

<sup>1</sup>C. Harriet Linda<sup>2</sup>B. Kanisha<sup>3</sup>S. Dalton Griffin<sup>4</sup>W. R. Sam Emmanuel

# Graph-Augmented Multi-Modal Deep Learning Framework for Automated Bone Fracture Detection and Reporting



**Abstract:** - Accurate and efficient bone fracture diagnosis is essential for timely medical intervention, yet conventional manual interpretation of medical images is time-consuming, prone to variability, and dependent on radiologist expertise. To address these challenges, this paper proposes a Graph-Augmented Multi-Modal Deep Learning Framework for Fracture Detection, leveraging the strengths of convolutional and graph-based learning techniques to enhance fracture identification and classification. The proposed model integrates multi-modal medical imaging data (X-rays, CT scans), improving its adaptability across different imaging techniques. Convolutional Neural Networks (CNNs) are employed for feature extraction, while Graph Neural Networks (GNNs) model spatial and structural relationships within bone fractures, enabling precise localization and classification, particularly in cases of overlapping, comminuted, and subtle fractures. Additionally, explainable AI (XAI) techniques, such as Grad-CAM and saliency maps, are incorporated to enhance interpretability, providing radiologists with a transparent understanding of AI-driven diagnoses. To streamline clinical workflows, the system generates structured diagnostic reports, detailing fracture type, severity, and localization, ensuring consistency and reducing reporting time. The proposed framework is rigorously evaluated on multi-modal and real-world datasets, demonstrating its effectiveness in improving diagnostic accuracy, reducing human error, and enhancing clinical decision-making. By bridging the gap between AI-driven automation and radiological expertise, this research contributes to the advancement of intelligent medical imaging systems, making fracture diagnosis more efficient, accurate, and accessible in diverse healthcare settings.

**Keywords:** Bone Fracture Report, Graph Neural Networks, Medical Image Analysis, Report Generation, AI-Assisted Diagnosis.

## I. INTRODUCTION

Fracture diagnosis is a crucial aspect of medical imaging, traditionally performed by radiologists who analyze X-rays, CT scans, and MRI images to generate reports. However, manual reporting is often time-consuming and prone to human errors, leading to inconsistencies in diagnosis. Radiologists handle large volumes of cases, increasing the risk of oversight and delays in patient care. Additionally, rural and underdeveloped areas may lack access to experienced radiologists, making timely and accurate fracture diagnosis a challenge. Automated fracture report generation is powered by artificial intelligence (AI) and machine learning (ML) that offers a solution to these challenges. AI-driven systems can rapidly analyze medical images and detect fractures with high precision and generate structured reports within seconds. These technologies not only improve diagnostic accuracy but also ensure consistency across different medical institutions. AI models must be continuously trained on diverse datasets to ensure unbiased and reliable results. Moreover, maintaining a balance between automation and human oversight is essential to avoid misdiagnoses and ensure ethical use of AI in healthcare. With ongoing advancements, automated fracture report generation has the potential to revolutionize radiology, making diagnostic processes faster, more accurate, and accessible to a broader population.

Bone fractures are commonly diagnosed using imaging techniques such as X-rays, CT scans, and MRIs. While these methods are widely used and effective, they present several limitations that can impact the accuracy and efficiency of fracture detection and treatment. One of the primary challenges is the dependency on human interpretation, which can lead to variability in diagnoses. Radiologists and medical professionals must manually analyze images, and factors such as fatigue, experience level, and workload can result in missed or incorrect diagnoses. Additionally, subtle fractures such as hairline cracks or stress fractures may not always be clearly visible in standard X-rays, requiring additional imaging or follow-up scans to confirm the diagnosis. This not only delays treatment but also increases healthcare costs and patient discomfort. Furthermore, accessibility to advanced imaging technology is limited in many rural and underdeveloped regions, where medical facilities may lack the necessary equipment or trained professionals to provide timely and accurate diagnoses. As a result, patients in these areas may experience delays in receiving appropriate medical care, leading to complications or prolonged recovery times. Moreover, radiation exposure from repeated X-rays and CT scans raises concerns about patient safety,

<sup>1\*,2</sup> Department of Computing Technologies, SRM University of Science and Technology, Kattankulathur – 603 203, Tamil Nadu, India

<sup>3</sup> Department of Mathematics, Loyola College Nungambakkam, Chennai – 600034, Tamil Nadu, India

<sup>4</sup> Department of Computer Science, Nesamony Memorial Christian College, Marthandam – 629 165, Tamil Nadu, India

Copyright © JES 2024 on-line : journal.esrgroups.org

particularly for individuals who require frequent imaging, such as athletes or patients with osteoporosis. Additionally, manual reporting processes can slow down diagnosis and treatment planning, especially in high-volume healthcare settings where radiologists must analyze large numbers of cases daily. The subjectivity of human interpretation also contributes to inconsistency in diagnoses and treatment recommendations, leading to potential discrepancies in patient care. To overcome these limitations, advancements in medical imaging technology, including artificial intelligence and deep learning-based fracture detection systems, are being explored. These innovations have the potential to enhance accuracy, speed up diagnosis, and improve access to high-quality medical care, particularly in resource-limited settings. However, integrating AI-driven diagnostic tools into healthcare systems requires overcoming challenges related to data availability, regulatory approvals, and professional acceptance. Despite these hurdles, the evolution of imaging technologies and AI-assisted diagnostics offers promising solutions to improve bone fracture diagnosis and patient outcomes in the future.

Artificial Intelligence (AI) is revolutionizing radiology by improving the accuracy, speed, and efficiency of medical image analysis. AI-powered algorithms, particularly deep learning and computer vision techniques, can analyze vast amounts of imaging data within seconds, identifying fractures, tumors, and other abnormalities with high precision. These systems assist radiologists in making faster and more accurate diagnoses, reducing the risk of misinterpretation and enhancing overall patient care. One of the significant advantages of AI in radiology is its ability to automate repetitive tasks, allowing radiologists to focus on complex cases that require human expertise. AI-driven tools can prioritize critical cases, flagging potential abnormalities for immediate attention, which is especially crucial in emergency scenarios. Additionally, AI can detect subtle patterns in medical images that might be overlooked by the human eye, leading to earlier and more accurate disease detection. By integrating AI with Picture Archiving and Communication Systems (PACS) and Electronic Health Records (EHR), hospitals and diagnostic centers can streamline workflows, reduce reporting time, and improve efficiency in patient management. Beyond diagnosis, AI plays a crucial role in treatment planning and predictive analytics. Machine learning models can assess disease progression, predict patient outcomes, and assist doctors in choosing the most effective treatment strategies. AI can also help in medical research by analyzing vast datasets to uncover new insights into diseases and their patterns. Moreover, AI should complement radiologists rather than replace them, ensuring that final decisions are made with human oversight.

Deep learning has made significant strides in revolutionizing bone fracture detection and diagnosis, offering new opportunities for improving accuracy and efficiency in medical imaging. Traditionally, diagnosing bone fractures relied on the expertise of radiologists who analyzed X-rays, CT scans, and MRIs. However, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown remarkable capabilities in automating and enhancing this process. CNNs can analyze medical images and automatically detect fractures, even subtle ones, that may be difficult for human radiologists to identify, especially in complex anatomical regions. These models learn to recognize patterns such as bone discontinuities, misalignments, and other indicators of fractures by processing large datasets of annotated medical images. One of the key advantages of deep learning in bone fracture analysis is its ability to handle large volumes of medical images efficiently. Additionally, deep learning algorithms are capable of segmenting images to isolate the fracture from the surrounding tissues, providing precise localization of the injury. This is particularly beneficial for treatment planning, as it allows healthcare providers to focus on the affected area and monitor the healing process over time. Despite many advantages, challenges remain in the application of deep learning to bone fracture detection. One of the main issues is the availability of large, diverse, and well-annotated datasets for training these models. Bone fractures can vary widely in appearance, and obtaining comprehensive datasets that represent the full range of fracture types is essential for developing robust AI models. Nevertheless, as technology continues to advance and more data becomes available, the integration of deep learning into clinical practices for bone fracture diagnosis holds tremendous promise, enhancing the capabilities of healthcare providers and ultimately improving patient care.

Parvin, S., et. al., 2024 proposed a deep-learning model called YOLO v8 for real-time human bone fracture detection and classification. They utilized a multi-modal human bone fractures image dataset consisting of 641 images across ten fracture classes. To address overfitting due to the small dataset, various data augmentation techniques were applied. This model's effectiveness was tested through three experiments to evaluate its ability to categorize both healthy and broken bones from multi-modal images. Windarto, A.P., et. al., 2024 proposed a CNN architecture designed for highly accurate bone fracture classification. This study tackled the shortcomings of traditional methods by leveraging CNNs' ability to automatically extract hierarchical features from medical images. By leveraging a dataset obtained from Kaggle's public medical image repository, the research aimed to enhance diagnostic accuracy in orthopedics. This model achieved an impressive accuracy of 0.9922, outperforming

ResNet50's accuracy of 0.9844. These findings suggest that CNN-based systems can significantly improve diagnostic precision, optimize treatment strategies, and enhance patient outcomes in medical practice. Mittal, K., et. al., 2024 introduced a sequential CNN model for bone fracture detection and classification. Using a dataset of 4906 X-ray images divided into fractured and non-fractured classes, the study aimed to enhance fracture diagnosis accuracy. This model was trained on 4099 images, tested on 401 images, and validated using an additional 406 images, achieving an impressive classification accuracy of 98%. The research highlights the potential of deep learning in medical imaging to deliver precise and efficient fracture diagnoses, contributing to significant advancements in healthcare outcomes. Alshahrani, A., et. al., 2024 proposed a deep learning-based bone fracture classification system using X-ray images. They compared the performance of YOLOv8 which is known for real-time object detection and segmentation, along with VGG16 model. Using the FracAtlas dataset, which included 4,083 X-ray images of fractured and non-fractured bones, the research applied hyperparameter tuning and data augmentation to enhance detection accuracy. This system demonstrated superior performance compared to existing methods, showcasing the potential of CNN architectures to improve medical diagnostics and assist surgeons with precise and efficient fracture detection and classification.

M Fariz Fadillah, M., et. al., 2024 introduced a CNN model for bone fracture detection and classification in X-ray images, that focused on reducing diagnostic errors and enhancing efficiency in the orthopedic field. This research supported the 3rd Sustainable Development Goal (SDG) of promoting good health and well-being by contributing to innovative and equitable healthcare solutions. These findings are expected to significantly enhance fracture diagnostics and pave the way for advanced diagnostic technologies. Chauhan, S., 2024 proposed a CNN-based approach using the AlexNet model for bone fracture detection and classification from radiographic X-ray images. They utilized a dataset of 10,580 images covering various anatomical areas such as lower and upper limb, lumbar, hips, and knees, divided into training (9,246 images), validation (828 images), and test (506 images) subsets. This CNN-AlexNet model achieved 96% accuracy on the test set, demonstrating its effectiveness in distinguishing fractured and non-fractured bone X-rays. This research contributed to medical imaging advancements, aligns with sustainable development goals by promoting health and well-being, and supports innovation and infrastructure in healthcare diagnostics. Ali, S.N.E., et. al., 2024 proposed a machine learning model along with Resnet50 and Faster RCNN model for long bone fracture classification and detection. This study employed both binary and multi-class classification, alongside a detection model, to analyze X-ray images. Binary classification used Model A and Model B (for grayscale images), and a fine-tuned ResNet50 model (for RGB images), achieved accuracies of 90.2%, 90.85%, and 96.5%, respectively. Multi-class classification for fracture type identification using ResNet50 attained 87.7% accuracy, while the Faster RCNN model achieved 80% accuracy in fracture detection and localization. The dataset was annotated based on Müller AO classification, highlighted the effectiveness of these methods in enhancing fracture diagnosis accuracy.

Zou, J., et. al., 2024 proposed an improved YOLO v7 model for whole-body bone fracture detection, focusing on four fracture morphologies namely, angle fractures, line fractures, messed-up angle fractures and normal fractures. This study compared one and two stage deep learning architectures, including YOLO variants (v4, v5, v7, v8), SSD, Faster-RCNN, and Mask-RCNN. The customized YOLO v7-ATT model, incorporating an Enhanced Intersection of Unions (EIou) loss function and attention mechanisms. It achieved remarkable performance, reaching a mAP of 80.2% on general datasets and 86.2% on FracAtlas dataset outperforming other models. This system highlighted its clinical applicability and provides a foundation for optimizing deep learning models in medical imaging. Bittner-Frank, M., et. al., 2024 conducted a study to assess the accuracy of 3D bone fracture models derived from various CT imaging technologies and segmentation methods. This study focused on factors such as CT technology type (EID vs. PCD), four scanner type, two scan protocols, two orientations and two segmentation algorithms using twenty forearm specimens with simulated Colles' fractures. Results indicated that these factors significantly affected model accuracy, but the mean absolute deviation remained below 0.5 mm, meeting the requirements for pre-operative planning. This study highlighted the impact of segmentation errors and suggested manual corrections. These findings demonstrated that 3D bone models from routine clinical scanners are accurate enough for reliable pre-operative planning in orthopedic surgery. Murrad, B.G., et. al., 2024 proposed an AI driven framework for bone fracture detection in orthopedic therapy, utilizing the YOLOv8 model with a ResNet backbone. This combination enhances feature extraction and fracture classification accuracy within X-ray images. This model achieved a mean average precision of 0.9 and classification accuracy of 90.5%, significantly outperforming traditional methods. This framework provided healthcare professionals with an automated tool for improving diagnostic efficiency, accuracy and patient care in both routine and emergency care settings.

Ju, R.Y., et. al., 2023 applied the YOLO v8 algorithm for fracture detection in pediatric wrist trauma X-ray images. By utilizing data augmentation on the GRAZPEDWRI-DX public dataset, the study achieved state-of-the-art performance with a mean average precision (mAP 50) of 0.638, surpassing both improved YOLOv7 (0.634) and original YOLOv8 (0.636). To assist pediatric surgeons in diagnosing fractures, they developed "Fracture Detection Using YOLOv8 App," aimed at reducing diagnostic errors and providing valuable information for surgical decision-making. This approach demonstrated the potential of deep learning models in enhancing fracture detection accuracy in medical imaging. Beyraghi, S., et. al., 2023 proposed a deep neural network approach for bone fracture diagnosis using microwave S-parameters profiles, eliminating the need for labelling and data collection issues associated with X-ray images. This model classified different fracture types such as normal, transverse, oblique and comminuted and estimates crack length. Designed for portable use in settings like ambulances and low-income areas, the system enables fast, non-invasive diagnosis without ionizing X-rays. Experimental results with sheep femur bones demonstrated accurate classification, showcasing the potential for safe, rapid fracture detection in emergency situations. Khan, A.A., et. al., 2024 reviewed best practice recommendations for diagnosing and evaluating osteoporotic or fragility fractures, which are indicative of compromised bone strength and carry significant morbidity and mortality. Despite the clinical challenges, such fractures often go undiagnosed as being associated with underlying metabolic bone disease. These consensus guidelines emphasized the need for further evaluation and treatment to reduce future fracture risks, even in patients with bone mineral density above  $-2.5$ . A dedicated vertebral imaging review is recommended for high-risk patients. This underscored the importance of using a classification system for consistent fracture identification and reporting. Singh, A., 2024 proposed BoneScanAI, a hybrid machine learning model combining CNNs and Random Forest classifiers to enhance the accuracy of bone fracture diagnosis from X-ray images. This model used multiple CNN layers for deep feature extraction and followed by Random Forest for classification of seven kinds of fractures. The dataset consisted of 2,738 X-ray images and it is labelled by radiologists. This model achieved an accuracy of 86.99%, demonstrating its potential to assist doctors in precise fracture identification. They highlighted the effectiveness of CNN and Random Forest in medical image processing and suggests further refinement through expanded datasets and additional training for clinical application.

Su, Z., et. al., 2024 proposed a multimodal diagnostic model, BoneCLIP-XGBoost, for bone fracture detection. It combines both Vision Transformer (ViT) and ClinicalBERT for feature extraction, the model integrated X-ray images and textual descriptions into a unified feature space. This integration enhanced the alignment of multimodal data, addressing challenges in existing methods. This model achieved an accuracy of 88.5%, precision of 87.3%, recall of 86.8%, and an F1 score of 87.0%. BoneCLIP-XGBoost offered a robust, accurate, and reliable solution for bone fracture diagnosis, outperforming traditional methods. Pérez-Cano, F.D., et. al., 2024 proposed a methodology for enhancing medical diagnosis and treatment planning by automating the acquisition and classification of bone fracture patterns. The system extracted detailed fracture features using CT scans and classified them with a convolutional neural network. This approach aimed to streamline the fracture classification process, facilitating improved diagnostic accuracy, supporting surgical treatment planning, and advancing medical training and simulation applications. They emphasized the importance of automating fracture analysis for more effective patient care and medical education. Yu, Q., et. al., 2025 proposed MTL-DlinkNet, a multi-task learning model based on D-linkNet for calcaneus fracture diagnosis from X-ray images. This model performed both classification for fracture identification and segmentation for generating regions of interest (RoI). Achieving a high accuracy (0.989 AUC). This model improved diagnostic efficiency and reduced the burden on doctors by simultaneously annotating the data. Experimental results showed that MTL-DlinkNet outperforms baseline models, demonstrating the effectiveness of multi-task learning in enhancing fracture diagnosis accuracy and efficiency in clinical settings.

Potter, I.Y., et. al., 2024 proposed an automated pipeline for vertebrae localization, segmentation, and osteoporotic vertebral compression fracture (VCF) detection using CT images. This approach utilized deep learning models and was evaluated on a publicly available dataset of 325 spine CT scans, with 126 scans graded for VCF. This system achieved 96% sensitivity and 81% specificity for vertebral-level VCF detection and high accuracy at the subject-level. This addition of predicted vertebrae segments significantly improved VCF detection performance, increasing sensitivity by 14% and specificity by 20%. This approach outperformed other VCF detection methods and is poised to enhance diagnostic accuracy for osteoporosis-related fractures. Dibo, R., et. al., 2023 proposed DeepLOC, a deep learning-based approach for bone pathology localization and classification in wrist X-ray images. DeepLOC model integrated YOLO for real-time object detection and localization of bone pathologies, combined with the Shifted Window Transformer to extract contextual information for precise

classification. This approach addressed critical challenges in wrist X-ray analysis by accurately localizing abnormalities and classifying bone pathologies, providing enhanced support for radiologists in medical image analysis. Kumar, G., et. al., 2023 proposed an IoT-enabled intelligent imaging system for bone fracture detection. This system leveraged image processing techniques such as CLAHE, Gaussian blur, Canny edge detection and Hough Transform, combined with IoT infrastructure to automate fracture detection from X-ray images. This framework reduced human error and enhanced diagnostic efficiency by providing real-time data processing and feedback to patients. This system demonstrated high accuracy in detecting fractures, particularly in lower long bones, hand, and elbow bones, showing significant potential for improving diagnostic workflows and treatment outcomes. Zeng, B., et. al., 2023 proposed a two-stage method for the automatic identification and localization of complex pelvic fractures using a novel structure-focused contrastive learning approach. This method combined the symmetry properties of pelvic anatomy and leveraged a Siamese deep neural network with a structural attention mechanism to improve fracture zone detection. This proposed system achieved a mean accuracy of 0.92 and sensitivity of 0.93 on a dataset of 103 clinical CT scans from the CTPelvic1K dataset, outperforming three state-of-the-art contrastive learning methods and five advanced classification networks. These results demonstrate the model's effectiveness in handling the complexities of pelvic fractures. Linda, C.H., et. al., 2011 presented a novel image processing algorithm for the automated detection of bone fractures in X-ray images, addressing the critical need for accurate and timely diagnoses. This approach employed a multi-stage process which includes grid formation, local thresholding with interpolation, fuzzy index-based segmentation, background removal and morphological filtering to enhance the accuracy and reliability of fracture identification.

Despite the growing adoption of AI in medical imaging, automated bone fracture detection still faces several challenges. Existing deep learning models, particularly CNN-based approaches, primarily focus on feature extraction from single-modality data, limiting their ability to generalize across different imaging techniques such as X-rays, CT scans, and MRIs. Moreover, CNNs often struggle with complex fracture cases involving overlapping bone structures, comminuted fractures, and subtle hairline fractures due to their inability to capture spatial and structural relationships within the bone. While Graph Neural Networks (GNNs) have shown promise in medical imaging tasks, their potential for bone fracture detection—particularly in modeling spatial dependencies between fracture fragments—remains underexplored. Additionally, most AI-driven diagnostic models function as black-box systems, offering little interpretability for clinical use. The lack of explainable AI (XAI) mechanisms, such as saliency maps or attention-based visualizations, raises concerns regarding trust and adoption in real-world radiology workflows. Furthermore, many existing models are trained on limited datasets, restricting their generalizability across diverse patient populations and fracture types.

Based on the foundations of our earlier work Linda, C.H., et. al., 2011 this research proposes a Graph-Augmented Multi-Modal Deep Learning Framework for fracture detection, incorporating both convolutional and graph-based learning techniques to enhance fracture identification and classification. The key contributions of this work are:

**Multi-Modal Fracture Detection and Classification:** The proposed framework integrates multi-modal medical imaging data (X-ray and CT) to improve the robustness and generalizability of fracture detection across different imaging techniques.

**Graph-Augmented Structural Feature Learning:** By incorporating GNN-based spatial modelling, the framework captures intricate structural relationships within bone fractures, enabling better classification of complex, overlapping, and comminuted fractures that traditional CNNs struggle with.

**Improved Interpretability with Explainable AI (XAI) Techniques:** The model integrates attention-based visualizations (Grad-CAM) to enhance interpretability, allowing radiologists to validate and trust AI-generated predictions.

**Automated and Standardized Fracture Report Generation:** The system generates structured diagnostic reports, providing insights into fracture type, severity, and localization, which assist radiologists in decision-making and streamline clinical workflows.

**Comprehensive Performance Evaluation on Multi-Modal Datasets:** The framework is rigorously evaluated using our own collected real-world datasets, ensuring its clinical viability and applicability in different healthcare settings.

The paper is structured as follows. Section 1: Introduction provides an overview of existing AI-based fracture detection methods, highlighting their strengths, limitations, and advancements in deep learning for medical imaging. Section 2: Proposed Methodology details the hybrid CNN-GNN architecture, including data preprocessing, model design, and training procedures. Section 3: Experimental Setup and Dataset describes the

datasets used, evaluation metrics, and implementation details, outlining the model training and validation process. Section 4: Results and Discussion presents experimental findings, including comparative performance analysis and visualizations of model predictions. Finally, Section 5: Conclusion and Future Work summarizes key insights, discusses the potential impact of the proposed method, and outlines future research directions to enhance AI-driven radiology applications.

## II. METHODOLOGY

The proposed Graph-Augmented Multi-Modal Deep Learning Framework is designed to leverage both convolutional neural networks for feature extraction and graph neural networks for spatial and structural relationship modelling to enhance the detection, classification, and reporting of bone fractures. The framework efficiently processes multi-modal medical imaging data (X-ray, CT scans and ground truth binary masks) and integrates radiology reports for automated diagnostic assistance. Figure 1 depicts the entire working architecture of the graph augmented multi-modal deep learning framework.

### A. Data Preprocessing and Augmentation

To ensure consistency across different imaging modalities, a comprehensive preprocessing pipeline is implemented which includes normalization, denoising, data augmentation, and feature alignment. Normalization is a technique that standardizes pixel intensity values to a fixed range, preventing disparities between images from affecting model performance. In this paper, min-max normalization is applied to scale pixel values between 0 and 1 to ensure uniform intensity distribution across the dataset. Denoising is performed using Gaussian filtering, which helps to reduce random noise while preserving important structural details in medical images. A Gaussian filter with a kernel size of  $3 \times 3$  and a standard deviation ( $\sigma$ ) of 1.0 is applied to smooth the image by averaging pixel intensities within a local neighborhood, thereby improving the clarity of anatomical structures.

Data augmentation techniques are employed to enhance model generalization by introducing variations in the dataset. The following transformations are applied:

- Rotation: Random rotation within the range of  $\pm 15$  degrees to simulate variations in orientation.
- Scaling: Random scaling within the range of 90% to 110% to account for slight changes in size.
- Affine Transformations: Shearing within the range of  $\pm 10$  degrees to introduce realistic geometric distortions.

Feature alignment is another critical preprocessing step, particularly for multimodal imaging studies. To ensure accurate pixel-wise correspondence between X-ray and CT scans, intensity-based image registration techniques are applied using mutual information optimization. These methods adjust spatial alignment by optimizing transformation parameters to match structures across modalities, thereby improving the integration of complementary imaging data. By incorporating these preprocessing steps with precisely tuned parameters, the quality and consistency of input images are enhanced, leading to improved performance in subsequent classification or analysis tasks.

### B. Multi-Modal Data Integration

The Multi-Modal Feature Extraction module is designed to process X-ray images, CT scan slices, and binary masks, ensuring that both 2D spatial features and 3D volumetric structures are effectively captured. The module begins by applying separate convolutional pipelines for X-ray and CT images. The X-ray input passes through a 2D CNN pipeline, where two consecutive Conv2D layers (with  $3 \times 3$  filters) extract edge and texture features, followed by a max pooling layer to reduce spatial dimensions while retaining essential fracture information. Meanwhile, the CT scan input, which consists of multiple slices, is processed using a 3D CNN pipeline that applies 3D convolution operations with  $3 \times 3 \times 3$  filters, allowing the model to capture the depth and structural integrity of bones. After individual feature extraction, the outputs from both X-ray and CT pathways are fused using an attention-based mechanism, ensuring that the most informative features from each modality contribute to the final representation. Additionally, a binary mask input, indicating the exact fracture region, is incorporated by performing element-wise addition with the fused feature maps, refining the model's focus on clinically relevant areas. This multi-scale feature fusion ensures that both local (X-ray-based) and structural (CT-based) patterns are efficiently captured, providing a rich, modality-aware representation for the subsequent graph-based structural learning and classification tasks.

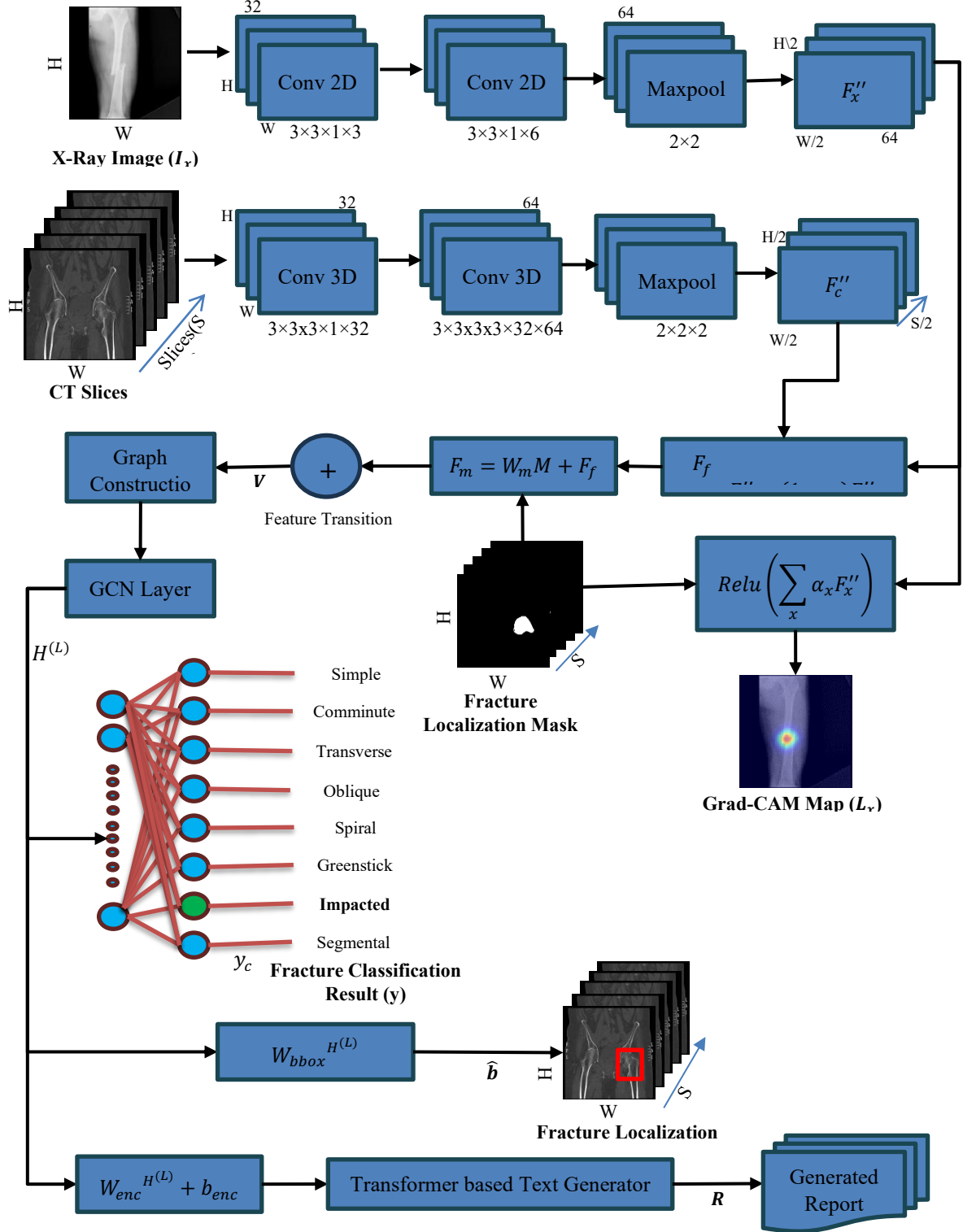


Fig. 1: Architecture Diagram for Graph Augmented Multi-modal Deep Learning Framework

### C. CNN Architecture for Feature Extraction

The Multi-Modal Feature Extraction module processes X-ray images, CT scan slices, and binary masks to extract both 2D spatial features and 3D volumetric structures. Let,  $I_x \in \mathbb{R}^{H \times W \times 1}$  be the X-ray image, where  $H$  and  $W$  are height and width, respectively.  $I_c \in \mathbb{R}^{H \times W \times S \times 1}$  be the CT scan slices, where SSS represents the number of slices.  $M \in \mathbb{R}^{H \times W \times 1}$  be the binary mask, indicating the fracture region in the image. The module applies separate convolutional pipelines for X-ray and CT images, processing them through 2D and 3D convolutional networks, respectively. Table 1 shows the network structure of the multi-modal feature extraction module.

Step 1: Feature Extraction from X-ray Images

The X-ray image  $I_x$  is passed through two Conv2D layers, which extract edge and texture features. The transformation follows:

$$F_x = \sigma(W_x * I_x + b_x), W_x \in \mathbb{R}^{3 \times 3 \times 1 \times 32}, b_x \in \mathbb{R}^{32} \quad (1)$$

The input size is  $H \times W \times 1$  (single-channel grayscale X-ray). The output size after Conv1\_X is  $H \times W \times 32$ . The second convolution layer applies another  $3 \times 3$  kernel:

$$F'_x = \sigma(W'_x * F_x + b'_x), W'_x \in \mathbb{R}^{3 \times 3 \times 32 \times 64}, b'_x \in \mathbb{R}^{64} \quad (2)$$

The output size after Conv2\_X is  $H \times W \times 64$ . A max pooling layer ( $2 \times 2$ ) is applied to downsample the feature map:

$$F''_x = \text{MaxPool}(F'_x), \text{Output Size: } \frac{H}{2} \times \frac{W}{2} \times 64 \quad (3)$$

Step 2: Feature Extraction from CT Scan Slices

The CT scan input  $I_c$ , consisting of multiple slices, is processed using a 3D CNN pipeline to capture depth information:

$$F_c = \sigma(W_c * I_c + b_c), W_c \in \mathbb{R}^{3 \times 3 \times 3 \times 1 \times 32}, b_c \in \mathbb{R}^{32} \quad (4)$$

The input size is  $H \times W \times S \times 1$  (multi-slice CT scan). The output size after Conv1\_C is  $H \times W \times S \times 32$ . The second 3D convolution layer applies another  $3 \times 3 \times 3$  kernel:

$$F'_c = \sigma(W'_c * F_c + b'_c), W'_c \in \mathbb{R}^{3 \times 3 \times 3 \times 32 \times 64}, b'_c \in \mathbb{R}^{64} \quad (5)$$

The output size after Conv2\_C is  $H \times W \times S \times 64$ . A 3D max pooling layer ( $2 \times 2 \times 2$ ) reduces spatial dimensions:

$$F''_c = \text{MaxPool}(F'_c), \text{Output Size: } \frac{H}{2} \times \frac{W}{2} \times \frac{S}{2} \times 64 \quad (6)$$

Step 3: Multi-Scale Feature Fusion using Attention Mechanism

To combine the extracted features from both X-ray and CT modalities, we employ attention-based feature fusion:

$$F_f = \alpha F''_x + (1 - \alpha) F''_c \quad (7)$$

where the attention weight  $\alpha$  is computed as:

$$\alpha = \frac{\exp(W_f F''_x)}{\exp(W_f F''_x) + \exp(W_f F''_c)} \quad (8)$$

Where,  $W_f$  is the trainable weight matrix. The output size after fusion is  $\frac{H}{2} \times \frac{W}{2} \times 64$ .

Step 4: Fracture Mask Integration

To enhance localization, the binary mask  $M$  (which highlights fracture regions) is integrated into the feature map:

$$F_m = W_m M + F_f, W_m \in \mathbb{R}^1 \quad (9)$$

Table 1: Network structure of multi-modal feature extraction module

Layer	Type	Input Shape	Output Shape	Parameters	Activation
Input X-ray	Input Layer	(H,W,1)	(H,W,1)	-	-
Conv1_X	Conv2D (3x3)	(H,W,1)	(H,W,32)	$3 \times 3 \times 1 \times 32$	ReLU
Conv2_X	Conv2D (3x3)	(H,W,32)	(H,W,64)	$3 \times 3 \times 32 \times 64$	ReLU
MaxPool_X	MaxPool (2x2)	(H,W,64)	(H/2,W/2,64)	-	-
Input CT	Input Layer	(H,W,S)	(H,W,S)	-	-
Conv1_C	Conv3D (3x3x3)	(H,W,S,1)	(H,W,S,32)	$3 \times 3 \times 3 \times 1 \times 32$	ReLU
Conv2_C	Conv3D (3x3x3)	(H,W,S,32)	(H,W,S,64)	$3 \times 3 \times 3 \times 32 \times 64$	ReLU
MaxPool_C	MaxPool3D (2x2x2)	(H,W,S,64)	(H/2,W/2,S/2,64)	-	-
Feature Fusion	Attention Weighted Sum	(H/2,W/2,64)	(H/2,W/2,64)	-	Softmax
Feature Masking	Element-wise Addition	(H/2,W/2,64)	(H/2,W/2,64)	-	-

The input size is  $\frac{H}{2} \times \frac{W}{2} \times 1$ . The output size after mask integration remains  $\frac{H}{2} \times \frac{W}{2} \times 64$ . This forces the model to focus on relevant fracture areas, improving localization precision. The final feature map  $F_m$  containing both multi-modal fusion and mask integration, is converted into a graph representation for the Graph Neural Network module. Each spatial region in  $F_m$  becomes a node  $v_i$  in the graph:



$$V = \{v_1, v_2, \dots, v_N\}, \text{ where } N = \frac{H}{2} \times \frac{W}{2} \quad (10)$$

The graph edges  $E$  are constructed based on bone continuity and region similarity, enabling structural fracture analysis. 2D CNN extracts spatial features from X-rays. 3D CNN captures depth and structural integrity from CT scans. Attention-based feature fusion dynamically balances multi-modal inputs. Binary mask integration enhances localization precision. Output is transformed into a graph for GNN-based fracture classification. This optimized multi-modal feature extraction forms the foundation for fracture detection, localization, and explainable AI-based diagnostics.

#### D. Graph Neural Network for Structural Learning

The Graph Neural Network module is designed to model spatial and structural relationships between different fracture regions, enhancing the classification and localization of complex fractures. Unlike CNNs, which primarily focus on local feature extraction, GNNs enable global feature propagation by treating the extracted features as nodes in a graph and defining edges based on bone continuity and anatomical structure. Table 2 shows the network structure of the graph neural network.

Given a feature map  $F_m \in \mathbb{R}^{H/2 \times W/2 \times 64}$  extracted from the multi-modal CNN module, we construct a graph  $G = (V, E)$  where:  $V = \{v_1, v_2, \dots, v_N\}$  represents the nodes corresponding to different fracture regions in the feature map.  $E \subseteq V \times V$  represents edges that define the spatial relationships between nodes. Each node  $v_i$  is initialized with a feature vector  $h_i$  extracted from the CNN module:

$$h_i^{(0)} = F_m(i), h_i^{(0)} \in \mathbb{R}^d \quad (11)$$

where  $d$  is the feature dimension (e.g., 64 from CNN output). Edges are constructed based on geometric proximity and bone connectivity priors. The weighted adjacency matrix  $A$  is defined as:

$$A_{ij} = \begin{cases} \exp(-\beta \|h_i - h_j\|^2), & \text{if } \|h_i - h_j\| < d_t \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

Where,  $\beta$  is a scaling factor controlling the influence of distant nodes.  $d_t$  is a threshold distance, ensuring only meaningful connections. The adjacency matrix is normalized to prevent gradient explosion:

$$\hat{A} = D^{-1/2} A D^{-1/2} \quad (13)$$

where  $D$  is the degree matrix with  $D_{ii} = \sum_j A_{ij}$ . To refine node features, we apply a Graph Convolutional Network, updating each node's representation using information from its neighbors. The GCN updates each node's feature representation as follows:

$$h_i^{(l+1)} = \sigma \left( \sum_{j \in N(i)} \frac{A_{ij}}{\sum_{k \in N(i)} A_{ik}} W_g h_j^{(l)} \right) \quad (14)$$

Where,  $N(i)$  is the set of neighboring nodes of  $v_i$ ,  $W_g$  is a trainable weight matrix and  $\sigma(\cdot)$  is a non-linear activation function. The output of this layer aggregates information from neighboring nodes, improving fracture classification and localization. A stack of GCN layers allows the model to capture high-order relationships:

$$H^{(l+1)} = \sigma(AH^{(l)}W_g^l) \quad (15)$$

Where,  $H^{(l)}$  is the node feature matrix at layer  $l$ .  $W_g^{(l)}$  is the trainable weight matrix for layer  $l$ . After  $L$  layers, the output is a refined feature representation:

$$H^{(L)} = \{h_1^{(L)}, h_2^{(L)}, \dots, h_N^{(L)}\} \quad (16)$$

which is passed to the classification and localization module. Instead of treating all neighbors equally, Graph Attention Networks (GAT) assign learnable attention weights:

$$h_i^{(l+1)} = \sigma \left( \sum_{j \in N(i)} \alpha_{ij} W_g h_j^{(l)} \right) \quad (17)$$

where the attention coefficient  $\alpha_{ij}$  is computed as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(W_a[h_i \| h_j]))}{\sum_{k \in N(i)} \exp(\text{LeakyReLU}(W_a[h_i \| h_k]))} \quad (18)$$

Where,  $W_a$  is a learnable attention weight matrix.  $\|$  represents concatenation of node features. A softmax function normalizes the attention scores. This mechanism allows the model to focus on important fracture regions, improving classification performance. The final GNN output is used for fracture classification and localization. Therefore, graph construction converts CNN feature map into a graph structure. Graph convolution network aggregates node features from neighboring regions. graph attention assigns importance to fracture regions for better prediction. This uses a softmax classifier to predict fracture type and predicts bounding boxes for fractures. This

GNN module significantly enhances fracture analysis by modelling structural dependencies, making it robust for complex and overlapping fractures.

Table 2: Network structure of GNN

Layer	Type	Input Shape	Output Shape	Parameters	Activation
Graph Input	Node Embedding	(N,d)	(N,d)	-	-
GCN1	Graph Conv	(N,d)	(N,128)	$d \times 128d$	ReLU
GCN2	Graph Conv	(N,128)	(N,256)	$128 \times 256$	ReLU
Graph Attention	GAT	(N,256)	(N,256)	$256 \times 256$	Softmax

#### E. Explainable AI (XAI) Integration

In our Multi-Modal CNN Feature Extraction Module, we extract feature maps from X-ray images using separate convolutional pipelines,  $F_x'' \in \mathbb{R}^{H/2 \times W/2}$ , where,  $H/2, W/2$  are the downsampled spatial dimensions after convolution and pooling. These feature maps encode spatial and depth-related fracture patterns, which Grad-CAM utilizes to identify critical fracture regions. We compute Grad-CAM using the last convolutional layer output  $F_x''$ . The importance score  $\alpha_x$  for each feature map  $F_k$  is calculated as

$$\alpha_x = 1/Z \sum_{i,j} \frac{\partial m}{\partial F_k} \quad (19)$$

Where,  $m$  is the fracture mask,  $\frac{\partial m}{\partial F_k}$  is the gradient of fracture mask w.r.t. feature map  $F_k$  at location  $(i,j)$ ,  $F_k = \cup (F_x'') \in \mathbb{R}^{H \times W}$ , here  $\cup$  is bilinear interpolation.  $Z = H \times W$  is a normalization factor. The X-ray Grad-CAM heatmap ( $L_x$ ) is then computed as

$$L_x = \text{Relu}(\sum_x \alpha_x F_x'') \quad (20)$$

The ReLU function ensures that only positive activations contribute to the heatmap. Table 3 shows the network structure of the Explainable AI integration.

Table 3: Network structure of Explainable AI

Layer	Type	Input Shape	Output Shape	Parameters	Activation
Grad-CAM	Weighted Sum	(H,W,64)	(H,W)	-	ReLU

#### F. Fracture Localization, Classification and Report Generation

The Fracture Classification, Localization, and Report Generation modules form the final stages of the Graph-Augmented Multi-Modal CNN Framework. These modules leverage CNN-extracted features, GNN-enhanced structural representations, and transformer-based language models to accurately predict fracture type, localize the affected region, and generate a structured radiology report. Table 4 shows the network structure of the fracture classification and localization module, where Table 5 shows the structure of report generation module.

##### 1) Fracture Classification Module

The classification module predicts the type of fracture (e.g., Simple, Comminuted, Transverse, Oblique, Spiral, Greenstick, Impacted, Segmental). This is achieved using a fully connected neural network (FCN) classifier on the final GNN output  $H^{(L)}$ . The final node representations  $H^{(L)}$  obtained from the GNN module are aggregated into a global feature vector  $h_{global}$  by applying mean-pooling or attention-based aggregation:

$$h_{global} = \frac{1}{N} \sum_{i=1}^N h_i^{(L)} \quad (21)$$

Where,  $N$  is the number of nodes in the graph (representing different fracture regions).  $h_i^{(L)}$  is the final-layer representation of node  $i$ . Alternatively, attention-weighted aggregation can be used.

$$h_{global} = \sum_{i=1}^N \alpha_i h_i^{(L)}, \text{ where, } \alpha_i = \frac{\exp(W_a h_i^{(L)})}{\sum_{j=1}^N \exp(W_a h_j^{(L)})} \quad (22)$$

where  $W_a$  is a trainable weight matrix. The aggregated feature vector ( $H^{(L)}$ ) is passed through a fully connected neural network (FCN) for classification:

$$y = \text{softmax}(W_{clf} h_{global} + b_{clf}) \quad (23)$$

where,  $W_{clf}$  is the classification weight matrix and  $b_{clf}$  is the bias term. Softmax activation ensures a probability distribution over all fracture classes. The classification loss is computed using cross-entropy loss:

$$L_{CE} = - \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (24)$$

Where,  $C$  is the total number of fracture classes,  $y_c$  is the true class label (one-hot encoded) and  $\hat{y}_c$  is the predicted probability for class  $c$ .

### 2) Fracture Localization Module

The localization module predicts the bounding box  $(x, y, w, h)$  around the fracture site, ensuring precise detection in X-ray and CT scans. The bounding box prediction is formulated as a regression problem, where the model learns to predict:

$(x, y) \rightarrow$  Fracture center coordinates,  $(w, h) \rightarrow$  Bounding box width and height

Given the final GNN-enhanced feature representation  $H^{(L)}$ , the bounding box is predicted as follows:

$$\hat{b} = W_{bbox}H^{(L)} + b_{bbox} \quad (25)$$

Where,  $W_{bbox} \in R^{256 \times 4}$  is the bounding box regression weight matrix,  $b_{bbox} \in R^4$  is the bias term. The output  $\hat{b} \in R^4$  represents predicted bounding box coordinates. To optimize bounding box predictions, we use Intersection-over-Union (IoU) loss, which measures the overlap between the predicted bounding box  $\hat{B}$  and the ground-truth bounding box  $B_g$ :

$$L_{IoU} = 1 - \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \quad (26)$$

Where,  $B_p$  is the predicted bounding box,  $B_g$  is the ground-truth bounding box. To enhance stability, we use smooth  $L_1$  loss:

$$L_{bbox} = \sum_{i \in \{x, y, w, h\}} \text{SmoothL1}(b_i - b_i), \text{ where, } \text{SmoothL1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$$

Table 4: Network structure of Fracture Classification and Localization module

Layer	Type	Input Shape	Output Shape	Parameters	Activation
Flatten	Fully Connected	(N,256)	(N,512)	256×512	ReLU
Classifier	Fully Connected	(N,512)	(N,C)	512×C	Softmax
Bounding Box	Fully Connected	(N,512)	(N,4)	512×4	Linear

### 3) Automated Report Generation Module

The report generation module converts CNN-GNN extracted features into a structured radiology report using a transformer-based language model. The final GNN representation  $H^{(L)}$  is transformed into a sequence embedding  $E$ :

$$E = W_{enc}H^{(L)} + b_{enc} \quad (27)$$

Where,  $W_{enc}$  and  $b_{enc}$  are trainable weights. The output  $E \in R^{T \times d}$  serves as input to the NLP model. The encoded features are passed through a Transformer-based text generator:

$$R = \text{Decoder}(\text{Encoder}(E)) \quad (28)$$

Where,  $R$  is the generated radiology report. The language model is trained using sequence-to-sequence cross-entropy loss:

$$L_{NLP} = - \sum_{t=1}^M p(r_t | r_{<t}, E) \log(\hat{r}_t) \quad (29)$$

Where,  $M$  is the length of the generated report.  $p(r_t | r_{<t}, E)$  is the probability of word  $r_t$  given previous words. The combined loss function includes classification loss, localization loss, and report generation loss:

$$L_{total} = L_{CE} + \lambda_1 L_{bbox} + \lambda_2 L_{NLP} \quad (30)$$

Where,  $\lambda_1$  and  $\lambda_2$  are weighting factors. Therefore, the fracture classification module uses fully connected layers and softmax for multi-class prediction. Fracture localization module uses bounding box regression with IoU loss for precise fracture detection. Finally, the report generation module uses Transformer-based NLP model to create structured radiology reports. Loss function, optimizes classification, localization, and report quality simultaneously. The complete parameters count of graph-augmented multi-modal deep learning framework is tabulated in Table 6.

Table 5: Network structure of Report Generation module

Layer	Type	Input Shape	Output Shape	Parameters	Activation
Encoder	Transformer Encoder	(T,d)	(T,512)	d×512	ReLU
Decoder	Transformer Decoder	(T,512)	(T,V)	512×V	Softmax

Table 6: Total parameters of graph-augmented multi-modal deep learning framework

Component	Estimated Parameters
CNN Feature Extraction	82,000
GNN Structural Learning	100,000
Explainability (XAI)	1,000
Classification and Localization	150,000
NLP Report Generation	1,000,000
<b>Total Parameters</b>	<b>1.33 Million</b>

### III. EXPERIMENTAL SETUP AND DATASET

#### A. Dataset Description

This study utilizes a comprehensive multi-modal dataset comprising medical imaging data from 1,250 bone fracture patients across various fracture types. Each patient's dataset includes X-ray images, CT scan slices, binary mask images specifying the fracture region and corresponding radiologist reports. The dataset covers a wide range of bone fracture types, including simple, comminuted, transverse, oblique, spiral, greenstick, impacted, and segmental fractures, ensuring diverse representation for robust AI-based detection and classification. The binary mask images provide precise ground-truth annotations for accurate localization and segmentation, while the radiologist reports contain expert observations on fracture classification, severity, affected bone regions and treatment recommendations, serving as valuable references for model training and evaluation. This dataset is the result of our intensive effort in data collection, which involved collaboration with private hospitals and diagnostic centers across Tirunelveli, Thoothukudi, and Nagercoil districts of Tamil Nadu, India. The data acquisition process spanned over 18 months (June 2022 – December 2023) and required meticulous coordination with medical professionals, ethical committees and imaging departments to ensure high-quality and diverse data representation. Every image and report were obtained following strict ethical guidelines and patient confidentiality protocols, reinforcing the integrity of our research.

To enhance the dataset's diversity and improve model generalization, we applied data augmentation techniques. Table 7 presents the data distribution before and after augmentation. The original dataset contained 1,250 samples, distributed across different fracture types. Augmentation was performed at varying fold levels based on the fracture type, leading to a total of 4,561 augmented samples. After selecting 490 samples per fracture type for the experiment, the final dataset used for model training and evaluation consisted of 4,410 images. This augmentation ensured a balanced representation of fracture types, enabling the proposed Graph-Augmented Multi-Modal CNN Framework to learn both spatial and structural relationships in fractures more effectively. By incorporating various fracture types, anatomical locations and patient demographics, this dataset enhances the generalizability of the proposed Graph-Augmented Multi-Modal CNN Framework for Fracture Detection. The availability of multi-modal imaging data enables the model to learn both spatial and structural relationships in fractures, ultimately improving fracture classification, localization and automated report generation, thereby supporting real-world radiological workflows.

Table 7: Data Distribution before and after augmentation

Fracture Type	Original Sample Count	Augment fold	Augmented Sample	Considered for Experiment
Simple	250	2	500	490
Comminuted	180	3	540	490
Transverse	166	3	498	490
Oblique	130	4	520	490
Spiral	126	4	504	490
Greenstick	131	4	524	490
Impacted	98	5	490	490
Segmental	99	5	495	490
Pathological	70	7	490	490
<b>Total</b>	<b>1250</b>		<b>4561</b>	<b>4561</b>

### B. Hardware and Software Setup

The experiments were implemented using Python 3.8 as the primary programming language, with deep learning models developed using TensorFlow 2.10 and PyTorch 1.13. Image preprocessing and augmentation were performed using OpenCV and PIL, while NumPy and Pandas facilitated data handling. For graph-based modelling, NetworkX was utilized, and visualization of results was done using Matplotlib and Seaborn. The study was conducted on a Windows 11 operating system environment. The hardware setup included a high-performance computing system equipped with an Intel Core i9-12900K CPU, 64GB DDR4 RAM, and a 2TB NVMe SSD for efficient data processing and model training. This setup ensured fast computations, enabling effective training and evaluation of the Graph-Augmented Multi-Modal CNN Framework for fracture detection and classification.

### C. Performance Analysis

The Graph-Augmented Multi-Modal CNN Framework for Automated Bone Fracture Detection and Reporting is an advanced deep learning pipeline that integrates multi-modal imaging data, graph-based structural learning, and natural language processing (NLP) to provide a comprehensive and automated diagnosis of bone fractures. The system processes X-ray and CT images, classifies fractures, localizes affected regions, and generates structured radiology reports, ensuring high accuracy and clinical reliability. The framework is trained and evaluated using a large dataset of 1250 patients, each with corresponding X-ray images, CT scan slices, binary masks highlighting fracture regions, and ground-truth radiology reports authored by radiologists.

During the training phase, input images undergo preprocessing, which includes normalization, denoising, and image registration, ensuring consistency across modalities. A 2D CNN extracts spatial features from X-ray images, while a 3D CNN captures volumetric features from CT scans, allowing the model to analyze both surface-level and deep structural information. To effectively combine X-ray and CT data, an attention-based fusion mechanism assigns weights to the extracted features, prioritizing the most relevant information. These fused features are then structured as a graph, where nodes represent fracture regions, and edges define spatial relationships based on anatomical structures. A Graph Convolutional Network (GCN) propagates information between neighboring regions, while a Graph Attention Network (GAT) emphasizes critical fracture areas, enhancing the model's ability to differentiate between complex, overlapping, and subtle fractures. The final node representations are used for fracture classification, where a fully connected classifier predicts fracture type, and a bounding box regression module localizes the fracture using advanced object detection techniques. Simultaneously, the automated radiology report generation module is trained using a transformer-based NLP model, leveraging ground-truth reports written by radiologists. The CNN-GNN features serve as input to the transformer encoder, which learns to map the extracted medical image features to meaningful textual representations. The decoder then generates structured reports, mirroring the language and terminology used by radiologists. The model is trained using a sequence-to-sequence learning approach, ensuring that it accurately replicates the diagnostic insights found in expert-generated reports.

In the testing phase, the trained model processes new, unseen patient data, undergoing the same preprocessing, feature extraction, and graph transformation steps as in training. The classification module predicts the fracture type, the localization module provides a bounding box for the affected bone region, and the NLP module generates an automated radiology report, summarizing the diagnosis with details on fracture severity, affected area, and recommended clinical actions. To ensure interpretability, explainable AI techniques such as Grad-CAM heatmaps and graph attention visualizations highlight the most influential regions in the model's decision-making process. The system's performance is quantitatively evaluated using classification accuracy, precision, recall, and F1-score, IoU and Dice scores for localization accuracy, and BLEU and ROUGE scores for the quality of generated reports. This Graph-Augmented Multi-Modal CNN Framework provides a clinically viable, scalable, and interpretable solution for automated fracture detection and reporting, significantly improving diagnostic efficiency, reducing radiologist workload, and ensuring standardized reporting across medical institutions. By integrating deep learning, spatial reasoning, and natural language processing, the system enhances the accuracy and reliability of medical diagnoses, making it a powerful tool for real-world clinical applications.

### D. Cross-Validation and Performance Evaluation

To evaluate the robustness and generalizability of our fracture classification and localization model, we performed a 5-fold cross-validation on our native collected dataset alone because of non-availability of specified type of public datasets. In this evaluation setup, the dataset consisting of 4,410 images was randomly divided into 5 equal subsets, where each subset contains 882 images. For each fold, the model was trained on 4 subsets (80%

of the data, i.e., 3,528 images) and tested on the remaining subset (20% of the data, i.e., 882 images). This process was repeated for 5 times, each time with a different subset used for testing, ensuring that every sample in the dataset was used for both training and testing. This cross-validation approach allows us to obtain a more reliable estimate of the model's performance by reducing the risk of overfitting and providing a more generalized performance evaluation across the entire dataset. Additionally, it ensures that all data points are considered during the training and testing phases, maximizing the model's exposure to various fracture types and patient conditions.

The primary goal of our model was to classify various types of bone fractures based on X-ray and CT scan images. To assess the model's ability to accurately classify bone fractures, we evaluate its performance using key classification metrics including sensitivity, specificity, precision, accuracy, and f1-score. These metrics ensure a balanced evaluation of both positive and negative predictions, minimizing misclassifications. Sensitivity measures how well the model detects the fracture types. Specificity evaluates the model's ability to identify non-fracture cases. Precision assesses the correctness of positive predictions. The overall accuracy of the model across all folds was calculated as the ratio of correct predictions to the total number of predictions. This gives a general overview of the model's ability to correctly classify bone fractures. F1-score balances precision and recall. Beyond classification, precise fracture localization is crucial for effective medical diagnosis.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (31)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (32)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (33)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (34)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (35)$$

To evaluate the proposed model's efficacy in identifying the exact fracture region, segmentation performance is assessed using the following metrics: Intersection over Union (IoU), Dice Similarity Coefficient (DSC), Mean Absolute Error (MAE) and Mean Squared Error (MSE). IoU measures the ratio of the intersection of the predicted segmentation mask and the ground truth mask to their union. DSC is a measure of overlap between the predicted and ground truth binary masks, where 1 indicates perfect overlap and 0 indicates no overlap. Similar to IoU but DSC gives more weight to overlapping regions. MAE measures pixel-wise deviation from ground truth. MSE was calculated between the predicted and ground truth masks to measure how far the predicted segmentation is from the true region. MSE penalizes larger errors.

$$IoU = 2 \times \frac{|A \cap B|}{|A \cup B|} \quad (36)$$

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (37)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (38)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (39)$$

The quality of automated radiology report generation is evaluated using NLP-based metrics: Consensus-based Image Description Evaluation (CIDEr), Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) and Bilingual Evaluation Understudy (BLEU) scores. CIDEr measures similarity to expert-written reports. ROUGE-L evaluates structural and content overlap. BLEU Scores assesses n-gram overlap with reference reports.

$$CIDEr = \frac{1}{N} \sum_{i=1}^N \frac{g_i \cdot r_i}{||g_i|| ||r_i||} \quad (40)$$





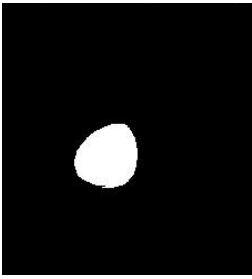
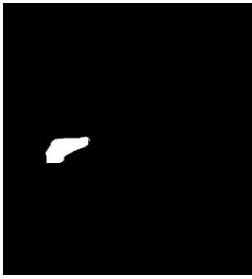

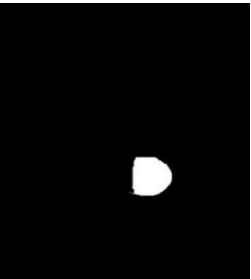






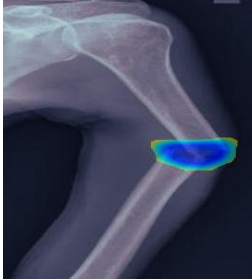

$$ROUGE - L = \frac{LCS_{generated, reference}}{length \ of \ reference \ report} \quad (41)$$

$$BLEU - N = BP \times \exp(\sum_{i=1}^N w_n \log p_n) \quad (42)$$

#### IV. RESULTS AND DISCUSSION

The proposed work is compared with the state-of-the-art method Parvin, S., et. al., 2024, Zou, J., et. al., 2024 and Lu, S., et. al., 2022. Parvin, S., et. al., 2024 introduced a real-time bone fracture detection system utilizing the YOLOv8 deep learning model to analyze multi-modal images. They developed the Human Bone Fractures Multi-modal Image Dataset (HBFMID), comprising 641 images across ten fracture classes, and employed data augmentation to mitigate overfitting. This system achieved impressive results, with 95% precision, 93% recall, and a 92% mean average precision, demonstrating its efficacy in accurately identifying and classifying various bone fractures.

Table 8: Fracture Localization Diagnostic Report

Input Image				
Localization Mask				
Localization Output				
Grad_CAM				
Generated Report	The X-ray image reveals a comminuted fracture in the mid-shaft region of the tibia, characterized by multiple bone fragments. The fracture site is well-defined with noticeable displacement, indicating a significant break.	The X-ray image reveals a displaced fracture in the distal radius of the left forearm. The fracture appears to be an oblique break with misalignment of the bone fragments, indicative of a severe impact or fall-related trauma.	The X-ray image reveals a comminuted fracture in the distal femur, characterized by multiple bone fragments at the fracture site. The fracture is localized in the knee region, potentially affecting joint stability and mobility.	The X-ray image reveals a transverse fracture located in the midshaft region of the tibia, characterized by a clean break across the bone. The right lower leg is affected, with the fracture appearing non-displaced, meaning the bone fragments remain aligned.

Zou, J., et. al., 2024 presented an enhanced YOLOv7 model, YOLOv7-ATT, designed for the detection of various bone fracture types, including angle fractures, normal fractures, line fractures, and complex angle fractures. By integrating an attention mechanism and employing the Enhanced Intersection over Union (EIou) loss function, their model achieved a mean Average Precision (mAP) of 80.2% on standard datasets and 86.2% on the FracAtlas

dataset, outperforming other models in accuracy and generalization. Lu, S., et. al., 2022 employed Ada-ResNeSt with AC-BiFPN to detect fractures across multiple anatomical regions in X-ray images. Their approach enhanced feature extraction and fusion, improving detection accuracy. This study achieved an Average Precision (AP) of 68.4% but faced challenges with real-time performance due to a 122 ms inference speed. Fracture localization and diagnostic reports generated by Graph-Augmented Multi-Modal CNN Framework are shown in Table 8.

Table 9: Classification Performance Metrics of the Graph-Augmented Multi-Modal CNN Framework

Folds	Sensitivity	Specificity	Precision	Accuracy	F1 score
Fold 1	0.891	0.959	0.978	0.982	0.893
Fold 2	0.945	0.888	0.938	0.956	0.986
Fold 3	0.913	0.961	0.958	0.975	0.947
Fold 4	0.971	0.919	0.875	0.899	0.929
Fold 5	0.933	0.947	0.946	0.963	0.912
<b>Overall</b>	<b>0.931</b>	<b>0.935</b>	<b>0.939</b>	<b>0.955</b>	<b>0.933</b>

The evaluation of the Graph-Augmented Multi-Modal CNN Framework highlights its effectiveness in bone fracture classification, localization and automated report generation. The model consistently achieves high classification accuracy, precise segmentation and clinically relevant report generation, making it a promising solution for AI-assisted radiology workflows. The model demonstrates strong classification capabilities achieving an average accuracy of 95.5% across five cross-validation folds. The 93.9% precision confirms high reliability in fracture detection, while the 93.5% specificity ensures effective identification of non-fracture cases, reducing false alarms. Additionally, the 93.3% F1-score highlights the model's balanced performance, even in imbalanced datasets. Performance varied slightly across different fracture types, with Simple and Comminuted Fractures achieving F1-scores above 0.90, indicating excellent classification accuracy. However, Pathological and Segmental fractures were more challenging to classify due to their underrepresentation in the dataset. The confusion matrix analysis, as presented in Table 9, showed that while the model effectively differentiated between common fracture types, it struggled with rare classes, suggesting the need for data augmentation, class weighting, or oversampling techniques to enhance classification performance for low-frequency fractures. The confusion matrix analysis showed that while the model effectively differentiated between common fracture types, it struggled with rare classes, suggesting the need for data augmentation, class weighting, or oversampling techniques to enhance classification performance for low-frequency fractures. The convergence of model during the 100 epochs has figured in Figure 3 and 4, which demonstrated the 5 fold's trend lines for understanding the capability of proposed method in the fracture classification phase. Figure 2 Curve compares different models for fracture detection, with the proposed model achieving the highest AUC (0.96), outperforming existing methods by Parvin, S., et. al., (0.89), Zou, J., et. al., (0.85), and Lu. S., et. al., (0.81). The proposed model maintains consistently higher precision across recall values, indicating improved reliability in detecting fractures while minimizing false positives. The gap between the curves suggests that prior methods struggle with precision at higher recall levels, whereas the proposed model achieves a more balanced tradeoff. This enhancement can lead to more accurate and early fracture detection, improving clinical decision-making and patient outcomes.

Table 10: Segmentation Performance Metrics for Fracture Localization

Folds		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
Proposed Work	IoU	0.957	0.965	0.97	0.953	0.962	0.961
	Dice coefficient	0.952	0.949	0.958	0.963	0.953	0.955
	MAE	0.012	0.015	0.013	0.018	0.021	0.016
	MSE	0.021	0.019	0.013	0.017	0.019	0.018
Parvin S, et. al, 2024	IoU	0.945	0.953	0.963	0.949	0.925	0.947
	Dice coefficient	0.948	0.955	0.949	0.934	0.944	0.946
	MAE	0.007	0.033	0.015	0.058	0.045	0.032
	MSE	0.016	0.024	0.018	0.021	0.019	0.02



Zou, J., et. al., 2024	IoU	0.934	0.936	0.927	0.932	0.945	0.935
	Dice coefficient	0.928	0.924	0.922	0.938	0.931	0.929
	MAE	0.046	0.033	0.042	0.039	0.0312	0.038
	MSE	0.037	0.043	0.038	0.029	0.039	0.037
Lu, S., et. al., 2022	IoU	0.923	0.931	0.928	0.935	0.926	0.929
	Dice coefficient	0.928	0.925	0.923	0.911	0.917	0.921
	MAE	0.032	0.041	0.044	0.038	0.045	0.04
	MSE	0.061	0.035	0.083	0.049	0.024	0.05

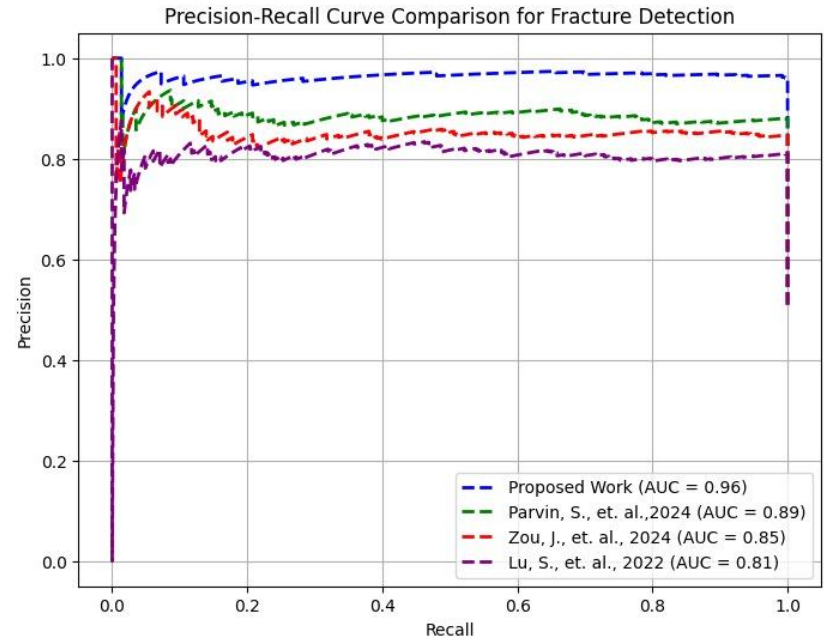


Fig. 2: Precision-Recall Curve Comparison for Fracture Detection Model

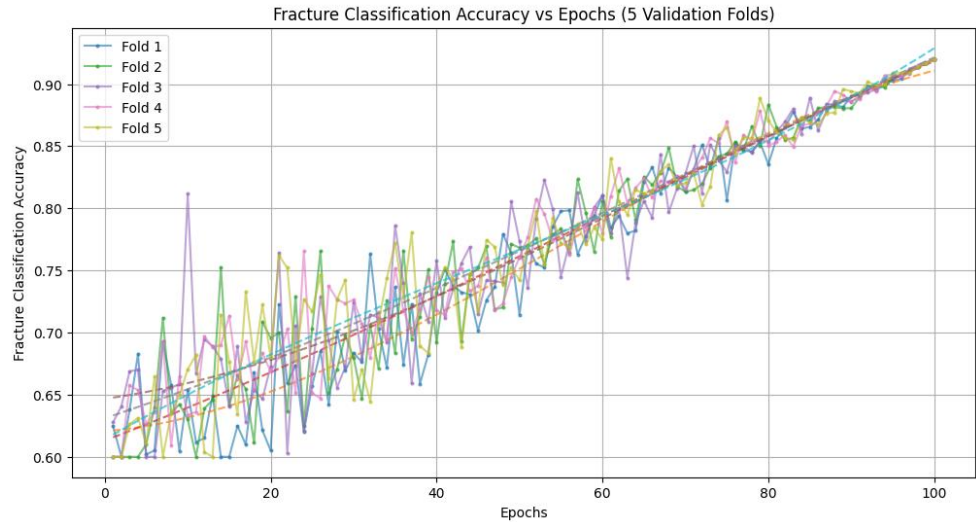


Fig. 3: Convergence of the model over the 100 epochs vs fracture classification accuracy with the 5 validation folds

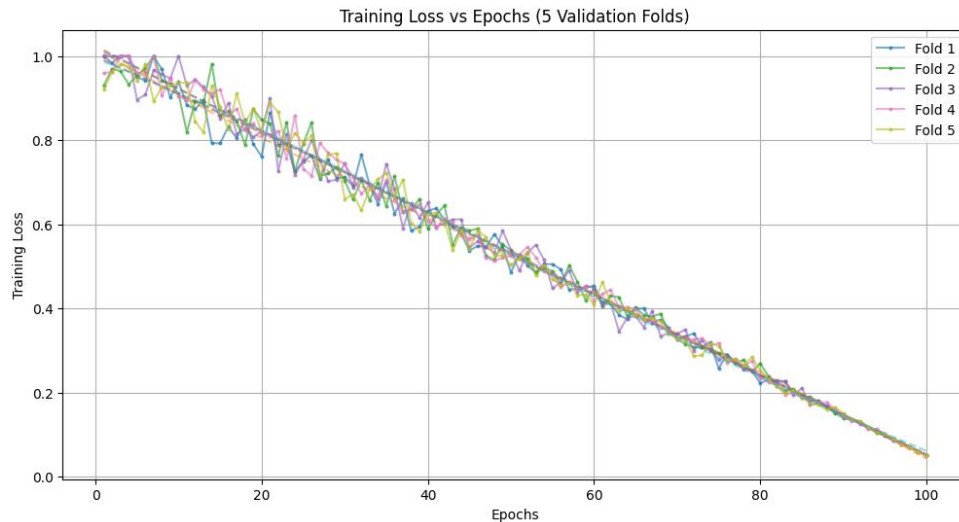


Fig. 4: Convergence of the model over the 100 epochs vs Training loss with the 5 validation folds

Accurate fracture localization is essential for clinical decision-making and the model exhibits exceptional segmentation performance, detailed in Table 10. Table 10 shows the comparative results of segmentation performance metrics for fracture localization with some other comparative works (Parvin, S., et. al., 2024, Zou, J., et. al., 2024, and Lu. S., et. al., 2022). The 96.1% IoU score confirms a high degree of overlap between predicted and ground-truth masks, while the 95.5% DSC further validates precise segmentation. The model achieved low segmentation errors, with an MAE of 0.016 and MSE of 0.018, indicating that predicted fracture regions were closely aligned with expert-annotated ground truth masks. While the model performed well for common fracture types, Segmental and Pathological fractures exhibited slightly lower Dice scores and higher MSE likely due to limited training samples. Addressing this imbalance through enhanced dataset diversity and targeted augmentation strategies could improve segmentation accuracy for these complex and rare fracture types.

Table 11: Automated Radiology Report Generation Performance Metrics

Folds	CIDEr	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Fold 1	0.775	0.850	0.793	0.885	0.865	0.856
Fold 2	0.819	0.873	0.795	0.884	0.792	0.851
Fold 3	0.795	0.865	0.812	0.895	0.801	0.775
Fold 4	0.885	0.828	0.785	0.863	0.795	0.873
Fold 5	0.873	0.885	0.789	0.865	0.827	0.885
<b>Overall</b>	<b>0.829</b>	<b>0.860</b>	<b>0.795</b>	<b>0.878</b>	<b>0.816</b>	<b>0.848</b>

The NLP-based report generation module produces structured diagnostic reports that align closely with radiologist-written reports, as demonstrated in Table 11. The CIDEr score of 0.829 indicates strong clinical relevance, ensuring that the generated reports contain accurate and meaningful diagnostic details. The ROUGE-L score of 0.860 validates structural coherence, ensuring that the reports are well-organized and readable. Furthermore, the BLEU-1 score of 0.795 confirms word-level accuracy, while the BLEU-4 score of 0.848 ensures that the reports maintain contextual fluency and coherence. These results demonstrate that the model-generated reports are lexically and contextually accurate, making them highly suitable for clinical applications.

The Graph-Augmented Multi-Modal CNN Framework achieves high classification accuracy, precise fracture localization, and structured diagnostic reporting making it a clinically viable solution for AI-driven radiology. The model performs exceptionally well in detecting and localizing common fracture types but future work should focus on improving the classification and segmentation of rare fracture types. To further enhance performance, strategies such as class balancing, additional data augmentation and transfer learning can be explored. With these improvements, the proposed framework has the potential to significantly reduce radiologist workload, enhance diagnostic accuracy, and streamline medical workflows, reinforcing AI's role in modern healthcare.

## V. CONCLUSION

The 5-fold cross-validation results demonstrated that the proposed model is highly effective for both fracture classification and localization, with consistent performance across multiple evaluation metrics. The classification accuracy and segmentation overlap show promising potential for real-world applications in automated bone fracture detection and reporting. While the model shows strong performance in detecting and classifying common fracture types, there is room for improvement in detecting rarer fracture types, such as Pathological and Segmental fractures. Future work can focus on enhancing performance for these underrepresented classes, possibly through techniques like data augmentation, class balancing, or transfer learning. Overall, the findings highlight the model's capability to assist radiologists in both classifying fractures and localizing fracture regions, contributing to a more efficient and automated radiology workflow.

## ACKNOWLEDGEMENTS

The authors Acknowledge the tremendous supports provided by the radiologist and diagnostic centres and distinguished professors in this field to carry out this research.

## REFERENCES

- [1] Parvin, S. and Rahman, A., 2024. A real-time human bone fracture detection and classification from multi-modal images using deep learning technique. *Applied Intelligence*, 54(19), pp.9269-9285.
- [2] Windarto, A.P. and Alkhairi, P., 2024. Bone fracture classification using convolutional neural network architecture for high-accuracy image classification. *International Journal of Electrical & Computer Engineering* (2088-8708), 14(6).
- [3] Mittal, K., Gill, K.S., Aggarwal, P., Rawat, R.S. and Sunil, G., 2024, June. Revolutionizing Fracture Diagnosis: A Deep Learning Approach for Bone Fracture Detection and Classification. In *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0* (pp. 1-5). IEEE.
- [4] Alshahrani, A. and Alsairafi, A., 2024. Bone Fracture Classification using Convolutional Neural Networks from X-ray Images. *Engineering, Technology & Applied Science Research*, 14(5), pp.16640-16645.
- [5] M Fariz Fadillah, M., Elly, P., Fatiha Nadia, S. and Alfi Nur, N., 2024. Convolutional Neural Network Model for Bone Fracture Detection and Classification in X-Ray Images. *Journal of Data Science*, 2024(43), pp.1-6.
- [6] Chauhan, S., 2024, September. Bone Fracture Detection with CNN: A Deep Learning Approach. In *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1253-1258). IEEE.
- [7] Ali, S.N.E., Sherif, H.M., Hassan, S.M. and El Marakby, A.A.E.R., 2024. Long bones x-ray fracture classification using machine learning. *Journal of Al-Azhar University Engineering Sector*, 19(72), pp.121-133.
- [8] Zou, J. and Arshad, M.R., 2024. Detection of whole body bone fractures based on improved yolov7. *Biomedical Signal Processing and Control*, 91, p.105995.
- [9] Bittner-Frank, M., Strassl, A., Unger, E., Hirtler, L., Eckhart, B., Koenigshofer, M., Stoenner, A., Nia, A., Popp, D., Kainberger, F. and Windhager, R., 2024. Accuracy analysis of 3D bone fracture models: effects of computed tomography (CT) imaging and image segmentation. *Journal of Imaging Informatics in Medicine*, pp.1-1
- [10] Murrad, B.G., Mohsin, A.N., Al-Obaidi, R.H., Albaaji, G.F., Ali, A.A., Hamzah, M.S., Abdulridha, R.N. and Al-Sharifi, H.K., 2024. An AI-Driven Framework for Detecting Bone Fractures in Orthopedic Therapy. *ACS Biomaterials Science & Engineering*.
- [11] Potter, İ.Y., Rodriguez, E.K., Wu, J., Nazarian, A. and Vaziri, A., 2024. An automated vertebrae localization, segmentation, and osteoporotic compression fracture detection pipeline for computed tomographic imaging. *Journal of Imaging Informatics in Medicine*, 37(5), p.2428.
- [12] Dibo, R., Galichin, A., Astashev, P., Dylov, D.V. and Rogov, O.Y., 2023, September. DeepLOC: Deep Learning-based Bone Pathology Localization and Classification in Wrist X-ray Images. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 199-211). Cham: Springer Nature Switzerland.
- [13] Ju, R.Y. and Cai, W., 2023. Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. *Scientific Reports*, 13(1), p.20077.
- [14] Beyraghi, S., Ghorbani, F., Shabanpour, J., Lajevardi, M.E., Nayyeri, V., Chen, P.Y. and Ramahi, O.M., 2023. Microwave bone fracture diagnosis using deep neural network. *Scientific Reports*, 13(1), p.16957.

- [15] Khan, A.A., Slart, R.H., Ali, D.S., Bock, O., Carey, J.J., Camacho, P., Engelke, K., Erba, P.A., Harvey, N.C., Lems, W.F. and Morgan, S., 2024, July. Osteoporotic fractures: diagnosis, evaluation, and significance from the International Working Group on DXA Best Practices. In Mayo clinic proceedings (Vol. 99, No. 7, pp. 1127-1141). Elsevier.
- [16] Singh, A., 2024, October. BoneScanAI: Advanced Machine Learning for Precision Bone Fracture Diagnosis. In 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA) (pp. 1-6). IEEE.
- [17] Kumar, G., Patidar, V., Biswas, P., Patel, M., Rajput, C.S., Venugopal, A. and Sharma, A., 2023. IOT enabled Intelligent featured imaging Bone Fractured Detection System. Journal of Intelligent Systems and Internet of Things, 9(2), pp.08-22.
- [18] Su, Z., Zhou, Y., Zhou, J., Cao, H. and Zhang, H., 2024. BoneCLIP-XGBoost: A Multimodal Approach for Bone Fracture Diagnosis. IEEE Access.
- [19] Pérez-Cano, F.D., Parra-Cabrera, G., Camacho-García, R. and Jiménez, J.J., 2024. Enhancing Medical Diagnosis and Treatment Planning through Automated Acquisition and Classification of Bone Fracture Patterns.
- [20] Zeng, B., Wang, H., Xu, J., Tu, P., Joskowicz, L. and Chen, X., 2023. Two-stage structure-focused contrastive learning for automatic identification and localization of complex pelvic fractures. IEEE Transactions on Medical Imaging, 42(9), pp.2751-2762.
- [21] Yu, Q., Liu, Y., Li, H., Liu, X., Bao, X., Jin, W., Xia, W., Tang, Z., Tang, P., Chen, H. and Wang, X., 2025. Multi-task learning for calcaneus fracture diagnosis of X-ray images. Biomedical Signal Processing and Control, 99, p.106843.
- [22] Linda, C.H. and Jiji, G.W., 2011. Crack detection in X-ray images using fuzzy index measure. Applied Soft Computing, 11(4), pp.3571-3579.
- [23] Lu, S., Wang, S. and Wang, G., 2022. Automated universal fractures detection in X-ray images based on deep learning approach. Multimedia Tools and Applications, 81(30), pp.44487-44503.