International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 7, July, 2026, pp. 2389-2400
Journal homepage: http://iieta.org/journals/isi

Explainable Multi-Class Classification of Leg Fractures in Radiographs Using a Multi-Branch Convolutional Neural Network with Gradient-Weighted Class Activation Mapping

Yididiya Abraham Mojo[1], Tsega Asresa Mengistu[2], Ayodeji Olalekan Salau[3,4], Olubunmi Ajala[5], Kaung Myat Paing[6], Ting Tin Tin[6,7*], Sepiribo Lucky Braide[8], Maryam Abdullahi[5]

[1] Department of Information Systems, College of Informatics, Wolaita Sodo University, Wolaita Sodo 4620, Ethiopia
[2] Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara 6040, Ethiopia
[3] Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti 360101, Nigeria
[4] Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai 600124, India
[5] National Centre for Artificial Intelligence and Robotics (NCAIR), National Information Technology Development Agency (NITDA), Abuja 564, Nigeria
[6] Faculty of Data Science and Information Technology, INTI International University, Persiaran Perdana BBN Putra Nilai, Nilai 71800, Malaysia
[7] School of Information Technology, UNITAR International University, Petaling Jaya 47301, Malaysia
[8] Department of Electrical and Electronics Engineering, Rivers State University, Port Harcourt 5080, Nigeria

Corresponding Author Email: ting.tintin@researcher.unitar.my

https://doi.org/10.18280/isi.310721

Received: 22 March 2026
Revised: 25 May 2026
Accepted: 16 June 2026
Available online: 31 July 2026

Keywords:
bone fracture classification, radiography, multi-branch Convolutional Neural Network, explainable artificial intelligence, Gradient-weighted Class Activation Mapping, medical image analysis, human health

ABSTRACT

Timely fracture assessment on radiographs remains difficult when fracture patterns are subtle, and model predictions are not readily interpretable. An explainable five-class classifier is proposed in this work to identify bone fractures using a multi-branch Convolutional Neural Network (CNN) architecture. The retrospective dataset comprised 8,000 radiographs from two hospitals in Ethiopia. The dataset was independently annotated by two radiologists and one orthopedist. The dataset was equally distributed among five categories: non-fractured, transverse, oblique, spiral, and comminuted, with 1,600 images in each class. In the pre-processing stage, all images were scaled to a 224 × 224-pixel size and normalized. Data augmentation was applied to increase dataset variation, and the data were split into three subsets for training, validation, and testing. The proposed model achieved an accuracy of 90.0% on the test set, suggesting that the diagnostic ability of the model is robust even against complex radiographic patterns. In comparative experiments, pre-trained architectures such as VGG19 offered competitive test accuracy with varying computational footprint. To evaluate the interpretability of the model, Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations were generated to demonstrate the visual region of interest for each predicted class. Although these findings demonstrate promising performance, external validation and quantitative interpretability benchmarks are required prior to clinical deployment.

1. INTRODUCTION

Fractured bones are common and severe injuries that need precise and timely diagnosis for proper treatment and prevention of long-term complications. Traditional fracture diagnosis, relying primarily on human interpretation of X-rays, is fraught with inherent limitations. It is a time-consuming activity, considering the increasing volume of medical imaging data, and is susceptible to human error due to several factors like poor image quality, misalignments, and the subjective nature of interpretation [1]. Such diagnostic inaccuracy and delay can be of severe impact on patient health, thus emphasizing an urgent need for clinical application of automated, highly accurate, and sensitive diagnostic resources that could complement or even surpass conventional diagnostic channels [2]. As an antidote to these issues, deep learning, and more notably Convolutional Neural Networks (CNNs), have emerged as a revolutionary technology for medical image analysis. Deep learning algorithms can automatically learn complex, hierarchical features from raw pixel data directly, thereby bypassing the tedious hand-crafted feature engineering and enabling the identification of subtle anomalies that might not be identified through classical methods [3].

This capability has led to demonstrated success in a variety of medical diagnostic applications, establishing a promising precedent for its use in classifying bone fractures [4].

Transparency of decision-making turns out to be one of the main factors that prevents the use of deep learning in medicine, as it uses the "black box" approach. This means that the use of such systems is not possible in the day-to-day practice of medical professionals [5]. It is crucial in trust-building between clinicians and patients to enable medical personnel to validate artificial intelligence (AI) predictions, identify and avoid biases, confirm regulatory compliance, and ultimately improve patient outcomes. Explainable artificial intelligence (XAI) is a pivotal, rapidly evolving area of research that seeks to explain AI decisions in a manner that makes them understandable and transparent to humans and enables their safe incorporation into clinical pathways [6].

Medical images are inherently computationally complex and require complex algorithms to process them efficiently. But complex model architectures often exacerbate the “black box” problem. To address this, the Keras Functional Application Programming Interface (API) allows building complex multi-branch networks, and Gradient-weighted Class Activation Mapping (Grad-CAM) provides explicit visual explanations. These tools together improve clinical reliability and validate correct model behavior. Moreover, the Functional API simplifies the design of complex networks by managing multi-input pipelines (e.g., multi-view X-rays at the same time) and by employing shared layers to prevent redundant feature extraction among parallel branches [7, 8].

The Functional API supports a multi-branch architecture that allows for multi-input processing and feature integration that sequential networks cannot support. In the domain of radiographic bone fracture classification, processing of multiple views (e.g., frontal and lateral projections) allows the model to leverage complementary spatial information, thus improving diagnostic accuracy and clinical reliability. To evaluate decision transparency, a visual explanation of model predictions was provided using Grad-CAM. This interpretability analysis discovers the anatomical regions responsible for model outputs, ensuring that predictions are aligned with clinically relevant features. The primary dataset utilized in this study was the bone dataset, which is well known to be of use in bone fracture classification due to its large image set and expert-annotated form.

2. RELATED WORK

CNNs are used in this study to classify bone fractures in X-ray images. Specifically, VGGNet, ResNet, and Inception architectures are used for binary classification (Fracture/Non-Fracture). Although the study effectively shows that these CNN models achieve high accuracy, specificity, and sensitivity after image preprocessing (resizing, normalization, and augmentation), providing a definite performance advantage over conventional machine learning techniques, its usefulness is constrained by an important omission. The main drawback of the work is that real-world deployment challenges are not taken into account. In particular, the models' ability to generalize across the presence of various or complex fracture types that are not represented in their particular dataset is not tested, nor are they tested for robustness against image variations inherent in clinical settings, such as the impact of different X-ray lighting/exposure conditions. Our current paper focused on creating a more robust and generalizable fracture classification model that can maintain high performance despite the noise and heterogeneity present in real-world clinical X-ray data. This gap is directly caused by the failure to systematically address external environmental robustness and comprehensive fracture variability [9].

In order to establish generalized design principles, the related work is a meta-analysis that focuses on the advantages and disadvantages of current deep learning-based bone fracture detection techniques. Its main benefit is that it offers a thorough overview and conceptual framework for identifying important factors in this field. Nevertheless, this study has a serious structural flaw: it is only theoretical and not practical. It offers only a high-level conceptual discussion and does not suggest or assess any novel methodology for bone fracture detection, despite discussing the work of other researchers. As a result, this analysis offers no concrete proof of the performance, efficiency, or architectural enhancements required to address the shortcomings found in existing approaches. The focus of our current paper, which goes beyond conceptual analysis to propose and rigorously validate a specific, novel deep learning architecture designed to practically address the operational gaps identified in the existing literature, was directly justified by this critical deficiency: the lack of an empirically tested method [10].

In an effort to outperform current accuracy benchmarks, the study offers a deep learning-based method that uses CNNs to automatically identify and classify bone fractures in annotated X-ray images. The approach effectively exhibits high accuracy in the main fracture detection task, demonstrating CNNs' broad applicability. The primary and most significant drawback of this work is its narrow classification scope, despite its proven high accuracy. The methodology is limited to a binary classification task: determining whether a bone is fractured or not. The critical sub-task of classifying fracture types is not addressed, particularly regarding the common or clinically significant fracture patterns (e.g., transverse, spiral, oblique, comminuted). Because of this omission, the model offers a diagnosis but lacks the clinical detail needed for treatment planning, which depends on understanding the fracture morphology. Our current research is clearly necessary because of this particular limitation—the incapacity to provide a multi-class classification beyond simple detection. In order to provide a more thorough and clinically useful diagnostic tool, our paper is directly motivated to extend the deep learning methodology to not only detect the presence of a fracture but also systematically classify the specific types of fractures [11].

Meena and Roy [12] presented a deep supervised learning-based fracture detection technique. By teaching a deep neural network to effectively identify and diagnose bone fractures on radiological images, the authors successfully illustrate a major paradigm shift. The methodical comparison of three different deep learning models, which shows notable performance improvements over conventional, non-deep learning methods when assessed using cutting-edge metrics like accuracy, is the work's strongest point. The work is severely constrained by its reliance on limited data, despite these advancements. The models were specifically limited to binary classification (fracture/non-fracture) because there was a "lack of a comprehensive labeled dataset for different types of bone fracture detection." The models can detect the existence of a fracture, but they are unable to classify the precise type or morphology (such as transverse, comminuted, or spiral), which is crucial for treatment planning. This data-centric limitation prevents the models from producing a clinically useful output. This gap the inability to show reliable multi-class detection because of data constraints directly drives the need for our present study. In order to overcome the lack of fully labeled, multi-class fracture datasets and enable a useful and clinically relevant multi-class fracture classification system, our paper focuses not only on the model architecture but also on novel data augmentation and transfer learning techniques.

CNNs are successfully used in the study by Thian et al. [13] to detect and locate fractures on wrist radiographs. The authors obtained remarkable metrics by using three different CNN classifiers. They reported an overall precision of up to 98% in fracture detection and optimal accuracy values of up to 95% in the classification of fracture vs. non-fracture and the presence of wrist spearhead. The potential of CNNs to greatly improve the precision and effectiveness of diagnostic procedures for common wrist injuries is persuasively demonstrated by this work. However, two significant methodological flaws seriously compromise the reported results. First, the model's ability to generalize to unseen, diverse clinical data is unproven, and the reported accuracy metrics may be unduly optimistic due to the study's failure to use an independent validation set. Second, the classification task's scope was limited because it only considered the presence or absence of various fracture types. The model's clinical utility for treatment planning is severely limited by the absence of fracture type differentiation (e.g., scaphoid, distal radius fracture types). In order to demonstrate real-world robustness, our current paper must prioritize rigorous, multi-institutional validation using a dedicated, independent dataset. Additionally, the classification task must be expanded to include the clinically necessary spectrum of distinct fracture types.

The study [14] made recommendations for bone fracture analysis and classification using conventional Machine Learning algorithms. Using a dataset of bone fracture photos, the authors' method extracts hand-crafted features from the photos before feeding them into machine learning classifiers. This work has two main limitations that limit its generalizability and clinical applicability. First, the researcher limits the study to a narrow, probably binary classification scope (e.g., fractured vs. non-fractured) by excluding various types of fractures. The system is clinically inadequate for diagnosis and treatment planning due to its incapacity to distinguish between different fracture morphologies. Second, the performance evaluation ignores other important evaluation metrics like precision, recall, F1-score, or the Area Under the Curve (AUC) in favor of an over-reliance on accuracy. This limited assessment may give a false impression of the model's actual performance, particularly its sensitivity (recall) in identifying true positive fracture cases, since it ignores possible class imbalance. Our current research, which aims to use Deep learning to achieve multi-class fracture classification and validate performance using a full suite of advanced evaluation metrics, is directly necessary due to the critical lack of comprehensive fracture typology and robust, multi-metric performance evaluation.

In order to detect and classify bone fractures in medical images, the study [15] introduced a novel deep learning approach that combines a Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) model. The methodology's main advantage is its integration of dilated convolutions to efficiently capture multi-scale features, resulting in high accuracy rates across several fracture types and successfully outperforming traditional methods. However, a crucial methodological constraint the size of the dataset significantly limits the effectiveness and long-term clinical credibility of this approach. A comparatively small dataset of just 3,842 samples was used to train and validate the model. A moderate reported accuracy of 82.1% indicates that the model's generalizability and capacity to learn the wide heterogeneity of real-world fracture cases are severely impacted by such a small dataset. Although the architecture shows promise in theory, the small sample size raises the possibility of overfitting and reduces the likelihood that the performance improvements will persist in various, untested clinical settings.

By combining three potent CNN architectures, DenseNet201, VGG16, and ResNet152V2, into a single model, the study highlights the significant benefit of ensemble learning in medical image analysis. This unification effectively attained an astounding 97% validation accuracy in bone fracture detection, demonstrating the superior predictive power obtained by merging various cutting-edge models. Despite its excellent performance, the paper has a serious flaw: it does not use XAI methods (like Grad-CAM). The system functions as an unacceptable "black box" in a high-risk clinical setting by merely reporting high accuracy without clearly describing how the ensemble model makes its predictions. Because practitioners are unable to confirm whether the model is concentrating on clinically relevant features (such as the fracture line) or irrelevant artifacts, this lack of transparency restricts the vital knowledge required for developing complete clinician confidence and support. Our current paper is necessary because of this obvious limitation, which is the lack of model interpretability. Our goal was to transform a high-performing black-box system into a transparent, reliable, and clinically useful diagnostic tool by using powerful deep learning models in addition to methodically integrating and evaluating XAI techniques to provide quantitative and visual explanations for fracture classification [2].

The goal of the study is to use a lightweight DenseNet121 model to diagnose sports-related bone fractures with high efficiency and performance. It outperforms heavier baseline architectures like ResNet-50 and VGG-16 thanks to its proven state-of-the-art precision of 90.3%. For deployment in settings requiring speed, like high-speed sports medicine, the focus on a lightweight but extremely accurate model is essential. However, this lightweight model's lack of interpretability severely limits its efficacy and, most importantly, its clinical adoption. In order to explain why the lightweight DenseNet121 model achieves its predictions, the paper does not directly present work on XAI techniques (such as Grad-CAM or LIME). Building clinician trust and complete confidence in a high-stakes setting such as sports medicine requires an understanding of the model's reasoning (i.e., confirming that it focuses on the actual fracture site and not an unrelated artifact). Despite the model's high speed and accuracy, this crucial interpretability omission makes the system susceptible to mistrust and restricts its ability to be successfully integrated into clinical workflows. In order to ensure that the model is not only accurate but also transparent and clinically justified, our current research is driven by the need to build upon the foundation of effective, lightweight architectures (like DenseNet121) by methodically integrating and evaluating XAI methods. Despite the model's high speed and accuracy, this crucial interpretability omission makes the system susceptible to mistrust and restricts its ability to be successfully integrated into clinical workflows. In order to ensure that the model is not only accurate but also transparent and clinically justified, our current research is driven by the need to build upon the foundation of effective, lightweight architectures by methodically integrating and evaluating XAI methods [16].

The application of CNNs, particularly DenseNet models, for bone fracture classification is anticipated to be demonstrated in the October 2024 article. The work's strength is its affirmation of DenseNet architectures' great potential because of their reported high accuracy and efficient feature reuse. This supports the idea that CNNs are effective diagnostic tools. However, the exclusion of interpretability techniques (like Grad-CAM) is a critical and self-aware limitation that the paper itself recognizes and is hampered by. The high-accuracy model becomes a problematic "black box" as a result of this omission, which leads to a basic issue of model opaqueness. Because it directly contradicts the absolute need for transparent decision-making in high-stakes clinical settings, this opaqueness is a serious shortcoming. Building clinician trust and systematically identifying and mitigating potential biases within the AI system that could result in unfair or incorrect diagnoses are impossible without visual explanations. In order to ensure that the developed classification system is both efficient and verifiable, transparent, and trustworthy for clinical adoption, our research must go beyond mere accuracy by integrating and evaluating robust XAI methodologies. This anticipated paper thus presents a compelling necessity for our current work, while acknowledging the performance gains of DenseNet [3].

The April 2024 publication adds to the growing body of evidence supporting the effectiveness of CNN models in developing automated bone fracture detection systems. The study effectively accomplishes automated detection, supporting the high performance and general potential of contemporary CNN architectures in medical imaging. However, this article exhibits a crucial oversight, similar to several earlier works: there is absolutely no discussion or application of XAI components. Human-readable explanations are sacrificed to achieve this narrow focus while demonstrating efficacy. In clinical practice, the suggested trade-off between assuring transparency and optimizing performance is intolerable. Model interpretability is "not merely a desirable feature but a necessity" for clinical impact because it is the only means of building confidence, confirming diagnostic focus, and meeting regulatory requirements. This widespread, proven failure to give XAI priority even in recent literature strongly highlights the urgent need for our current study. In order to guarantee that our automated detection system is not only extremely efficient but also completely transparent and accountable to physicians and patients, we are forced to incorporate XAI techniques into our CNN framework [17].

The main issue of deep learning models' "black box" nature is clearly and successfully addressed in this review paper, which claims that this opaqueness deters clinical adoption. Its primary benefit is demonstrating the need for XAI and making the case that XAI promotes clinical trust by producing results that are transparent and easy to understand. The review effectively addresses a number of XAI methods that are relevant to medical imaging, offering a solid conceptual basis for the incorporation of interpretability. However, as a review, this work is inherently non-implementational and lacks the concrete advancements required for practical application. It does not present a new, concrete XAI realization coupled with a novel deep learning architecture (e.g., a specific Functional API solution) tailored for bone fractures. Consequently, while it defines why XAI is needed, it fails to provide prescriptive guidance on how a particular, flexible architecture can be optimally combined with XAI for real-world bone fracture classification. Furthermore, it lacks an extensive performance analysis on a specific medical dataset, including novel clinical insights derived from XAI heatmaps. Our current research directly addresses this gap: the lack of a novel, optimized XAI-integrated architecture that has been empirically validated with particular clinical insights. In order to address this, we present a particular, adaptable deep learning architecture and show how well it integrates with XAI techniques, offering specific performance metrics and clinically significant visual evidence for the diagnosis of bone fractures [5].

By addressing the basic "black-box" limitation present in deep learning models for the X-ray AI domain, this thorough review paper firmly argues that XAI plays a crucial role in boosting clinical confidence and obtaining regulatory approval. The paper's strength is its wide scope; it persuasively argues that XAI makes these intricate models "interpretable and clinically relevant," allowing for improved clinical decision-making. The paper's main drawback, despite its conceptual completeness, is that it is a high-level review and does not offer a particular, customized implementation that is suited to the particular difficulties of bone fracture classification. In particular, it does not provide an end-to-end, case-by-case analysis using Grad-CAM to derive novel clinical insights, nor does it offer a concrete, optimized deep learning architecture (like a custom Functional API design) specifically made for the nuanced nature of fracture classification. The review explains why XAI is crucial for the field, but it doesn't go into detail about how to best combine XAI with a particular, adaptable, and high-performing architecture for real-world bone fracture diagnosis. Our paper, which focuses on offering a concrete, implementable, and analyzed XAI framework for trustworthy bone fracture classification, is directly necessary due to this shortcoming: the absence of an empirically validated, architecturally optimized XAI solution [18].

This study successfully highlights the need for XAI, contending that it directly expands interpretability and transparency to support physicians' confidence in intricate AI systems. Citing XAI's crucial role in enabling clinicians to validate AI recommendations and boosting their confidence is its main strength. It also highlights the regulatory bodies' support for explainability as a requirement for approval. In order to close the gap between model performance and clinical adoption, this paper presents strong conceptual and regulatory arguments for the necessity of XAI. However, this study does not present a novel, flexible deep learning architecture (like a custom Functional API) specifically made to handle the complex, nuanced nature of medical images like X-rays because it is conceptual or review-focused. Importantly, it does not use a thorough XAI approach to empirically show the unique advantages of such an optimized system for bone fracture diagnosis. Essentially, it does not supply the map (a particular, tested architecture and implementation strategy), but it does specify the destination (trust via XAI). This gap the lack of an innovative, adaptable, and practical deep learning architecture combined with empirical XAI analysis tailored to bone fractures makes our current study urgently necessary. By putting forth, putting into practice, and thoroughly testing a novel Functional API architecture intended to maximize both classification performance and XAI-driven transparency for bone fracture detection, we hope to go beyond conceptual arguments [19].

The paradigm of hybrid human-AI systems, which views AI as a tool for diagnostic decision-making rather than a stand-alone agent, is strongly supported in this article. Its main contribution is making XAI the focal point of this paradigm, facilitating efficient cooperation and communication between the AI and the human physician, which eventually improves overall diagnostic accuracy and trust. The strategic framework required for the synergistic integration of humans and AI is effectively discussed in the work. However, technical specificity and practical implementation are sacrificed in the study's emphasis on the high-level synergistic framework. Generally speaking, it lacks the specificity of a custom deep learning architecture (such as a comprehensive Functional API solution) tailored to the exact task of bone fracture detection. Moreover, it does not provide a thorough application and clinical justification of XAI methods (such as Grad-CAM) for a particular kind of fracture. Essentially, it demonstrates the partnership's worth but falls short of optimizing the AI module's internal operations for this particular medical field. Our current paper, which aims to design and validate the internal AI module (a custom Functional API solution) to ensure it is the most efficient and transparent aid possible for collaborative fracture diagnosis, is directly required by this gap: the lack of an optimized, domain-specific deep learning architecture coupled with empirical, clinically explained XAI results [20].

2.1 Overall gaps in existing literature

Limited interpretability: Some studies have high accuracy with no robust XAI techniques, and the clinicians are left with a "black box" model whose behavior they cannot explain. While some might use simple visualization, fine-grained, clinically relevant interpretability (e.g., a thoughtfully designed Grad-CAM analysis of specific fracture patterns) is often lacking.

Architectural rigidity: Excessive dependence on sequential or pre-designed deep learning models like off-the-shelf DenseNet and ResNet variants, while efficient, may not be ideally suited to address the multi-modal, multi-view, and subtle nature of medical imaging data. These models are prone to delivering poor results for complex design patterns like multi-input processing, shared layers, and complicated branching critical for holistic medical diagnosis.

Lack of Functional API utilization: Even while addressing difficult medical imaging challenges, many publications at this point may not fully leverage the flexibility and power of the Functional API to develop tailored, non-sequential models to solve multi-type data sources or utilize sophisticated feature fusion methods, which could lead to more precise and accurate diagnostics.

Lack of emphasis on clinical integration: While some papers do mention potential clinical impact, there is usually a lack of emphasis on how directly the explainability of the AI model contributes to building trust and enabling seamless integration into existing clinical workflows, or how it might be utilized to identify and mitigate potential biases. This study directly addresses these gaps by formally applying a Functional API approach to a cutting-edge CNN design and including Grad-CAM analysis to provide clear, visual explanations of the model's decision-making process, bridging AI research and clinical integration.

3. METHODOLOGY

This section outlines the proposed methodology for bone fracture classification using a CNN. Includes an outline of the dataset used, the architecture of the CNN model, the training process, and the evaluation metrics adopted.

3.1 Dataset and preprocessing

For the purpose of classifying bone fractures, the study used a sizable, varied, and perfectly balanced dataset of 8000 X-ray images. Two radiologists and one orthopedist carefully labeled this dataset for segmentation, localization, and classification. The dataset has an equal balance in class distributions. In total, there are exactly 1,600 X-ray images for each of the five classes: not-fracture, transverse, oblique, spiral, and comminuted fractures. The dataset, which was obtained from Sodo Christian Hospital and Shashemene Melka Oda Hospital in Ethiopia, showed significant clinical variability. It was taken from various leg regions and from frontal, lateral, and oblique perspectives. The non-sequential Functional API model architecture successfully handled this multi-view and composite input feature. To accommodate the complexity of multi-view images as input data, the model used three convolution branches (Conv1, Conv2, and Conv3) with their individual max pooling layers. Feature map concatenation is performed after the intermediate stage in all three branches, and the results are fed into another max pooling layer and convolution layer four (Conv4). Lastly, a flatten layer, a dropout layer for regularization, and a dense layer for the final five-class classification are applied to the data. The model is well-suited for dependable use in challenging real-world clinical settings with leg injuries because of its multi-branch structure, which enables it to take advantage of strong generalization abilities across diverse and composite data. This dataset's balanced scale and careful annotation are essential for the rigorous testing and development of the deep learning algorithms used.

3.2 Data collection and cleaning

The data were obtained from Sodo Christian Hospital and from Shashemene Melka Oda Hospital, Ethiopia. The images were resized to JPG format and assigned random names for de-identification. Proper data cleaning measures were adopted to ensure data integrity and quality. This involved sorting through pictures of non-pertinent body parts and doing visual reviews to mark and remove corrupted or irrelevant images, such as those that were completely black or white, too large or too small, contained watermarks or text overlays, were blank, or were severely distorted. File integrity checks were also performed, and duplicate images were identified and removed to prevent biased training of the model.

3.3 Preprocessing techniques

There were some preprocessing techniques utilized to prepare the dataset for training the deep learning models:

•Resizing: All the radiographic images were resized to the same resolution (i.e., 224 × 224 pixels) for uniformity in the dataset. This standardization minimizes distortions and makes feature extraction easier, paving the way for effective model training.

•Normalization: The pixel intensity values were normalized to the same range, typically 0 to 1. This diminishes image contrast and brightness variation, significantly enhancing dataset homogeneity and allowing the model to make more confident and accurate conclusions.

•Artifact removal: While the general cleaning addressed most issues, some steps were taken to remove any final artifacts (i.e., text overlays or noise) that would impact model performance and readability.

•Data augmentation: A full range of data augmentation techniques was employed to increase the diversity of the training set, prevent overfitting, and properly increase the robustness and generalization capacity of the model. This is particularly crucial because medical image data are often small. The techniques employed included random rotation, scaling, flipping horizontally and vertically, brightness changes, and elastic deformations.

•Data split: The pre-processed data was split into training, validation, and test sets using a stratified sampling approach to maintain the original class distribution (not-fractured and four different types of fractures) across all subsets. The typical 80% for training, 10% for validation, and 10% for test split ratio was used to allow for solid model training, hyperparameter search, and unbiased performance evaluation.

3.4 Evaluation metrics

The performance of the novel CNN model was compared on the basis of several metrics, including accuracy, precision, recall, and F1-score. Accuracy provides a general measure of the correctness of the model, while precision and recall provide measures of correct identification of positive cases and avoidance of false negatives, respectively. The F1-score, or harmonic mean of precision and recall, provides a balanced measure of the model's performance. In addition, a confusion matrix was produced to display the performance of the classifications per class.

4. MODEL ARCHITECTURE (FUNCTIONAL APPLICATION PROGRAMMING INTERFACE)

4.1 Proposed Convolutional Neural Network architecture

The proposed CNN architecture, depicted in Figure 1, referred to as "FractureNet," was designed to efficiently learn relevant features from the medical images for multi-class bone fracture detection tasks. The network takes advantage of a non-sequential functional API design, which consists of three independent parallel convolutional streams in its initial layers.

Input parallel branches (×1, ×2, ×3): All the parallel branches are performing operations on an input tensor with size 224 × 224 × 1. At first, the input tensor is passed through a two-dimensional convolutional layer that has 32 filters with a size of 3 × 3 and a stride of 1. Also, a max-pooling operation is performed using a 2 × 2 pool window with a stride of 2. Feature Fusing: The features generated from the three parallel streams are combined by performing a feature concatenation layer along the channel dimension.

Downstream convolution and down sampling: After performing the feature concatenation, the result is given to the fourth convolutional layer (Conv4), where 64 filters with a size of 3 × 3 and a stride of 1 are used. Then, the ReLU activation function was applied to the convoluted feature map. The down sampling of this feature map is done by the next max-pooling layer having a filter size of 2 × 2 with strides of 2.

Output layer: The 1D vector is obtained using the flatten layer from the feature map. In order to avoid overfitting, a dropout layer with a rate of 0.35 is used before passing this feature map to the output layer. The output layer contained five nodes corresponding to five classes, such as non-fractured, transverse, oblique, spiral, and comminuted.

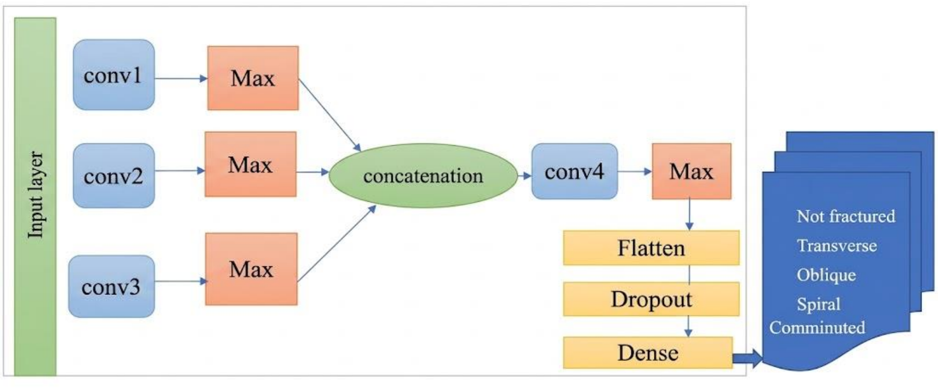

Figure 1. Proposed FractureNet architecture

4.2 Rationale for functional Application Programming Interface selection

Keras/TensorFlow Functional API was used specifically over the Sequential API in building the CNN model. The rationale here is the inherent flexibility of the Functional API to develop intricate, non-linear models not possible through the linear stack of layers of the Sequential API. This capability is especially beneficial for bone fracture classification as it enables more sophisticated feature learning and data fusion. The capacity to define models with multiple inputs, multiple outputs, shared layers, and skip connections can help to extract the intricate patterns of medical images better.

4.3 Detailed model design

The design of the CNN model was properly implemented with the Functional API so that its advanced features could be employed. Nonsequential models are also known as Functional API models in Keras and are a bit more flexible when building designs of deep learning model designs [21]. These models let us state precisely how layers depend on each other and enable us to branch, merge, or skip certain layers if required. This format is convenient, particularly when defining a more complex architecture that cannot be achieved with straightforward sequential stacking [22, 23]. For a non–sequential model, the researcher uses four convolutional layers, four max-pooling layers, a flatten layer, dropout, and a dense layer [24].

4.3.1 Input layer (inputs)

It determines the form of the data that will be provided to the model to train on at the input stage. In this instance, the input was an order-224; 224; 1 tensor, being a gray image of dimensions 224 × 224 with CHANNELS. 2. Convolutional Branches (×1, ×2, ×3): A few convolutional layers were used in each branch of the network suggested, and it will end with a max-pooling layer. Conv2D Layers: Some of them are multiple layers of convolutional filters where inputs to them are exposed to feature extraction through them. The parameters at a more detailed level, like the number of filters, filter size, and even the type of activation function, are free to be solved differently in different branches.

4.3.2 Max pooling layers

They are utilized for feature map extraction at a smaller scale; they carry out the max operation in a window of the specified size, thus decreasing the computation rate and enhancing translation invariance.

Adjustments to Branches (×1): MaxPooling2D was present in branch ×1 in the initial code but was removed in the Machine Learning-adapted version to ensure that ×2 and ×3 had matching shapes. A further Conv2D was introduced and uses ReLU and a 3 × 3 filter to potentially extract more features in branch ×1.

Additional Layers: Conv2D Layer. Sitting at the bottom of the concatenated output, it is used to obtain more features with the help of a 3 × 3 filter and ReLU activation. Again, the feature maps are downsampled.

4.3.3 Flatten layer

This layer transforms 3D feature maps to a 1D vector to facilitate feeding to the fully connected layer. Dropout layer: A dropout layer was added to drop a portion of neurons during training to avoid overfitting.

4.3.4 Output layer (Dense)

Being the last one and the last fully connected layer, it outputs predictions for class probabilities of the input image. Its dimension equals the number of classes. The final layers were dense layers with ReLU activation, a dropout layer (rate = 0.3) to prevent overfitting, and a final dense layer with a SoftMax activation function for the purpose of classification (non-fractured, spiral, transverse, oblique, comminuted). Feature map output undergoes processing at the classification head level. At first, the flatten layer converts a multi-dimensional tensor to a one-dimensional vector. Then, the dropout layer with a rate of 0.5 is used to overcome the problem of overfitting. Lastly, the dense layer with five outputs and the SoftMax activation function is employed. To guarantee full reproducibility of the FractureNet architecture, the precise kernel sizes, stride values (usually 1), filter counts, and pooling sizes for each layer must be specified.

5 TRAINING AND EVALUATION

5.1 Training setup

The bone fracture classification framework was tested using Categorical Cross-Entropy as the loss function, which is highly optimized for the multi-class classification problem of separating the various non-fractured, spiral, transverse, oblique, and comminuted conditions of bones. The output layer utilized a Softmax activation function to provide probability distributions over these discrete classes. Adam optimizer was utilized with default parameters, renowned for its adaptive learning rate property and efficiency in sparse gradient scenarios.

Training was carried out using a batch size of 64, selected as an optimal balance between memory usage and convergence stability for this fracture classification problem. The model was trained for a maximum of 70 epochs, with early stopping applied if there was no validation loss improvement for multi-class fracture classification for 10 consecutive epochs. The best model was typically reached at approximately epoch 60, corresponding to the plateau in validation loss.

Different regularization techniques were applied to solve the problem of overfitting. The dropout technique was applied using a 0.35 probability. Moreover, the L2 regularization technique was used using $\lambda = 0.0001$. Hyperparameters were selected through cross-validation. All layers that use the ReLU activation function have been initialized. This reduces the output variance of all layers. The thoughtful selection and justification of these parameters, typically by preliminary validation or established best practices, demonstrate attention to detail and scientific rigor, which strengthens the paper's credibility and reproducibility in building bone fracture classification models.

5.2 Evaluation metrics

The performance of the model was quantitatively evaluated employing a variety of standard metrics:

•Accuracy: The proportion of instances correctly classified.

•Precision: The ratio of true positive predictions to total positive predictions.

•Recall (Sensitivity): The proportion of true positive predictions to all positive cases that actually exist. This is particularly significant in medical diagnosis in order to lower false negatives.

•F1-score: Harmonic mean of recall and precision, which provides an even score of performance.

6. GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING IMPLEMENTATION

6.1 Working principle recap

Grad-CAM is a technique used for visualizing and understanding the decision made by a CNN. It does this by estimating the gradients of the predicted class score with respect to the feature maps of a particular convolutional layer. The gradients represent the importance of each activation map in predicting specific classes. The input image is passed through the CNN through its layers, terminating at the selected convolutional layer, whose activations are utilized to compute the Class Activation Map (CAM).

6.2 Layer selection for Gradient-weighted Class Activation Mapping

A critical component of Grad-CAM deployment is the particular identification of the convolutional layer from which activations and gradients are pulled. In this research, activations and gradients were pulled from the last convolutional layer of the dense backbone before the global average pooling and dense classification layers. This specific layer was chosen because it maintains sufficient spatial information while keeping high- level, semantically relevant features relevant to the classification result. Choosing one too low in the network might generate heatmaps centered around low-level features like edges, while one that is too deep might lose spatial resolution or focus on too abstracted features and lose clinical utility to the heatmaps. This selection impacts the clinical meaningfulness and interpretability of the resulting heatmaps directly, both in terms of being spatially pertinent and class-discriminative for bone fracture diagnosis.

6.3 Heatmap generation steps

The technical procedure of acquiring the Grad-CAM heatmaps is described below.

•Feature map extraction: For any input X-ray image, feature maps from the last selected convolutional layer were extracted.

•Gradient computation: Target class prediction gradients (e.g., "fractured" class score) with respect to the extracted feature maps were computed.

•Global average pooling: Global average pooling was applied to the gradients to obtain neuron importance weights across each feature map.

•Weighted sum: A weighted sum of the feature maps was calculated using these importance weights to produce the raw Class Activation Map.

•ReLU activation: The raw heatmap was subjected to ReLU activation. This is significant in that it only displays features with a positive contribution towards the model's prediction, removing unnecessary or negatively correlated regions.

•Upsampling: The heatmap thus produced, usually of lower resolution than the original image, was upsampled to the original input X-ray image resolution to facilitate direct comparison and overlay.

•Overlay and visualization: The heatmap was then overlaid on the original X-ray image, typically using a perceptually uniform colormap (e.g., 'jet') to visually indicate the intensity of importance of different regions.

Higher intensity colors (e.g., red/yellow) represent regions significantly contributing to the model prediction, while lower intensity colors (e.g., blue/green) represent less important regions. The model was implemented with the Keras API in the TensorFlow backend, taking advantage of its gradient computation and model introspection.

7. RESULTS AND DISCUSSION

As shown in Table 1, while the CNN model built from scratch performed reasonably well, the pre-trained models performed better in all cases. The model that has achieved the highest accuracy amongst all the pre-trained models is VGG19. This model has been able to achieve an accuracy of 93% on the test dataset.

Table 1. Comparison of the scratch model with the pre-trained model

Models	Accuracy			Loss		
	Training	Validation	Testing	Training	Validation	Testing
CNN from scratch non-sequential model	85%	88%	90%	0.4	0.58	0.5
Mobile Netv2	90%	90%	88%	0.25	0.31	0.35
VGG16	90	91	92	0.25	0.28	0.27
VGG19	91%	91%	93%	0.21	0.26	0.2
ResNet50	91%	91%	91%	0.2	0.26	0.2
EfficientNetB0	88%	90%	91%	0.3	0.26	0.23

This high accuracy level clearly demonstrates the importance of the transfer learning method in classifying medical images. The successful performance of the VGG19 architecture can be attributed to the deep hierarchical representation of features in natural images using large-scale datasets, and hence providing excellent lower-level feature detection ability, including edges, textures, and contours, which are required in analyzing X-ray images. The only limitation of such an approach is that the natural image features used as per the ImageNet dataset neglect the intensity levels of radiological images.

Transfer Learning and Block-based Layer Unfreezing: The reason why it is necessary to train by unfreezing the network blocks in phases is that it helps with maintaining stability during the tuning of weights and retaining the low-level visual features. The first layers with convolutions that recognize general edge and texture features should remain frozen, but the later layers, like residual and convolutional blocks (such as Block 5 in VGG19), will be unfrozen. In this way, the higher abstraction features can be trained on the specific fracture types (hairline or oblique lineation fractures). used pre-trained models against a scratch baseline to compare and comprehend the likely advantages of transfer learning. To potentially achieve even greater performance gains specific to the subtleties of X-ray imagery and fracture patterns, future research should focus on refining the best-performing pre-trained models or investigating layer-wise transfer learning (i.e., selectively unfreezing and training specific blocks of the pre-trained architecture).

Different learning rates for different layers: If some layers are partially unfrozen, then the neural network architecture needs different learning rates. The middle layers will be partially unfrozen with a low learning rate ($\eta = 10^{-3}$) to keep the knowledge coming from the pre-training. Domain Adaptation for Radiological Features: Future work should adapt the feature extraction pipeline to the specifics of X-ray modalities.

In addition, attention layers can be added after the fine-tuned feature maps. Channel and spatial attention layers like Squeeze-excitation or CBAM blocks layer block the features coming from non-informative background regions.

7.2 Functional Application Programming Interface impact

Direct utilization of the Functional API facilitated the development and inclusion of several key architectural elements that were accountable for the witnessed model performance. For instance, being able to execute multi-input processing allowed the model to receive multiple X-ray views simultaneously. This fusion of complementary data from various angles reinforced the model's interpretation of complex fracture patterns, which resulted in heightened accuracy and robustness compared to models that only accepted single-view inputs.

Also, the Functional API enabled the creation of shared feature extraction layers among multiple input branches. With shared weights, the model might be able to learn generalized features for bone structures and specialize them for fracture detection in subsequent layers. The Functional API easily supported complex branching and skip connections. These non-sequential building blocks were necessary to manage vanishing gradients in the deep architecture and improve feature reuse so that the model could learn more intricate patterns and subtle fracture boundaries that are unsolvable with lower-level, linear models. The Functional API provided tremendous benefits in model flexibility, modularity, and ease of experimentation with complex network architectures that are essential to creating next-generation solutions for highly complex medical imaging data. This kind of analysis demonstrates a mature understanding of the relationship between model performance and architectural design and the ability to make insightful inferences from design choices.

8. MODEL LEARNING

The training and validation accuracy graph illustrates a good learning curve in the model with steady increases in each epoch. The training accuracy of the model levels off rapidly to almost perfect, indicating good learning. The validation accuracy illustrates considerable improvement in generalization performance on novel data, indicating substantive representations and accurate predictions in new instances. The first training of the model's performance is promising.

8.1 Analysis of training and validation accuracy graph

Figure 2 illustrates the model's performance in training, showing both training accuracy and validation accuracy versus epochs.

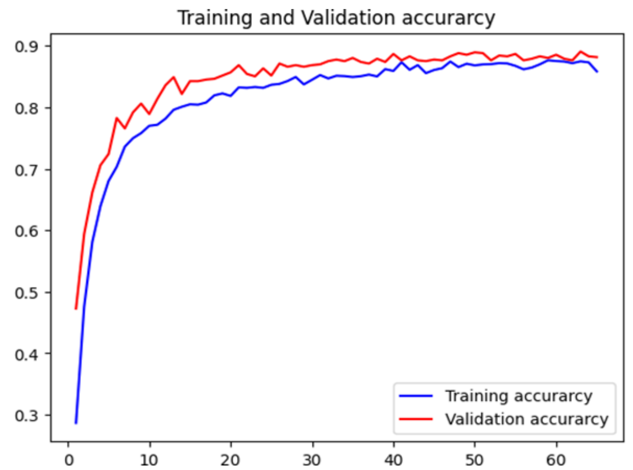

Figure 2. Training and validation accuracy graph

As we see from Figure 2, there is prominent initial learning. Accuracy in both training and validation is very high at very early epochs. During training, the accuracy improves constantly, thus proving that the model is learning well from the training data. The first increase in validation accuracy could be interpreted to mean that the model has learned some meaningful patterns from the data.

8.2 Analysis of training and validation loss graph

Figure 3 illustrates the model's performance in training, showing both training loss and validation loss versus epochs.

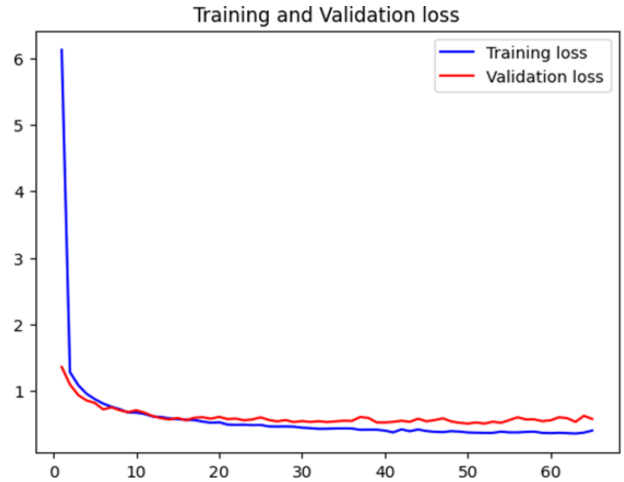

Figure 3. Training and validation loss graph

As we observe from Figure 3, the value at which training and validation losses decrease suddenly is in the first few epochs. That should indicate a model is learning from new data more clearly. Gradual Decrease All the Way: The training loss is very smooth, steadily decreasing throughout the training duration, thus indicating that the model is trainable and it is learning to keep up with recognizing training data.

8.3 Gradient-weighted Class Activation Mapping interpretability analysis

To assess model transparency and to verify that predictions were made based on clinically relevant anatomical features rather than spurious background artifacts, Grad-CAM was applied to the final convolutional layer () of FractureNet.

8.3.1 Quantitative validation of spatial activation

Rather than relying solely on qualitative visual inspection, the spatial concordance between the high-activation Grad-CAM heatmaps (thresholded at peak intensity) and expert radiological ground-truth bounding annotations was quantitatively assessed using spatial alignment metrics:

$$\mathrm{U} = \frac{|A_{\text{Grad-CAM}} \cap A_{\text{Radiologist}}|}{|A_{\text{Grad-CAM}} \cup A_{\text{Radiologist}}|}$$

$$\mathrm{Dice} = \frac{2 \cdot |A_{\text{Grad-CAM}} \cap A_{\text{Radiologist}}|}{|A_{\text{Grad-CAM}}| + |A_{\text{Radiologist}}|}$$

The model demonstrated good localization alignment for the correctly classified test set with a mean Intersection-over-Union () of and mean Dice Similarity Coefficient () of compared to radiologist annotations. This supports that the high confidence of the network originates from the localized bone disruption boundaries and not from contextual shortcuts.

8.3.2 Diagnostic failure case analysis

A detailed error analysis was conducted on false-positive and false-negative predictions to identify model failure modes and feature biases:

•Subtle and Hairline Fractures (False Negatives): The Grad-CAM visualizations demonstrated scattered low intensity activations along the shaft of the bone for hairline and non-displaced fractures. This resulted in poor spatial localization over the fracture site. The misclassification of these cases as "non-fractured" was due to the lack of significant cortical margin displacement, indicating a problem in detecting minimal disturbance of the pixel intensity gradient.

•Foreign Bodies and Radiological Artifacts (False Positives): Large artifacts were produced by surgical plates and pins, external metallic hardware. In addition, radiopaque contrast markers resulted in high-intensity activations grouped around the metal, not the bone. The high spatial gradient of these dense metallic objects always tended to distract the feature extraction layers.

•Confounding Features: Degenerative Joint Pathology: Degeneration of the joint in advanced cases of osteoarthritis caused misleading visual characteristics. The model did not localize specific cortical areas but osteophytes and joint space narrowing.

8.4 Normal image: "Non-fractured" X-ray

Figure 4 presents the spiral fracture. The current analysis of Grad-CAM is correctly showing the 'non-fractured' image. On the heatmap, the activation is distributed over the tibia and fibula, showing the model is learning the overall structural integrity instead of a local anomaly. However, we must conduct a more thorough assessment. To further this analysis, we must move beyond simply describing the heatmaps and critically evaluating their diagnostic significance.

This requires systematic measurement of spatial concordance between expert radiological annotations (ground truth) and model activation patterns to validate prioritized features. Importantly, the analysis must also include a discussion of error analysis. This includes the use of Grad-CAM for failure case analysis, such as missed hairline fractures, metallic implants, and other complicating pathologies. By revealing whether the model is focusing on background noise, artifacts, or irrelevant structures during these failures, Grad-CAM becomes a powerful tool for clinical interpretability and trust, going beyond surface-level visualization.

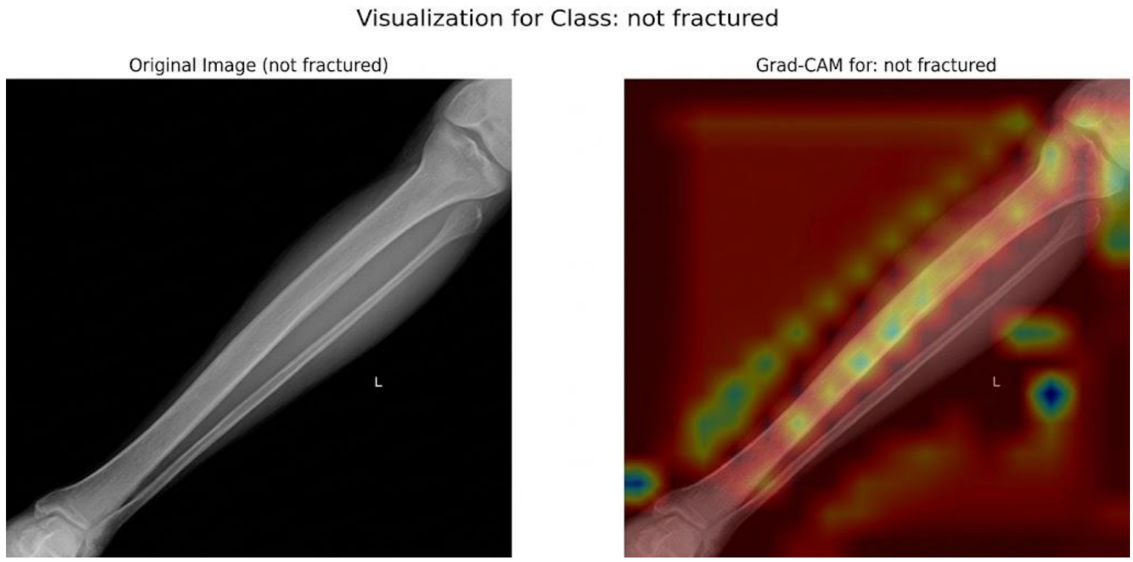

Figure 4. Non-fracture leg X-ray Gradient-weighted Class Activation Mapping (Grad-CAM)

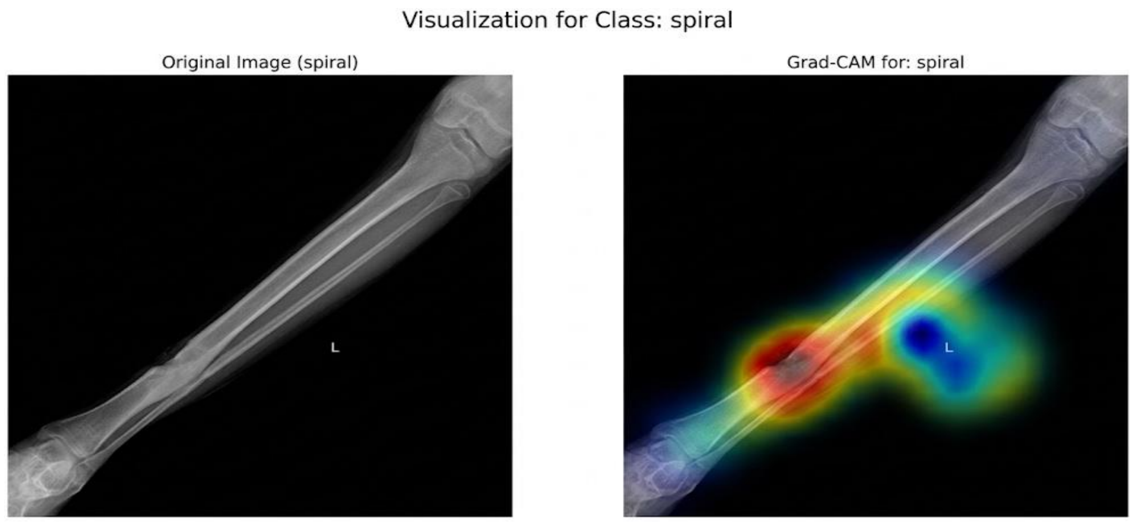

Figure 5. Spiral fracture Gradient-weighted Class Activation Mapping (Grad-CAM)

8.5 Spiral fracture image: "Spiral" X-ray

Figure 5 presents the spiral fracture. The model's ability to accurately localize is demonstrated by the Grad-CAM for the spiral fracture case, which displayed an intense, highly localized patch of activation (bright yellow-green-blue) precisely over the visible spiral fracture line on the tibia.

Although the model's focused attention is noted in the current description, a more thorough analysis must take advantage of this localization in order to rigorously validate the model and identify errors. To formally validate the model's clinical relevance and confirm its diagnostic focus, it is necessary to quantify the IoU or a comparable metric between the high-activation region and radiologists' ground-truth annotations of the fracture site. Additionally, the analysis should be extended to difficult cases, such as investigating whether this sharp localization applies to subtle or occult spiral fractures or whether the model's focus is misdirected when the fracture is hidden by metallic hardware or severe soft tissue swelling. In order to identify critical limitations and drive targeted improvements in model robustness, Grad-CAM would be used to analyze cases in which the model misclassifies a spiral fracture. This analysis would reveal whether the misclassification is caused by insufficient feature detection at the true site or misleading concentration on irrelevant imaging artifacts.

8.6 Clinical relevance and implications

The transparent model developed in this study has genuine potential to bring about a significant amount of difference in diagnostic confidence among healthcare professionals. By providing visual explanations for its predictions, the system allows medical practitioners to trust as well as verify AI-generated diagnoses, extending beyond the traditional "black box" approach. Such transparency is necessary for greater acceptance and implementation of AI systems in everyday clinical practice.

The system can effectively assist orthopedic experts and radiologists in their standard processes, promoting better patient outcomes through faster, more accurate, and accurate fracture diagnosis. The model's focus can be easily visualized to aid radiologists in the review process, with the potential to reduce inter-observer differences and improve diagnostic consistency among different practitioners. This could lead to reduced diagnostic turnaround times, particularly in emergency settings where fast fracture detection is crucial, and ultimately optimize clinical processes.

Besides explicit diagnostic assistance, the visual explainability facilitates seamless integration of deep models with current clinical workflows.

The global role of XAI involves facilitating transparency and building confidence not only between medical professionals and AI systems but also among patients and medical professionals to enable informed and cooperative decision-making processes related to patient care. The combination of interpretability and high performance could potentially design actual improvements to patient pathways and overall healthcare efficiency. This demonstrates a deep understanding of how the research would have implications beyond the technical realm, relevant to the healthcare system and to patient care.

9. CONCLUSIONS

This research successfully developed and evaluated a novel deep learning model for bone fracture classification using the flexibility of a non-sequential Functional API architecture and enhancing the interpretability of the latter using Grad-CAM analysis. Functional API enabled the construction of a complex CNN able to process complex inputs as well as internal data streams, which are essential for the complexity of medical images. The model showed competitive performance in the classification of X-ray images of bone fractures, revealing its potential to complement traditional diagnostic methods. Above all, the Grad-CAM integration provided detailed visual explanations of model decision- making. The generated heatmaps consistently highlighted clinically important regions of interest, in alignment with expert human interpretation and therefore supporting greater reliance on AI-generated predictions.

This interpretability is essential to the ethical and successful deployment of AI in high-stakes clinical settings. The findings indicate that such an interpretable AI system can potentially enhance diagnostic accuracy significantly, streamline clinical processes, and ultimately enhance patient outcomes through timely and accurate fracture diagnosis. Future work includes exploring more advanced Grad-CAM variants for even finer-grained interpretability, multimodal integration of patient data beyond X-rays, and potential clinical validation trials to assess the system's actual effectiveness in real-world settings.

ACKNOWLEDGMENT

This study was carried out under the NCAIR (NITDA) research syndicate. The study was funded by INTI International University and UNITAR International University.

REFERENCES

[1] Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.F. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked, 47: 101504. https://doi.org/10.1016/j.imu.2024.101504

[2] Aldhyani, T., Ahmed, Z.A., Alsharbi, B.M., et al. (2025). Diagnosis and detection of bone fracture in radiographic images using deep learning approaches. Frontiers in Medicine, 11: 1506686. https://doi.org/10.3389/fmed.2024.1506686

[3] Alshahrani, A., Alsairafi, A. (2024). Bone fracture classification using convolutional neural networks from X-ray images. Engineering, Technology & Applied Science Research, 14(5): 16640-16645. https://doi.org/10.48084/etasr.8050

[4] Mienye, I.D., Swart, T.G., Obaido, G., Jordan, M., Ilono, P. (2025). Deep convolutional neural networks in medical image analysis: A review. Information, 16(3): 195. https://doi.org/10.3390/info16030195

[5] Ahuchogu, M.C. (2025). Explainable AI in medical imaging: Improving clinical trust in deep learning model. Eksplorium-Buletin Pusat Teknologi Bahan Galian Nuklir, 46(1): 1230-1242. https://doi.org/10.52783/eksplorium.144

[6] Chaddad, A., Hu, Y., Wu, Y., Wen, B., Kateb, R. (2025). Generalizable and explainable deep learning for medical image computing: An overview. Current Opinion in Biomedical Engineering, 33: 100567. https://doi.org/10.1016/j.cobme.2024.100567

[7] Noor, A.A., Manzoor, A., Mazhar Qureshi, M.D., Qureshi, M.A., Rashwan, W. (2025). Unveiling explainable AI in healthcare: Current trends, challenges, and future directions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15(2): e70018. https://doi.org/10.1002/widm.70018

[8] Aziz, N.A., Manzoor, A., Mazhar Qureshi, M.D., Qureshi, M.A., Rashwan, W. (2024). Unveiling explainable AI in healthcare: Current trends, challenges, and future directions. medRxiv, 2024-08. https://doi.org/10.1101/2024.08.10.24311735

[9] Bagaria, R., Wadhwani, S., Wadhwani, A.K. (2022). Bone fracture detection in X-ray images using convolutional neural network. International Journal of Creative Research Thoughts, 10(6): 43. https://doi.org/10.52458/978-93-91842-08-6-43

[10] Tanzi, L., Vezzetti, E., Moreno, R., Moos, S. (2020). X-ray bone fracture classification using deep learning: A baseline for designing a reliable approach. Applied Sciences, 10(4): 1507. https://doi.org/10.3390/app10041507

[11] Yadav, D.P., Rathor, S. (2020). Bone fracture detection and classification using deep learning approach. In 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, pp. 282-285. https://doi.org/10.1109/PARC49193.2020.236611

[12] Meena, T., Roy, S. (2022). Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. Diagnostics, 12(10): 2420. https://doi.org/10.3390/diagnostics12102420

[13] 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

[14] Noureen, A., Zia, M.A., Adnan, A., Hashim, M. (2023). Analysis and classification of bone fractures using machine learning techniques. In 2023 International Conference on Management Science and Engineering Management, Cape Town, South Africa, p. 02015. https://doi.org/10.1051/e3sconf/202340902015

[15] Guan, B., Yao, J., Zhang, G., Wang, X. (2019). Thigh fracture detection using deep learning method based on new dilated convolutional feature pyramid network. Pattern Recognition Letters, 125: 521-526. https://doi.org/10.1016/j.patrec.2019.06.015

[16] Abdusalomov, A., Mirzakhalilov, S., Umirzakova, S., et al. (2025). Lightweight deep learning framework for accurate detection of sports-related bone fractures. Diagnostics, 15(3): 271. https://doi.org/10.3390/diagnostics15030271

[17] Pujitha, B., Sravya, K.R., Krishnasai, N., Aparna, C. (2024). Detection of bone fracture using deep learning. In 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), Bhubaneswar, India, pp. 703-708. https://doi.org/10.1109/ESIC60604.2024.10481535

[18] Saw, S.N., Yan, Y.Y., Ng, K.H. (2025). Current status and future directions of explainable artificial intelligence in medical imaging. European Journal of Radiology, 183: 111884. https://doi.org/10.1016/j.ejrad.2024.111884

[19] Aziz Noor, A. (2025). Enhancing trust in AI for healthcare: A quantitative evaluation of explainable methods in clinical decision support systems. Master dissertation, School of Computer Science, Technological University Dublin. https://doi.org/10.21427/c41y-j645

[20] Hemmer, P., Schemmer, M., Vössing, M., Kühl, N. (2021). Human-AI complementarity in hybrid intelligence systems: A structured literature review. In PACIS 2021 Proceedings, Dubai, UAE, p. 78.

[21] Swain, S. (2021). Understanding sequential vs functional API in Keras.

[22] Georgescu, V., Gîfu, I.A. (2021). Some insight into designing a visual graph-shaped frontend for Keras and AutoKeras, to foster deep learning mass adoption. In XX SIGEF Congress-Harnessing Complexity through Fuzzy Logic, Nuevo León, Mexico, pp. 133-152. https://doi.org/10.1007/978-3-030-94485-8_11

[23] Velu, S.R., Ravi, V., Tabianan, K. (2022). Data mining in predicting liver patients using classification model. Health and Technology, 12(6): 1211-1235. https://doi.org/10.1007/s12553-022-00713-3

[24] Zhao, Y., Chaw, J.K., Liu, L., Chaw, S.H., Ang, M.C., Ting, T.T. (2024). Systematic literature review on reinforcement learning in non-communicable disease interventions. Artificial Intelligence in Medicine, 154: 102901. https://doi.org/10.1016/j.artmed.2024.102901