
[26] Yang, T., Yu, X., McKeown, M. J., & Wang, Z. J. (2024). When federated learning meets medical image analysis: A systematic review with challenges and solutions. APSIPA Transactions on Signal and Information Processing, 13, Article e38. https://doi.org/10.1561/9781680837383
[27] Venkatesam, Anne Sreyas, Mohd Irfan, Lavish Kansal, Sunil Prashanth Kumar, Swaroop Challapalli, and Shanmugasundaram Hariharan. 2025. "Bone Fracture Detection through Advanced Neural Network Architectures" edited by N. Nwulu, K. Natarajan, R. Shanmugasundaram, P. K. Balachandran, M. Krishnamoorthy, and P. Mounica. E3S Web of Conferences 619:03011. doi:10.1051/e3sconf/202561903011.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19508956