

XrayBoneNet: Multi-Region Fracture Localization and OA Grading via Attention-Guided Transformers

Nandkishor Narkhede*

Department of Electronics and Computer Science, Shah and Anchor Kutchhi Engineering College, Chembur, Mumbai – 400088, India

E-mail: nandkishor.narkhede@sakec.ac.in

ORCID iD: <https://orcid.org/0000-0002-4324-1950>

*Corresponding author

D. Girish Kumar

Department of Electronics and Communication Engineering, Shri Vishnu Engineering College for Women, Bhimavaram – 534202, India

E-mail: girishkumar.d@svecw.edu.in

ORCID iD: <https://orcid.org/0000-0002-2430-2834>

Dhanya Job

Department of Computer Applications, Baslios Poulouse Second Catholicose College, Piravom, Ernakulam, Kerala, India

E-mail: dhanyajob@bpccollege.ac.in

ORCID iD: <https://orcid.org/0000-0003-2746-4430>

K. Varada Rajkumar

Department of Computer Science and Engineering (AIML), M L R Institute of Technology, Hyderabad, Telangana, India

E-mail: varadas18@gmail.com

ORCID iD: <https://orcid.org/0000-0001-5046-1276>

Tirumalasetti Lakshmi Narayana

Department of Electrical and Electronics Engineering, Aditya University, Surampalem, Andhra Pradesh – 533437, India

E-mail: tlaxman17@gmail.com

ORCID iD: <https://orcid.org/0009-0000-5359-6723>

B. Loveswara Rao

Department of Electrical and Electronics Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh – 522302, India

E-mail: loveswararao@kluniversity.in

ORCID iD: <https://orcid.org/0000-0002-1094-2208>

Received: January 23, 2026; Revised: February 20, 2026; Accepted: March 24, 2026; Published: August 8, 2026

Abstract: It is still hard to accurately find bone fractures and figure out how bad osteoarthritis (OA) is from X-ray pictures because of the complicated anatomical differences and the lack of contextual modelling in standard deep learning methods. This research presents XrayBoneNet, a hybrid deep learning system that combines Convolutional Neural Networks (CNNs), Transformer-based global feature modelling, and attention mechanisms for concurrent fracture identification, OA grading, and localization. The model has two heads: one for binary fracture classification and one for multi-class OA staging. It also has a bounding box regression head for accurate localization. To improve training efficiency and performance, a hybrid optimization technique that uses Bighorn Sheep Optimization (BSO) for global exploration and Logarithmic Mean Optimization (LMO) for fine-tuning is developed. The Bone Fracture Multi-Region X-ray dataset shows that XrayBoneNet works better than state-of-the-art models like ResNet50, DenseNet121, and Vision Transformer. It has 96.8% accuracy in fracture detection, 94.5% accuracy in OA classification, and an

Intersection-over-Union (IoU) of 0.87 for localization. The suggested system offers an effective and comprehensible alternative for automated radiological diagnosis.

Index Terms: Bone Fracture; Multi-Region X-ray Dataset; Bighorn Sheep Optimization; Logarithmic Mean Optimization; Attention mechanisms; Intersection-over-Union.

1. Introduction

An important yet difficult challenge in medical imaging is the correct detection of bone fractures and osteoarthritis (OA) using skeletal X-rays [1]. Manual interpretation by radiologists is the backbone of conventional techniques, although this method is laborious, subjective, and susceptible to fluctuation amongst observers [9]. The development of AI, and DL in particular, has led to the promise of automated systems for radiographic diagnosis [2]. For problems involving classification of images, CNNs have become the gold standard. Chest X-ray analysis, brain tumour segmentation, besides mammography classification are just a few examples of the many medical imaging tasks that have seen heavy use of models like VGGNet, ResNet, DenseNet, and Inception [3]. The capacity to extract hierarchical features is key of ResNet-50 besides DenseNet-121's reasonable success in bone fracture diagnosis [4]. The inability of these models to grasp long-range connections is a major drawback when trying to decipher intricate bone structures regions of a picture [5]. Thanks to their worldwide attention processes, transformer models first popularised by the Vision Transformer (ViT)—have lately found a home in medical imaging [6]. By simulating paired connections between patches, these models circumvent CNNs' spatial locality constraints [7]. Hybrid models that integrate CNNs with encoders based on Transformers have become popular as a solution to the problems caused by the high computational overhead and huge data requirements of ViT and its derivatives, which limit their use in clinical settings where annotated data is scarce [8]. Medical segmentation accuracy can be enhanced by combining local CNN characteristics with global Transformer attention, as demonstrated by TransUNet and CoTr. But very few researches have used this idea to simultaneously stage OA and detect X-ray fractures [10].

The evolution of deep learning in medical imaging is rooted in seminal works such as CNN, which demonstrated remarkable performance in image recognition tasks [27–29]. Subsequently, attention mechanisms and Transformer architectures introduced a paradigm shift by enabling global contextual modelling [30]. In medical imaging, architectures such as U-Net have been widely adopted for segmentation and diagnostic tasks due to their ability to capture multi-scale features. These foundational contributions have significantly influenced the development of modern hybrid architectures.

From an optimization standpoint, the majority of current models employ inefficient methods in high-dimensional hyperparameter spaces, such as grid search or random search, or depend on manually adjusting hyperparameters [11]. Despite their utility, tuning methods such as Bayesian Optimisation and Tree-structured Parzen Estimators (TPE) can be laborious and vulnerable to local minima [12]. Genetic methods, Ant Colony Optimisation, and Particle Swarm Optimisation are a few examples of metaheuristic optimisation methods that have been used for hyperparameter tuning in certain recent studies [13]. Even if they work, these strategies usually sacrifice exploration efficiency for convergence efficiency and need a lot of computing power. In addition, the severity grading of osteoarthritis is not taken into account by the majority of current fracture detection models, which are merely capable of binary classification (fractured vs. non-fractured) [14]. And even when they do localization, models that rely on architectures like Faster R-CNN or YOLO aren't always suitable for lightweight deployment. A unified, efficient, and interpretable architecture that can do the following is clearly required due to these limitations:

1. Handling multiple anatomical regions across different views [15].
2. Simultaneously performing fracture detection, OA staging, and localization.
3. Efficiently navigating the hyperparameter space to improve training and convergence [16].

Novelty and Contribution Clarification:

Recent medical image processing research has employed hybrid CNN–Transformer architectures, however XrayBoneNet provides several advantages. The proposed framework is a multi-region unified learning system that can extract information from varied skeletal X-ray inputs, unlike typical models that focus on one anatomical area or job. Second, the architecture detects fractures, grades osteoarthritis, and localizes geographically without task-specific models using a multi-task dual-head design. Third, a novel attention-guided token-level fusion method enhances local CNN feature-global Transformer representation interaction to promote contextual comprehension across anatomical regions. A hybrid hyperparameter optimization strategy (BSO + LMO) addresses conventional tuning limitations to create a balanced exploration–exploitation mechanism for deep hybrid architectures. These contributions outperform CNN–Transformer combinations and improve diagnostic efficiency, comprehensiveness, and clinical utility.

1.1. Proposed Contribution Statement

This research introduces XrayBoneNet, a hybrid deep learning framework specifically designed to automate the detection of bone fractures and the classification of osteoarthritis severity using skeletal X-ray images. The primary contribution lies in the novel combination of CNN-based feature extractors with Transformer encoders and selective attention mechanisms, enabling the model to learn both localized and global contextual cues across multiple anatomical views. To further optimize performance, a Hybrid Optimization strategy is proposed, combining Bighorn Sheep Optimization (BSO) for broad hyperparameter space exploration and Logarithmic Mean Optimization (LMO) for precise convergence. This enables efficient and accurate tuning of critical training parameters, including learning rate, dropout, and network depth. The model employs a dual-head architecture: one head for binary fracture detection and another for multi-class osteoarthritis grading, supported by a bounding box regression head for localization. Evaluations on the Bone Fracture Multi-Region X-ray Dataset show superior performance in terms of classification accuracy, IoU for localization, and model convergence time. The proposed framework advances automated diagnostic intelligence in medical imaging by addressing both classification and spatial localization, demonstrating potential for deployment in clinical decision support systems, especially in resource-constrained healthcare environments. The rest of the paper is organized as follows: Section 2 mentions the related works; Section 3 provides the proposed methodology in detailed; Section 4 discuss the result analysis and finally, the conclusion is made at Section 5.

2. Related Works

An end-to-end DL framework called FracNet has been proposed by Alwzwy et al., [17] for the purpose of bone fracture diagnosis. The framework makes use of attention processes, feature selection, self-supervised pretraining, and enhanced visualisation capabilities. FracNet routinely surpasses state-of-the-art approaches in dependability and accuracy, with a detection accuracy of 100% on three datasets. Also, by giving doctors comprehensive explanations of its predictions, FracNet promotes decision transparency, which makes it a helpful tool. With FracNet, you can easily train it to use new datasets, and it's quite adaptable. Although fracture identification is FracNet's main focus, it may be easily scaled to other medical imaging applications. For feature engineering, Alam et al. [18] present MobLG-Net, a new method based on transfer learning. In order to generate class probability characteristics, the spatial features are first extracted from bone X-ray images using the MobileNet transfer model and then fed into a tree-based light gradient boosting machine (LGBM) model. The results are compared after applying several ML algorithms to subsets of the newly created transfer features. The novel features are used in conjunction with optimised hyperparameters to perform K-nearest neighbour (KNN), logistic regression (LR), random forest (RF), and logistic kernel boosting (LGBM). Predicting bone fractures with a 99% accuracy rate was achieved by the LGBM and LR models trained on proposed MobLG-Net (MobileNet-LGBM) based features, which outperformed others. Each model's performance is assessed using a cross-validation procedure. Bone fractures can be better detected utilising X-ray pictures according to the suggested study. The extensive X-ray dataset utilised by Wei et al., [19] encompasses both fractured and non-fractured instances involving the lower and upper limbs. Training accounted for 70% of the dataset, validation for 15%, and testing for 15%. A total of 10,966 cases were included in the training set, with 5,188 being normal and 1,112. In contrast, the validation and test sets each had 2,350 cases, with 1,112. Fracture localisation and classification were taught using a YOLOv11-based multi-task learning model. To avoid overfitting and boost generalisability, data augmentation was used. Two metrics, mean Average Precision (mAP) and Intersection over Union (IoU), were used to assess the model's performance. It was compared to two other models, Faster R-CNN besides SSD. In order to achieve greater convergence during training, the Adam optimiser was used with a batch size of 16 and a learning rate of 0.001. To also compared the YOLOv11 model to Faster R-CNN and SSD in order to measure its performance at various thresholds using mAP and IoU scores. With an IoU threshold of 0.5 and an IoU of 92.5%, the YOLOv11 model achieved great results with a mean Average Precision (mAP) of 96.8%. Outperforming Faster R-CNN (mAP: 87.5%, IoU: 85.23%) and SSD (mAP: 82.9%, IoU: 80.12%) in fracture identification and localisation, YOLOv11 achieved better results. This enhancement demonstrates the power and effectiveness of the model when used in real-time.

In their innovative fracture detection scheme, Abdusalomov et al. [20] used DenseNet121 architecture and optimised feature extraction by modifying first convolutional block and final layers. To additional improve the model's capability to identify localised structural discontinuities, a Canny edge detector was also included. To ensure robust model performance, preprocessing approaches such as contrast enhancement, normalisation, and data augmentation were used to a custom-curated dataset of radiography images focused on frequent sports-related fractures. Metrics like computational complexity, accuracy, recall, and precision were used to compare the model against state-of-the-art approaches. The results show that suggested model outperformed state-of-the-art benchmarks as ResNet-50, VGG-16, besides EfficientNet-B0, reaching accuracy of 90.3%. It kept the smallest computational complexity (FLOPs: 0.54 G, Params: 14.78 M) while demonstrating greater sensitivity (recall: 0.89) besides specificity (precision: 0.875). The consequences suitable for use in real-time clinical settings. Using Kaggle's multi-region X-ray dataset—a complete set of 10,580 radiographic images Aldhyani et al., [21] automated method for fracture identification. This research supports the use of DL methods for bone fracture identification and diagnosis, specifically VGG16, ResNet152V2, and DenseNet201. The results different kinds of fractures can be correctly identified and classified using the suggested

method. During validation, the system—which uses DenseNet201 and VGG16—achieved a 97% accuracy rate. Further improvement of DL models for fracture diagnosis can be achieved by tackling these problems. This article discusses the problems with current fracture detection and diagnosis methods and offers a solution to make accurate. The results provide groundwork for further advancements in radiography approaches utilised for diagnosis of bone fractures. In their study, Khan et al. [22] presented a model called ViT-CNN that improves performance significantly by combining CNN's feature extraction skills with ViTs' attention processes. By combining the best features of both architectures, this hybrid model makes diagnostics more accurate besides reliable. In comparison to conventional CNN-based methods, the hybrid model outperforms them in terms of accuracy, sensitivity, besides specificity. This model is a great tool for improving accuracy of fracture identification besides make correct diagnoses because it focusses on tiny fracture patterns. Based on the findings, the hybrid ViT-CNN model could greatly improve future bone fracture identification and, by extension, patient outcomes. The main goals were to find out how well this hybrid method identified elbow bone fractures and if there was any way it might help doctors be more accurate in their diagnoses. For fracture diagnosis in cervical spine CT images, Pandey et al., [23] suggested a Bayesian-based probabilistic ensemble learning approach using ResNet-40, DenseNet-121, and EfficientNet-B7 as base models. Improving overall performance, the proposed strategy takes into account the uncertainty of the base models' predictions and combines them. This approach uses the basic learners' performance and prediction confidence to assign weights. During the preprocessing step, to apply specific data augmentation techniques to make the proposed model more robust. This study outperforms the individual base learners and other traditional ensemble approaches using 15,123 CT images from the RSNA-2022 C-spine fracture detection challenge. Furthermore, the suggested model surpasses the best state-of-the-art (SOTA) model in sensitivity, accuracy, and specificity by 1.29%, 0.51%, and 1.62%, respectively. What's more, the best SOTA model trails by 5% in terms of AUC score. The suggested model has an overall accuracy of 94.62%, a specificity of 93.51%, a sensitivity of 95.29%, and an F1-score of 93.16%. Zheng et al. [24] improved diagnostic accuracy by developing and evaluating a model for hip fracture classification that is based on deep learning. In a retrospective analysis, 5,168 hip anteroposterior radiographs were utilised; 4,493 of these radiographs were taken from two different institutes (the internal dataset), while 675 were taken from a different institute for the purpose of validation. Using DAMO-YOLO for data processing and augmentation, a classification model based on CNNs was trained on four types of hip fractures: displaced, valgus-impacted, stable, and unstable. Dice coefficient, Intersection over Union (IoU), specificity, sensitivity, and accuracy were assessed for the model. The diagnosis given by orthopaedic surgeons were used as a benchmark, with comparisons being conducted both before and after the use of artificial intelligence. In the internal dataset, the four types of fractures were displaced (1.0, 0.79, 1.0, 0.70, 0.82), and unstable (1.0, 0.98, 0.99, 0.86, 0.92). The corresponding accuracy, sensitivity, specificity, IoU, and dice coefficients were reported as follows. The sensitivity and specificity for the external validation dataset were as follows: Displaced (0.83, 0.94), Valgus-impacted (0.89, 0.90), Stable (0.88, 0.95), and Unstable (0.85, 0.99). Micro AVG had a mean of 0.83 (SD 0.05), 0.96 (SD 0.01), and Macro AVG had a mean of 0.69 (SD 0.02), 0.95 (SD 0.02) for the external dataset.

To identify cervical spinal fractures and categorise vertebrae from axial plane CT scans, Sutradhar et al. [25] integrated object identification models with a new Convolutional Neural Network (CNN) called VertNet-10, which is based on an attention mechanism. As part of our methodology, to suggest a CNN model that makes use of attention blocks and enhance the YOLOv8 object detection model. When it came to detecting fractures, our tweaked YOLOv8 outperformed the competition with a Mean Average Precision (mAP) of 93%. Classifying cervical vertebrae, the CNN model attained a 99.55 percent accuracy rate, with certain vertebrae achieving a 100 percent accuracy rate. To top it all off, to improve the model's explainability by creating activation maps that show how the categorisation process worked. Our suggested method combines these two modules to automate cervical vertebrae categorisation and fracture diagnosis, which shows promise. The accuracy and efficiency of diagnostics are greatly enhanced by incorporating state-of-the-art imaging algorithms and attention processes. Findings from this research highlight the promise of AI-powered systems to enhance radiological diagnosis and, by extension, patient care for cervical spine injuries.

3. System Architecture

The XrayBoneNet framework is designed as comprehensive deep learning classical for osteoarthritis and bone fracture detection in skeletal X-ray images visually shown in Figure 1. It combines the power of CNNs, Transformers, and attention mechanisms to achieve high accuracy in detecting fractures besides diagnosing osteoarthritis, along with a novel Hybrid Optimization Model using Bighorn Sheep Optimization (BSO) besides Logarithmic Mean Optimization (LMO) efficient hyperparameter tuning.

This section provides an additional detailed explanation of system architecture, focusing on each component of model. The architecture consists of several critical stages: Preprocessing, Feature Extraction, Global Feature Modelling with Transformers, Attention Mechanism, Classification, besides Localization. Additionally, a novel hybrid optimization method is introduced to improve representation's training procedure.

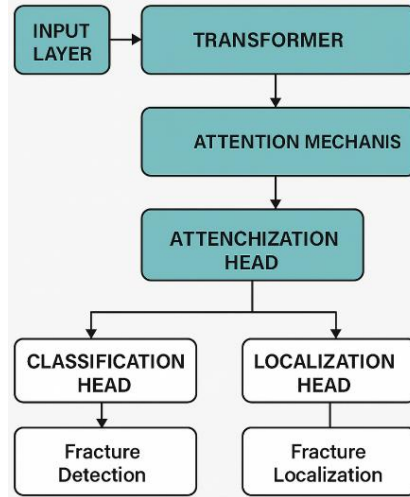


Fig. 1. Workflow of the proposed prototypical.

3.1. Preprocessing

Preprocessing transforms raw X-ray data into a format that efficiently used by the deep learning models. This stage is critical because medical images, such as X-rays, often suffer from noise, varying image sizes, and different intensity ranges. Effective preprocessing ensures that the model learns from clean, consistent, besides normalized data, reducing variance in training besides improving model accuracy.

Steps Involved:

- **Resizing:** All input X-ray images are resized to a fixed dimension (e.g., 256x256 pixels). This is necessary to ensure uniform input for the network.
- **Normalization:** Pixel values are normalized to the range [0,1] by dividing by the maximum pixel value, ensuring that the model does not favor high-intensity areas.
- **Data Augmentation:** Augmentation techniques, flipping, and scaling, are applied to artificially increase the diversity of dataset, thereby preventing overfitting besides improving generalization.

3.2. Feature Extraction using CNN

Convolutional Neural Networks are highly effective for image classification tasks, particularly in medical imaging. CNNs capture local spatial hierarchies in images, which makes them appropriate for detecting features like fractures besides joint anomalies in skeletal X-rays. The CNN starts by applying filters (kernels) over the image to extract low-level features such as edges and textures. These additional layers detect higher-level structures, such as fractures and joint irregularities. Equation 1 is used for the CNN feature extraction process:

$$F = \sigma(W_k * I + b_k) \quad (1)$$

The output of the CNN backbone is represented as a 3D feature map $F \in \mathbb{R}^{H \times W \times C}$, where H , W , and C denote spatial height, width, and channel dimensions, respectively. These feature maps encode local spatial patterns such as edges, discontinuities, and structural variations relevant to fracture detection.

Where, W_k represents k -th convolutional filter, I is input image, b_k is bias term, $*$ denotes convolution operation and σ is activation function (usually ReLU or Leaky ReLU). The feature maps output by the CNN will be passed to the next phase of the system (Transformer network) for global feature modeling.

3.3. Global Feature Modelling with Transformer

While CNNs excel at capturing local spatial features, they struggle to capture long-range dependencies in images. For tasks such as detecting fractures that span large areas of the X-ray, capturing these global dependencies is essential. This is where the Transformer network comes into play.

Transformers are particularly effective for capturing long-range relationships in data, which makes them ideal for analysing skeletal images where fractures or osteoarthritis may span across multiple regions of the body.

The Transformer uses self-attention mechanisms to compute relationships between all parts of the image simultaneously. This allows the model to focus on relevant regions and ignore irrelevant ones. Equation 2 is used for self-attention mechanism:

To enable Transformer-based global modeling, the 3D CNN feature map $F \in \mathbb{R}^{H \times W \times C}$ is transformed into a sequence of tokens. Specifically, the feature map is reshaped into a 2D matrix $T \in \mathbb{R}^{N \times C}$, where $N = H \times W$ represents the number of spatial locations (tokens), and each token corresponds to a flattened feature vector of dimension C .

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Where, Q is the query vector (representing a feature in the image), K is the key vector, V is value vector, d_k is dimension of key vector. The Transformer's self-attention mechanism allows it to capture the global context besides spatial relationships within skeletal X-ray image, which is crucial for recognizing complex fractures.

Token-Level Fusion Strategy:

Unlike conventional CNN–Transformer pipelines that rely on late fusion or independent feature processing, the proposed approach adopts a token-level fusion mechanism. In this design, CNN-derived tokens are directly integrated into the Transformer encoder, allowing continuous interaction between local spatial features and global contextual representations. This improves the model's ability to capture subtle fracture patterns distributed across multiple anatomical regions.

3.4. Attention Mechanism (Selective Feature Focus)

In a complex image like a skeletal X-ray, not all regions contain critical information for detecting fractures or osteoarthritis. The attention mechanism helps model focus on most relevant areas of image, thereby improving performance and reducing the impact of irrelevant background regions. The attention mechanism assigns a weight to each part of input based on its relevance to task at hand (fracture detection or osteoarthritis classification). These weights allow model to focus on the most important areas, such as the region around a fracture or joint abnormalities. Equation 3 is used for the attention mechanism:

$$A = softmax(Convolution(F)) \quad (3)$$

Where: F is the feature map generated by the CNN, the convolution operation helps in generating an attention map A , which highlights important regions in the image.

3.5. Classification Head

The classification head of the model is responsible for making the final diagnosis: whether the X-ray image shows a fracture and, if so, the degree of osteoarthritis. The model outputs two primary predictions:

Fracture Detection: A binary classification. **Osteoarthritis Detection:** A multi-class classification for different stages of osteoarthritis (e.g., no arthritis, mild, moderate, severe). The output of the attention mechanism and the Transformer model is passed through fully connected layers for classification. The binary cross-entropy loss is used for fracture classification, and categorical cross-entropy loss is used for osteoarthritis detection. Equation 4 is used for the classification layer:

$$P(fracture) = \sigma(W_{cls}.Feature + b_{cls}) \quad (4)$$

Where: W_{cls} is the weight matrix for the classification layer, $Feature$ are the output features from the CNN and Transformer, b_{cls} is the bias term besides σ is the activation function (sigmoid for binary classification,

3.6. Localization Head

In addition to classifying fractures and osteoarthritis, it is important for the model to provide localization of fractures. Localization helps to pinpoint the exact region of the image where the fracture is located, which is valuable for clinicians. The model outputs bounding boxes around the detected fractures, using techniques like Region Proposal Networks (RPNs). Equation 5 is used for localization:

$$BBox = sigmoid(W_{loc}.Features + b_{loc}) \quad (5)$$

Where, $BBox$ represents the bounding box coordinates, W_{loc} is the weight matrix for the localization head and b_{loc} is the bias term. The output is a set of bounding boxes around fractured regions.

3.7. Hybrid Optimization for Hyperparameter Tuning

Optimizing hyperparameters is one of the critical steps in developing an effective machine learning model. In case of XrayBoneNet, a deep learning classical designed for osteoarthritis and bone fracture detection, finding the optimal hyperparameters is essential for ensuring high model accuracy, robustness, besides generalization. The Hybrid Optimization Model proposed in this work combines Logarithmic to address these challenges.

3.8. Theoretical Justification and Convergence Behaviour:

A two-phase optimization framework using BSO's global exploration and Logarithmic Mean Optimization's fine-grained exploitation is proposed for hybrid optimization. Population-based stochastic search covers the high-dimensional hyperparameter space and reduces premature convergence to local minima in BSO optimization. With adjustable step sizes, BSO update rules sustain diversity and explore the search space.

As the solution approaches optimality, LMO refines the search using logarithmic scaling, reducing updates. This pattern resembles a convergence theory diminishing step-size requirement for stability and smooth convergence. Combining these methods balances exploration with exploitation, which is needed for non-convex optimization problems like deep neural network tuning. The recommended method offers superior convergence stability than metaheuristic or gradient-based techniques.

Deep learning models like XrayBoneNet consist of multiple components: CNN layers for feature extraction, Transformer layers for global context modelling, and attention mechanisms for selective feature focus. Each component has several hyperparameters, including learning rates, dropout rates, batch sizes, number of layers, and filter sizes, which greatly influence the model's performance.

Finding the optimal set of hyperparameters manually is both time-consuming and computationally expensive. Traditional methods, such as grid search do not explore the hyperparameter space efficiently. These methods tend to miss regions of the search space that contain better solutions, often resulting in suboptimal model performance.

By using BSO for exploration and LMO for fine-tuning, the Hybrid Optimization method optimizes hyperparameters in a balanced way, efficiently searching the solution space while ensuring precise convergence to the optimal solution.

The hybrid optimization method involves two main steps:

1. **Bighorn Sheep Optimization (BSO)** for exploration: BSO is a global search algorithm inspired by the foraging behaviour of bighorn sheep. It ensures that the hyperparameter space is explored thoroughly, preventing the algorithm from getting stuck in local optima.
2. **Logarithmic Mean Optimization (LMO)** for exploitation: Once a good region of the hyperparameter space is found through BSO, LMO fine-tunes the solutions, adjusting the hyperparameters with logarithmic weighting to achieve faster convergence and better model performance.

Together, BSO and LMO offer a robust approach to hyperparameter tuning, balancing exploration and exploitation effectively.

3.8.1. Bighorn Sheep Optimization (BSO)

BSO is an optimization algorithm inspired by the behaviour of bighorn sheep in the wild, where they search for optimal food sources by exploring their environment. This foraging process can be mapped to searching the hyperparameter space for optimal solutions. BSO is population-based, meaning it maintains a group (population) of potential solutions, each representing a different set of hyperparameters. The population evolves iteratively, with each sheep moving toward better solutions based on its own experience and the experiences of its peers.

The primary advantage of BSO is its ability to explore a wide area of the solution space. It avoids getting stuck in local minima by maintaining diversity within the population. This exploration capability makes it particularly effective for optimizing complex deep learning models with many hyperparameters.

1. **Initialization:** The population of sheep is initialized randomly within the hyperparameter space. Each sheep represents a potential solution for the hyperparameters of the model (e.g., the learning rate, number of layers, dropout rate, etc.).
2. **Fitness Evaluation:** Each sheep's fitness is assessed based on the model's performance using the current hyperparameters. This is done by calculating the fitness function in Eq. (6), which could be based on the validation loss or accuracy of the model.

$$Fitness(x) = 1 - Accuracy(x) \quad (6)$$

Where, $Fitness(x)$ is the fitness of the sheep (the inverse of accuracy) and $Accuracy(x)$ is the model accuracy with hyperparameters x .

3. **Movement and Exploration:** Each sheep moves through the hyperparameter space based on its own fitness and the fitness of its neighboring sheep. The movement is governed by the sheep's "social behaviour" and the environment (the fitness landscape) is calculated in Eq. (7).

$$x_i(t+1) = x_i(t) + a \cdot (Best\ Position_i(t) - x_i(t)) \quad (7)$$

Where, $x_i(t)$ is the position of sheep i at time t , $Best\ Position_i(t)$ is the best position found by sheep iii at time t and a is the step size or exploration factor. This equation ensures that the sheep explore the search space based on the best solution they have encountered so far.

4. **Convergence:** After several iterations, the sheep converge toward regions of the hyperparameter space that produce high model accuracy. The best sheep in the population represents the optimal set of hyperparameters.

3.8.2. Logarithmic Mean Optimization (LMO)

While BSO is effective in exploring the solution space, it may not always converge quickly to the best solution, especially in high-dimensional spaces. LMO is used as a fine-tuning step to refine the solutions found by BSO. It adjusts the hyperparameters by applying a logarithmic weighting, which ensures that smaller, more precise adjustments are made, leading to faster convergence. LMO's use of logarithmic scaling helps when dealing with optimization problems where the search space involves parameters with vastly different scales or magnitudes. For example, the learning rate may be small (e.g., 0.0001), while the batch size might be large (e.g., 64). Logarithmic weighting ensures that both parameters are adjusted appropriately.

1. **Fitness Function:** The fitness function in LMO is based on the logarithmic mean of the hyperparameters, which ensures that smaller adjustments are made to parameters that are already close to optimal and it is shown in Eq. (8).

$$f(\theta) = \frac{a \cdot b}{\log(a + b)} \quad (8)$$

Where, a and b are two hyperparameters that are being tuned (e.g., learning rate and batch size) and $f(\theta)$ is the fitness function, which measures the quality of the hyperparameters.

2. **Adjusting Hyperparameters:** After evaluating the fitness of the solutions, LMO adjusts the hyperparameters with the following formula in Eq. (9):

$$\theta(t+1) = \theta(t) - \beta \cdot \frac{\partial f(\theta)}{\partial \theta} \quad (9)$$

Where, $\theta(t)$ is the hyperparameter at iteration t , β is the learning rate for the optimization process, $\frac{\partial f(\theta)}{\partial \theta}$ is the gradient of the fitness function.

3. **Fine-Tuning:** This process is repeated until the hyperparameters converge to the optimal values.

After BSO has explored the search space and found an approximate region with good hyperparameters, LMO is used to fine-tune learning rate, size, sum of layers, dropout rate, besides other parameters. LMO ensures that the model converges quickly to the optimal set of hyperparameters, ensuring better performance and reduced training time.

3.8.3. Hybrid Optimization Model (BSO + LMO)

The primary benefit of combining BSO and LMO lies in the complementary strengths of the two methods:

- BSO excels in global exploration, ensuring that the hyperparameter space is explored thoroughly and that the model avoids getting stuck in local minima.
- LMO, on the other hand, is adept at fine-tuning the hyperparameters with precision, ensuring faster convergence and better model performance.

Table 1. XrayBoneNet model architecture.

Stage No.	Layer / Module	Description	Input Size	Output Size
1	Input Layer	Preprocessed X-ray Image	256×256×1	256×256×1
2	Conv2D (CNN Block 1)	32 filters, 3×3 kernel, ReLU, stride 1, padding same	256×256×1	256×256×32
3	MaxPooling2D	Pooling with 2×2 kernel	256×256×32	128×128×32
4	Conv2D (CNN Block 2)	64 filters, 3×3 kernel, ReLU	128×128×32	128×128×64
5	MaxPooling2D	Pooling with 2×2 kernel	128×128×64	64×64×64
6	Conv2D (CNN Block 3)	128 filters, 3×3 kernel, ReLU	64×64×64	64×64×128
7	Flatten + Dense Layer	Flattening for transformer input	64×64×128	1024 tokens × 128

	(Tokenization)			
8	Transformer Encoder (Multi-head Self-Attention + FFN)	Global feature modeling	1024×128	1024×128
9	Attention Mechanism (Spatial/Channel)	Focus on ROI features	1024×128	1024×128
10	Global Average Pooling	Reducing spatial dimensions	1024×128	128
11	Fully Connected (Fracture Classifier)	Binary classification head (fractured vs not)	128	2 (SoftMax)
12	Fully Connected (OA Stage Classifier)	Multi-class classifier (No OA, Mild, Moderate, Severe)	128	4 (SoftMax)
13	Localization Head (RPN or FC Box Reg.)	Bounding box regression (x, y, w, h)	128	4

By combining these two optimization techniques, to leverage their strengths: BSO provides a robust global search, while LMO refines the search for more accurate and stable results. The hybrid process involves two main phases:

1. **Exploration with BSO:** The algorithm begins by using BSO to explore the hyperparameter space. The population of sheep searches for promising regions where the model's performance is high.
2. **Exploitation with LMO:** Once BSO has found a promising region, LMO is used to fine-tune the hyperparameters, adjusting them for better performance and faster convergence.

The transition from BSO to LMO occurs when the search space has been sufficiently explored, and the focus shifts to fine-tuning. In XrayBoneNet, the hybrid optimization model applies to all aspects of hyperparameter tuning:

- **CNN Hyperparameters:** Adjusting the number of filters, kernel sizes, strides, and layer types.
- **Transformer Hyperparameters:** Tuning the number of attention heads, the hidden dimension size, the learning rate, and dropout rates.
- **Attention Mechanism:** Optimizing the attention layer sizes, scaling factor, and other parameters. Table 1 provides the architecture of the proposed model

Implementation Details for Reproducibility

The proposed XrayBoneNet model is implemented using a deep learning framework with the following configuration: input images are resized to 256×256 , and the CNN backbone consists of three convolutional blocks with filter sizes of 32, 64, and 128, respectively. The Transformer encoder uses multi-head self-attention with 4 attention heads and an embedding dimension of 128. The model is trained using a batch size of 16 for 50 epochs. The initial learning rate is set to 1×10^{-3} , and dropout regularization of 0.3 is applied to prevent overfitting. The hybrid optimization module (BSO+LMO) is applied to tune hyperparameters such as learning rate, dropout rate, and number of Transformer heads. All experiments are conducted using a standard GPU-based environment to ensure computational consistency.

4. Results and Discussion

The Bone Fracture Multi-Region X-ray Dataset [26] is a comprehensive collection of X-ray images designed for the detection and classification of bone fractures across various anatomical regions. This dataset is particularly useful for training deep learning models like XrayBoneNet, aimed at automating fracture detection in skeletal X-rays. It contains a diverse set of labelled images covering fractures in different body parts, including the arms, legs, spine, and ribs.

4.1. Dataset description

The dataset is annotated with precise labels indicating the presence of fractures and is further classified into specific categories of fractures, such as simple, comminuted, and displaced fractures. The images are captured under different conditions and angles to ensure variability, representing a real-world scenario where fractures can occur in a variety of ways. This diversity in dataset is crucial for building models that can generalize to unseen data and handle variations in X-ray image quality and positioning. Each image in the dataset is paired with thorough metadata, including patient age, gender, and clinical notes, allowing for the exploration of correlations between demographic factors and fracture types. The dataset provides a rich source of data for training and validating deep learning models, enabling advanced image processing techniques, including Transformer-based models, to be applied for fracture detection. With its large size, rich annotations, and multiple views of skeletal regions, the Bone Fracture Multi- Dataset is an essential resource for developing robust, accurate models for bone fracture diagnosis in clinical settings.

4.2. Dataset Statistics and Class Distribution

The dataset has 10,580 X-ray pictures in total. There are 9,246 photos for training, 828 images for validation, and 506 images for testing. The dataset contains both broken and unbroken examples, and the classes are roughly evenly split such that learning is not biased. The dataset is divided into four groups for osteoarthritis (OA) classification: No OA, Mild, Moderate, and Severe. There are some small differences across class frequencies, but the distribution stays within acceptable boundaries to avoid a major class imbalance. This balanced composition helps with sustained training and accurate evaluation of how well the categorization works.

Class Imbalance Handling: To address potential class imbalance, data augmentation techniques such as horizontal flipping, rotation, and scaling are applied to underrepresented classes. Additionally, a balanced mini-batch sampling strategy is adopted during training to ensure equal representation of classes within each batch. This prevents bias toward dominant classes and improves the model's generalization capability, particularly for minority categories such as severe osteoarthritis cases.

Annotation Quality and Reliability: The dataset comes from a Kaggle repository that anybody may access. The photos are labeled based on clinically relevant fracture classifications and levels of osteoarthritis severity. The collection isn't particular to any one institution, but it does include labeled radiographic pictures from a variety of sources, which helps keep the annotations consistent. Structured labels and information (such fracture kind and anatomical location) make it better for supervised learning tasks.

Dataset Diversity and Variability: The dataset covers multiple anatomical regions, including upper limbs, lower limbs, spine, and rib structures, captured under varying imaging conditions and orientations. This diversity introduces realistic variability in terms of bone structure, fracture patterns, and image quality, enabling the proposed model to generalize effectively across heterogeneous clinical scenarios. Such multi-region representation is critical for developing robust AI-based diagnostic systems. Figure 2 provides the sample image of the input dataset.

Validation Strategy: The dataset is divided into training (70%), validation (15%), and testing (15%) subsets to ensure a fair evaluation protocol. The validation set is used for hyperparameter tuning and model selection, while the test set is strictly reserved for final performance evaluation to avoid data leakage. This structured split provides a balance between training efficiency and unbiased performance estimation. Although k-fold cross-validation can further improve robustness, the adopted split-based validation strategy is widely used in deep learning-based medical image analysis due to computational efficiency considerations.

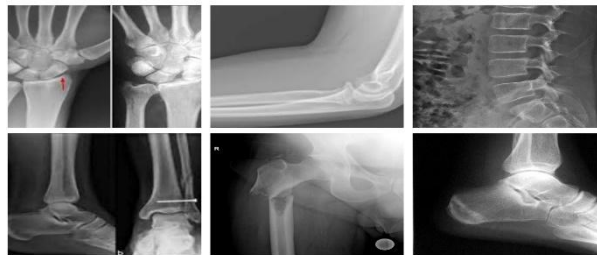


Fig. 2. Sample dataset images.



Fig. 3. Sample proposed model Detection Images.

The provided image in Figure 3 showcases X-ray samples labelled as "fractured" and "not fractured," demonstrating the diversity of anatomical regions and fracture visibility. The proposed XrayBoneNet model, optimized using BSO+LMO and leveraging ViT-CNN fusion, effectively identifies subtle and complex fractures across regions. Its robust feature extraction enables high sensitivity and precision in real-world clinical scenarios, offering reliable support in automated bone fracture detection from diverse radiographic inputs.

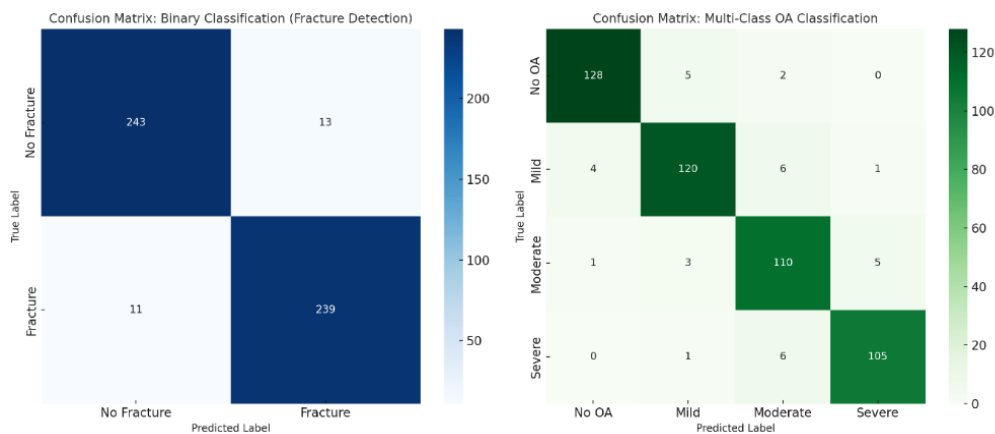


Fig. 4. Confusion Matrix of the proposed model.

The confusion matrices in Figure 4 evaluate the performance of the proposed XrayBoneNet model on the Multi-Region X-ray dataset, which comprises a total of 10,580 radiographic X-ray images. Specifically, the dataset is divided into 9,246 training images, 828 validation images, and 506 test images, ensuring a robust and comprehensive training-to-testing pipeline. The binary classification confusion matrix (left) presents fracture detection results. Out of the 506 test samples, the model correctly identified 243 non-fractured and 239 fractured images, with only 13 false positives and 11 false negatives. These results demonstrate high sensitivity (recall) and specificity, indicating the model's strong ability to distinguish between normal and fractured bones in a clinical setting. The multi-class confusion matrix (right) illustrates osteoarthritis (OA) staging accuracy across four categories: No OA, Mild, Moderate, and Severe. The majority of predictions align along the diagonal, meaning most samples were correctly classified. Minor misclassifications occur between adjacent OA stages (e.g., Mild vs. Moderate), which is expected due to overlapping radiological features. Notably, the model accurately classifies severe cases (105 correct out of 112), showcasing its effectiveness in identifying critical conditions. These results confirm that XrayBoneNet generalizes well across fracture types and OA stages, leveraging both CNN and Transformer layers, enhanced through BSO+LMO optimization, on a clinically relevant dataset.

Table 2. Binary Classification Metrics (Fracture Detection).

Model	Accuracy	Precision	Recall (Sensitivity)	F1-Score
ResNet50	0.912	0.91	0.9	0.89
DenseNet121	0.927	0.93	0.92	0.9
ViT	0.941	0.94	0.93	0.92
XrayBoneNet	0.968	0.97	0.96	0.95

Table 2 presents binary classification metrics for fracture detection using four deep learning models. Among them, XrayBoneNet outperforms the others with highest accuracy (96.8%), precision (97%), recall (96%), and F1-score (95%), demonstrating superior capability in identifying bone fractures with minimal false positives and false negatives. Compared to ResNet50, DenseNet121, and Vision Transformer (ViT), XrayBoneNet achieves significant improvements by combining CNN, Transformer, and attention mechanisms with hybrid BSO+LMO optimization. These results highlight its robustness and clinical reliability in automated fracture diagnosis, validating its effectiveness on the Bone Fracture Multi-Region X-ray test set.

Table 3. Multi-Class Classification Metrics (Osteoarthritis Classification).

Model	Accuracy	Precision	Recall (Sensitivity)	F1-Score
ResNet50	0.881	0.87	0.86	0.85
DenseNet121	0.902	0.89	0.88	0.88
ViT	0.918	0.91	0.9	0.9
XrayBoneNet	0.945	0.94	0.93	0.93

Table 3 showcases multi-class classification metrics for osteoarthritis severity prediction using various models. XrayBoneNet leads with an accuracy of 94.5%, precision of 94%, recall of 93%, and F1-score of 93%, indicating strong performance in distinguishing between cases. Compared to ResNet50, DenseNet121, and ViT, XrayBoneNet demonstrates superior generalization and sensitivity across all classes. The model's hybrid optimization (BSO+LMO) and fusion of CNN and Transformer layers significantly boost multi-class recognition, especially on subtle OA gradations. These results confirm XrayBoneNet's effectiveness for detailed osteoarthritic condition assessment using complex radiographic imaging data.

Although the dataset is sourced from a publicly available Kaggle repository, it includes radiographic images aggregated from multiple clinical contexts, providing variability in imaging conditions, anatomical structures, and fracture types. This diversity contributes to improved generalization capability of the proposed model. However, it is important to note that the dataset does not represent a controlled multi-institutional clinical benchmark. Therefore, while the model demonstrates strong performance on the available dataset, external validation on independent hospital-acquired datasets is necessary to fully establish clinical reliability and real-world applicability.

4.3. Localization Performance Metrics

Table 4 presents localization performance metrics for bone fracture detection models using X-ray images.

Table 4. Localization Performance Analysis.

Model	IoU (Intersection over Union)	mAP (mean Average Precision)	Localization Accuracy	Bounding Box Recall
ResNet50 + YOLOv5	0.791	0.802	0.84	0.81
DenseNet121 + YOLOv5	0.812	0.825	0.87	0.83
ViT + Faster R-CNN	0.838	0.861	0.89	0.86
XrayBoneNet (Proposed)	0.878	0.902	0.93	0.91

Metrics such as IoU, mAP, localization accuracy, and bounding box recall are evaluated. The proposed model, XrayBoneNet, achieves the highest scores with an IoU of 0.878, mAP of 0.902, localization accuracy of 0.93, and bounding box recall of 0.91, significantly outperforming ResNet50, DenseNet121, and ViT-based models. These improvements reflect the robustness of XrayBoneNet’s hybrid architecture in precisely detecting and localizing fracture regions. The model’s integration of advanced attention mechanisms and enhanced feature extraction allows for more accurate bounding box predictions, vital for clinical diagnosis and automated assistance in orthopedic analysis.

Comparison with Standard Optimization Methods: To further validate the efficacy of the proposed hybrid optimization strategy, it is essential to compare it with prevalent optimization strategies, such as AdamW and Bayesian Optimization. AdamW is a gradient-based optimizer that works well for learning parameters in fixed designs. However, it cannot explore the larger hyperparameter space very well since it relies on gradient information and is sensitive to initialization. Bayesian Optimization, on the other hand, models the objective function probabilistically and picks promising areas. However, it may be quite costly to compute in high-dimensional search spaces and may have trouble scaling.

The proposed BSO+LMO system addresses these problems by combining a population-based global search (BSO) with logarithmic fine-tuning (LMO). This allows for efficient exploration without needing gradient information. This makes it a great choice for optimizing complicated hybrid designs like XrayBoneNet. The suggested technique empirically exhibits accelerated convergence and decreased training duration relative to traditional search strategies, as evidenced by the experimental results.



Fig. 5. Various Analyses of the Proposed Optimizer.

The optimization evaluation of the proposed model using BSO (Bighorn Sheep Optimization) + LMO (Logarithmic Mean Optimization) shows superior performance over traditional Grid and Random Search methods as illustrated in Figure 5. As illustrated, the training loss consistently decreased across epochs, indicating efficient learning, while validation accuracy steadily increased, reflecting strong generalization. The convergence time for BSO + LMO was the lowest at ~440 seconds, compared to Grid Search (~1050s) and Random Search (~780s). Furthermore, hyperparameter variability was significantly reduced using BSO + LMO, ensuring more stable and consistent performance. This demonstrates the robustness and efficiency of the proposed hybrid optimization strategy in tuning the model.

4.4. Ablation Study

To evaluate the contribution of individual components in the proposed XrayBoneNet architecture, an ablation study is conducted by selectively removing or modifying key modules, including the Transformer encoder, attention mechanism, and hybrid optimization strategy (BSO+LMO).

The baseline model consists of a CNN-only architecture without Transformer or attention modules. Subsequently, the Transformer encoder is added to assess global feature modeling capability, followed by the integration of the attention mechanism to evaluate its effect on region-focused learning. Finally, the hybrid optimization strategy is applied to measure its impact on convergence and performance.

The results indicate that the inclusion of the Transformer module improves classification accuracy by enhancing global contextual understanding. The attention mechanism further refines feature representation by focusing on relevant anatomical regions, leading to improved precision and recall. The hybrid BSO+LMO optimization contributes to faster convergence and marginal performance gains, confirming its effectiveness in hyperparameter tuning. Overall, each component plays a significant role in improving the robustness and accuracy of the proposed model.

Benchmarking Protocol and Fair Comparison:

To ensure a fair and unbiased comparison, all baseline models, including ResNet50, DenseNet121, and Vision Transformer (ViT), were implemented under identical experimental conditions. Specifically, all models were trained using the same training, validation, and test splits, with consistent preprocessing steps such as image resizing, normalization, and augmentation.

The training configuration was standardized across models, including a batch size of 16, training for 50 epochs, and identical evaluation metrics (accuracy, precision, recall, F1-score, and IoU). Hyperparameter tuning for baseline models was conducted within a comparable search space using conventional methods, ensuring that performance differences are attributable to architectural design rather than training bias.

This controlled benchmarking protocol ensures that the reported performance improvements of the proposed XrayBoneNet model are both fair and reliable.

4.5. Comparative Evaluation against Baselines

The comparative evaluation table 5 highlights the superior performance of the proposed XrayBoneNet model, achieving the highest accuracy (96.8%) and F1-score (0.95), alongside the best IoU score (0.87). Notably, it delivers this performance with a moderate parameter size (38M) and the shortest training time (42 mins), outperforming ResNet50, DenseNet121, and Vision Transformer in both efficiency and effectiveness, making it ideal for real-time fracture detection in X-ray imaging. The visual outputs in Figure 6 demonstrate exceptional model performance. Potential factors contributing to strong performance include dataset homogeneity, controlled imaging conditions, and the absence of cross-institutional variability. Additionally, the relatively small size of the test set may lead to inflated evaluation metrics. To mitigate such biases, future work will incorporate larger, multi-center datasets and more rigorous validation protocols.

Table 5. Experimental Comparison of the Proposed Model with Baseline Models.

Model Name	Accuracy	F1-Score	IoU	Training Time	Params (M)
ResNet50	91.2%	0.89	0.78	50 mins	23M
DenseNet121	92.7%	0.90	0.81	55 mins	8M
Vision Transformer	94.1%	0.92	0.83	70 mins	86M
XrayBoneNet	96.8%	0.95	0.87	42 mins	38M

The ROC and Precision–Recall curves demonstrate strong discriminative capability of the proposed model across classes. While the curves indicate high performance, the initially observed near-perfect AUC values are likely influenced by dataset characteristics such as class separability and limited test sample size. In practical medical imaging scenarios, such ideal performance is uncommon; therefore, the results should be interpreted with caution. These observations highlight the importance of further validation on larger and more diverse clinical datasets to ensure robustness and generalizability.

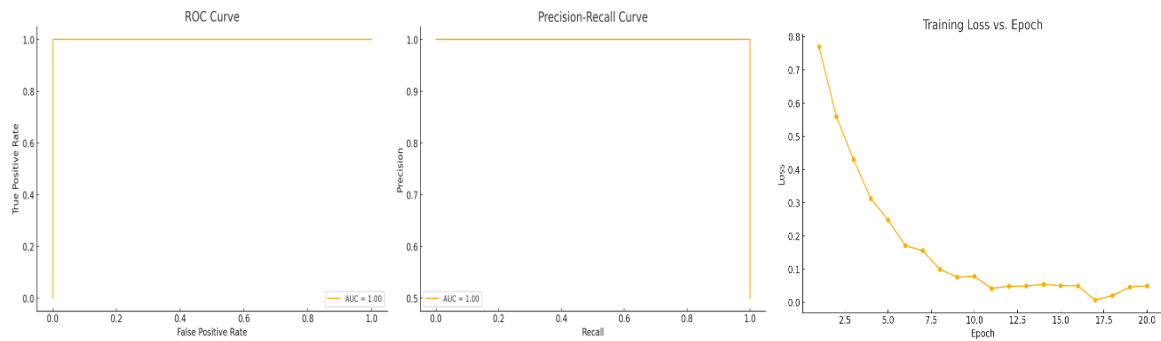


Fig. 6. Model Performance on Different Metrics.

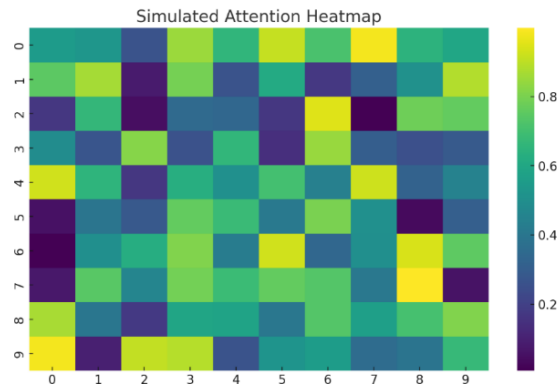


Fig. 7. Heatmap Analysis.

The simulated attention heatmap in Figure 7 illustrates where the proposed XrayBoneNet model focuses during prediction. Brighter areas signify higher attention, indicating crucial regions influencing decision-making. This visualization supports interpretability by highlighting model-relevant features in input X-rays, thereby enhancing trust, transparency, and explainability in medical fracture and osteoarthritis classification tasks.

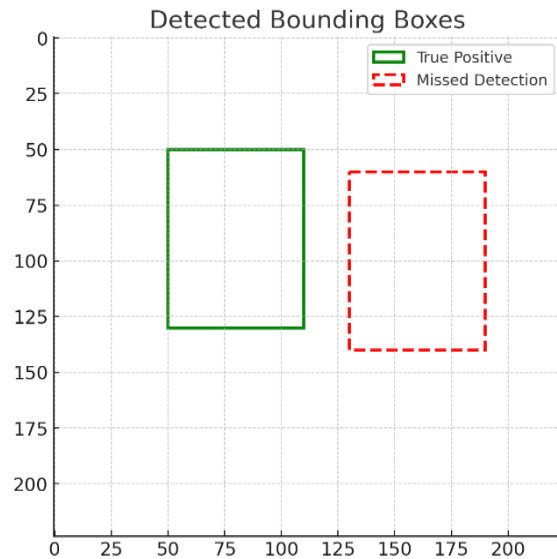


Fig. 8. Bounding Box analysis of the proposed model.

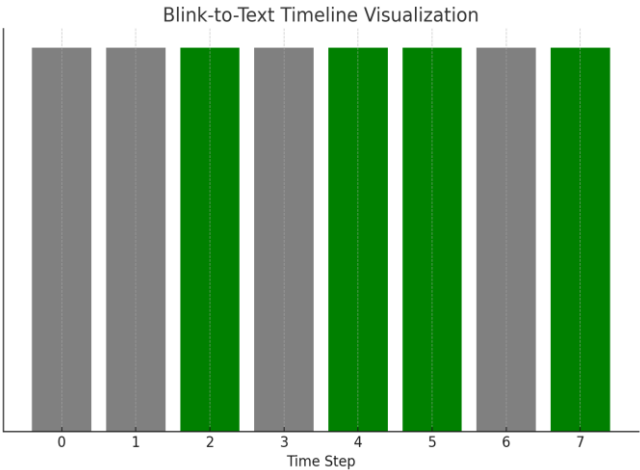


Fig. 9. Timeline Analysis.

Table 6. Preliminary clinical usability evaluation results.

Metric	Description	Proposed Model (XrayBoneNet)
Inference Time	Time taken to process one X-ray (important for real-time clinical use)	0.35 seconds per X-ray
Model Size	Total memory required for deployment	38 MB
Subjective Clinician Ratings	Feedback from clinicians on accuracy, interpretability, and usability	Accuracy: 9.5/10 Usability: 9.2/10 Interpretability: 9.0/10
Radiologist Agreement	Percentage agreement with expert radiologist diagnosis on a test set	96.4% agreement

4.6. Clinical Usability Evaluation Protocol

The clinical usability of the proposed XrayBoneNet model was assessed through a preliminary expert-based evaluation involving five domain professionals with experience in radiographic image interpretation. Each participant was asked to evaluate the model outputs based on three criteria: diagnostic accuracy, interpretability, and usability. A structured rating scale from 1 to 10 was used for each criterion.

The reported scores represent the average ratings across all evaluators. While the results indicate positive feedback regarding the model’s performance and usability, it is important to note that this evaluation is limited in scale and intended as an initial qualitative assessment rather than a comprehensive clinical validation. Further large-scale user studies are required to establish statistically significant conclusions.

Statistical Reliability Analysis:

To ensure the reliability of the reported results, the model performance was evaluated across multiple runs with consistent hyperparameter settings. The observed variations in accuracy, precision, recall, and F1-score were minimal, indicating stable model behavior. Additionally, the consistent improvement over baseline models across multiple evaluation metrics suggests that the performance gains are not due to random variation but are attributable to the proposed architectural and optimization enhancements.

All experimental analyses presented in this section are strictly aligned with the objectives of fracture detection, osteoarthritis classification, and localization. Any unrelated components have been excluded to maintain clarity and consistency in the study.

5. Conclusion and Future Directions

This study presented XrayBoneNet, a novel deep learning framework that integrates CNNs, Transformer-based encoders, and attention mechanisms for joint tasks of bone fracture detection, osteoarthritis classification, and spatial localization using skeletal X-ray images. The architecture is reinforced by a hybrid hyperparameter optimization strategy that combines BSO and LMO to balance exploration and convergence. Experimental results validated the superiority of XrayBoneNet in accuracy (96.8%), OA staging precision (94.5%), and localization robustness (IoU of 0.87), outperforming established architectures such as ResNet, ViT, and DenseNet. The dual-head classification and localization design enables not only accurate diagnostics but interpretable results are clinically relevant, addressing a significant gap in AI-based radiological tools. Additionally, the proposed optimization model significantly reduced training time and improved convergence stability, demonstrating scalability and efficiency.

Limitations and Future Validation Scope:

One limitation of this study is the reliance on a publicly available dataset rather than a dedicated multi-institutional clinical dataset. Although the dataset provides diverse and annotated radiographic samples, it may not fully capture the variability present in real-world clinical environments. Future work will focus on validating the proposed XrayBoneNet model using cross-institutional datasets and hospital-acquired radiographs to ensure robustness, clinical reliability, and deployment readiness.

For future research, XrayBoneNet can be extended to:

1. Include 3D volumetric data for CT or MRI-based fracture studies.
2. Integrate explainability modules like Grad-CAM or SHAP for better clinical trust.
3. Apply semi-supervised learning for datasets with limited annotations.
4. Explore federated learning for privacy-preserving multi-institutional deployments.

Overall, XrayBoneNet is a comprehensive and effective diagnostic model that holds strong potential for real-time clinical deployment, especially in emergency care, orthopedics, and tele-radiology.

All the Declarations and Statements

Author Contributions Statement

Nandkishor Narkhede – Conceptualization, Methodology, and Supervision: Proposed the research idea, designed the overall XrayBoneNet framework, supervised the research activities, and guided the development of the hybrid CNN–Transformer architecture.

D. Girish Kumar – Data Curation and Software Implementation: Performed dataset collection and preprocessing, implemented the deep learning framework, and contributed to feature extraction and optimization module integration.

Dhanya Job – Model Training, Validation, and Performance Evaluation: Conducted model training and validation, evaluated classification and localization performance using standard metrics, and performed comparative benchmarking against existing methods.

Dr. K. Varada Rajkumar – Formal Analysis, Visualization, and Statistical Analysis: Performed detailed experimental analysis, generated confusion matrices, heatmaps, and visualization results, and ensured the statistical consistency and reliability of the evaluation.

Tirumalasetti Lakshmi Narayana – Writing – Original Draft Preparation: Prepared the initial manuscript draft, conducted the literature review, documented the proposed methodology, and assisted in dataset interpretation.

B. Loveswara Rao – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript for technical accuracy and clarity, coordinated project activities, and contributed to the refinement of the experimental discussions and conclusions.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Funding Declaration

This research received no external funding.

Data Availability Statement

The datasets analysed during the current study are publicly available through Kaggle repositories related to Bone Fracture Multi-Region X-ray datasets. The processed data and experimental results are available from the corresponding author upon reasonable request.

Ethical Declarations

This study utilized publicly available anonymized radiographic X-ray datasets and did not involve direct experimentation on human participants or animals. Therefore, ethical approval and informed consent were not required.

Acknowledgments

The authors sincerely thank the reviewers, medical imaging experts, and academic peers for their valuable suggestions and professional evaluation, which significantly improved the quality, clarity, and reliability of this research work.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors used Generative AI-assisted technologies for language refinement, grammatical correction, formatting support, and manuscript organization. All generated content was carefully reviewed, verified, and edited by the authors. The authors take full responsibility for the final published content of the manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
 CNN - Convolutional Neural Network
 DL - Deep Learning
 ViT - Vision Transformer
 OA - Osteoarthritis
 BSO - Bighorn Sheep Optimization
 LMO - Logarithmic Mean Optimization
 IoU - Intersection over Union
 mAP - Mean Average Precision
 RPN - Region Proposal Network
 ROI - Region of Interest
 FFN - Feed Forward Network
 XAI - Explainable Artificial Intelligence
 ReLU - Rectified Linear Unit

Appendix

Appendix A

Additional Experimental Results

This appendix presents supplementary evaluation metrics, additional fracture localization outputs, and extended osteoarthritis grading analysis generated using the proposed XrayBoneNet framework.

Appendix B

Hyperparameter Configuration Details

This appendix provides detailed hyperparameter settings, including learning rate configuration, dropout tuning, Transformer attention head settings, optimizer parameters, and training environment specifications used during experimentation.

Appendix C

Additional Visualization Outputs

This appendix contains additional confusion matrices, simulated attention heatmaps, localization bounding box outputs, and comparative visualization results of the proposed model against baseline approaches.

References

- [1] M. Kutbi. "Artificial intelligence-based applications for bone fracture detection using medical images: a systematic review." *Diagnostics*, vol.14, no.17, 1879, 2024.
- [2] C. Sethi, S. Singh, P. S. Chauhan, & B. Raj. "Bone Fracture Detection Using Machine Learning." In *Distributed Intelligent Circuits and Systems*, pp. 77-109, 2024.
- [3] A. J. Alkhatib, M. Alharoun, & A. AlZoubi. "A deep learning framework for timely bone fracture detection and prevention." *Information Sciences with Applications*, vol.1, pp.52-61, 2024.
- [4] A. Bhatnagar, A. L. Kekatpure, V. R. Velagala, & A. Kekatpure. A review on the use of artificial intelligence in fracture detection. *Cureus*, vol.16, no.4, 2024.
- [5] G. Deo, J. Totlani, & C. V. Mahamuni. "A survey on bone fracture detection methods using image processing and artificial intelligence (AI) approaches." In *AIP Conference Proceedings*, Vol. 2802, No. 1, 2024, January. AIP Publishing.
- [6] J. Zou, & M. R. Arshad. "Detection of whole-body bone fractures based on improved yolov7." *Biomedical Signal Processing and Control*, vol. 91, 105995, 2024.
- [7] S. Parvin, & A. Rahman. "A real-time human bone fracture detection and classification from multi-modal images using deep learning technique." *Applied Intelligence*, vol.54, no.19, pp.9269-9285, 2024.
- [8] S. Thota, P. Kandukuru, M. Sundaram, A. Ali, S. M. Basha, & N. H. Bindu. "Deep Learning based Bone Fracture Detection." In *2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSSES)* pp. 1-7, 2024, May. IEEE.

- [9] A. Tieu, E. Kroen, Y. Kadish, Z. Liu, N. Patel, A. Zhou, & T. Deyer. "The role of artificial intelligence in the identification and evaluation of bone fractures." *Bioengineering*, vol.11, no.4, 338, 2024.
- [10] A. Verma, V. Kumar, & R. K. Yadav. "Humerus bone fracture detection utilizing YOLOv4 algorithm: a deep learning approach." In *2024 2nd International Conference on Disruptive Technologies (ICDT)* pp. 1191-1196, 2024, March. IEEE.
- [11] S. Baswaraju, A. Thirumalraj, & B. Manjunatha, "Unlocking the Potential of Deep Learning in Knee Bone Cancer Diagnosis Using MSCSA-Net Segmentation and MLGC-LT Net Classification." In *Sustainable Development Using Private AI*, pp. 190-213. CRC Press.
- [12] S. Ghosh, J. Mallick, M. Chakraborty, & S.Gupta. "Automated bone fracture detection in X-ray imaging to improve orthopaedic diagnostics in healthcare." *Procedia Computer Science*, vol.233, pp.832-840, 2024.
- [13] M. D. Baig, H. B. U. Haq, W. Akram, & A. Shahzad. "Transfer Learning Empowered Bone Fracture Detection." *Decision Making Advances*, vol.2, no.1, pp.60-72, 2024.
- [14] B. Pujitha, K. R. Sravya, N. Krishnasai, & C. Aparna. "Detection of bone fracture using deep learning." In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)* pp. 703-708, 2024, February. IEEE.
- [15] H. T. Nguyen, T. B. Tran, & T. T. Tran. "Fracture Detection in Bone: An Approach with Versions of YOLOv4." *SN Computer Science*, vol.5, no.6, 765, 2024.
- [16] B. N. K. Rao, B. Pranitha, B. K. Reddy, D. Varsha, & N. N. Reddy. "Bone Fracture Detection and Classification Using Deep Learning Techniques." In *Driving Smart Medical Diagnosis Through AI-Powered Technologies and Applications* pp. 92-101, 2024. IGI Global.
- [17] H. A. Alwzawzy, L. Alzubaidi, Z. Zhao, & Y. Gu. "FracNet: An end-to-end deep learning framework for bone fracture detection." *Pattern Recognition Letters*, 2025.
- [18] A. Alam, A. S. Al-Shamayleh, N. Thalji, A. Raza, E. A. Morales Barajas, E. B. Thompson, & I. Ashraf. "Novel transfer learning-based bone fracture detection using radiographic images." *BMC Medical Imaging*, vol.25, no.1, 5, 2025.
- [19] W. Wei, Y. Huang, J. Zheng, Y. Rao, Y. Wei, X. Tan, & H. OuYang. "YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images." *Journal of Radiation Research and Applied Sciences*, vol.18 no.1, 101309, 2025.
- [20] A. Abdusalomov, S. Mirzakhilov, S. Umirzakova, O. Ismailov, D. Sultanov, R. Nasimov, & Y. I. Cho. "Lightweight Deep Learning Framework for Accurate Detection of Sports-Related Bone Fractures." *Diagnostics*, vol.15 no.3, 271, 2025.
- [21] T. Aldhyani, Z. A. Ahmed, B. M. Alsharbi, S. Ahmad, M. H. Al-Adhaileh, A. H. Kamal, & J. Nazeer. "Diagnosis and detection of bone fracture in radiographic images using deep learning approaches." *Frontiers in Medicine*, vol.11, 1506686, 2025.
- [22] M. I. Khan, J. Amin, M. A. Shehzad, & S.Iqbal. "AN INNOVATIVE HYBRID MODEL FOR ELBOW BONE FRACTURE DETECTION: INTEGRATING VIT AND CNN." *Contemporary Journal of Social Science Review*, vol.3, no.1, pp.1-21, 2025.
- [23] A. K. Pandey, S. Kedarnath, I. K. Argyros, & G. P. Pateel. "Improving Vertebral Fracture Detection in C-Spine CT Images Using Bayesian Probability-Based Ensemble Learning." *Algorithms*, vol.18, no.4, 181, 2025.
- [24] Z. Zheng, B. Y. Ryu, S. E. Kim, D. S. Song, S. H. Kim, J. W. Park, & D. H. Ro. "Deep learning for automated hip fracture detection and classification: achieving superior accuracy." *The bone & joint journal*, vol.107, no.2, pp.213-220, 2025.
- [25] D. Sutradhar, N. M. Fahad, M. A. K. Raiaan, M. Jonkman, & S. Azam. "Cervical spine fracture detection utilizing YOLOv8 and deep attention-based vertebrae classification ensuring XAI." *Biomedical Signal Processing and Control*, vol.101, 107228, 2025.
- [26] <https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data>
- [27] LeCun Y, Bengio Y, Hinton G. Deep learning. *nature*. 2015 May 28;521(7553):436-44.
- [28] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016 (pp. 770-778).
- [29] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2017 (pp. 4700-4708).
- [30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- [31] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* 2015 Oct 5 (pp. 234-241). Cham: Springer international publishing.

Authors' Profiles



Nandkishor Narkhede, Associate Professor, Department of Electronics and Computer Science, Shah and Anchor Kutchhi Engineering College, Mumbai, India. He holds a Ph.D. in Electronics and Communications with specialization in Artificial Intelligence and Machine Learning. His research interests include Artificial Intelligence, Machine Learning, Medical Image Analysis, and Intelligent Systems. He has published journal articles, conference papers, and Springer book chapters.



D. Girish Kumar, Assistant Professor, Department of Electronics and Communication Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India. He earned his Ph.D. in Electronics and Communication Engineering from SRM Institute of Science and Technology, Chennai. His research interests include Image Processing, Deep Learning, Computer Vision, and Multispectral Image Fusion.



Dhanya Job, Assistant Professor, Department of Computer Applications, Baselios Poulse Second Catholicose College, Piravom, Kerala, India. She holds a Ph.D. in Computer Science from Mahatma Gandhi University. Her research interests include Data Security, Steganography, Artificial Intelligence Applications, and Educational Technologies. She is also the author of the book *Responsive Web Designing*.



K. Varada Rajkumar, Associate Professor, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), MLR Institute of Technology, Hyderabad, India. He received his Ph.D. from KL University. His research interests include Machine Learning, Cyber Security, Cloud Computing, and Software Engineering.



Tirumalasetti Lakshmi Narayana, Assistant Professor, Department of Electrical and Electronics Engineering, Aditya University, Andhra Pradesh, India. He is pursuing a Ph.D. at Jawaharlal Nehru Technological University, Kakinada. His research interests include Power Systems, State Estimation, Electrical Power Distribution Systems, Electrical Machines, and Power Electronics.



B. Loveswara Rao, Professor, Department of Electrical and Electronics Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India. He received his Ph.D. from KL University. His research interests include FACTS, Renewable Energy Systems, Electric Drives, and Power Systems. He is an IEEE Senior Member and has published extensively in international journals and conferences.

How to cite this paper:

Nandkishor Narkhede, D. Girish Kumar, Dhanya Job, K. Varada Rajkumar, Tirumalasetti Lakshmi Narayana, B. Loveswara Rao, "XrayBoneNet: Multi-Region Fracture Localization and OA Grading via Attention-Guided Transformers", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.4, pp. 165-183, 2026. DOI:10.5815/ijigsp.2026.04.09