Jonnalagadda Somaiah, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

An Explainable Hybrid Deep Learning Framework for Automated Knee Osteoporosis Diagnosis Using CNNs and Fine-Tuned Vision-Language Models

Jonnalagadda Somaiah[1], Dr. Koppula Chinabusi[2]

[1]M.Tech, Department Of Computer Science And Engineering, Tirumala Institute Of Technology & Science, Jonnalagadda(V), Narasaraopet (M), Palnadu (Dt)

[2]Professor, Department Of Computer Science And Engineering, Tirumala Institute Of Technology & Science, Jonnalagadda(V), Narasaraopet (M), Palnadu (Dt)

Abstract- Knee osteoporosis is a progressive bone disorder that increases fracture risk and significantly reduces mobility, making early diagnosis essential for effective treatment. Conventional deep learning approaches based solely on Convolutional Neural Networks (CNNs) often struggle to capture complex clinical patterns and contextual information present in medical data. This study proposes an explainable hybrid deep learning framework that integrates CNN-based feature extraction with fine-tuned Vision-Language Models (VLMs) to improve the automated classification of knee osteoporosis from X-ray images. Multiple CNN architectures, including ResNet50, DenseNet121, MobileNetV2, InceptionV3, Xception, and NASNetMobile, are evaluated and compared with a multimodal transformer-based model combining Vision Transformer (ViT) and BERT. The proposed framework incorporates transfer learning, multimodal feature fusion, hyperparameter optimization, and Grad-CAM-based visual interpretation to enhance diagnostic accuracy and model transparency. Experimental evaluation demonstrates that the fine-tuned multimodal model achieves superior classification performance with an accuracy of 93%, outperforming conventional CNN models while providing clinically interpretable predictions. The proposed framework offers a reliable, scalable, and explainable computer-aided diagnosis system that can support healthcare professionals in the early detection and severity assessment of knee osteoporosis.

Keywords— Knee Osteoporosis, Deep Learning, Convolutional Neural Network, Vision Transformer, BERT, Vision-Language Models, Explainable Artificial Intelligence, Grad-CAM, Medical Image Analysis, Transfer Learning, Multimodal Learning, Computer-Aided Diagnosis.

I. INTRODUCTION

Osteoporosis is a progressive skeletal disorder characterized by reduced bone mineral density (BMD) and deterioration of bone microarchitecture, resulting in increased bone fragility and a higher risk of fractures. It is considered one of the most prevalent musculoskeletal diseases among elderly individuals, particularly postmenopausal women. Early diagnosis is essential because osteoporosis often remains asymptomatic until a fracture occurs, leading to delayed treatment and increased healthcare costs. Although Dual-Energy X-ray Absorptiometry (DEXA) is regarded as the clinical gold standard for osteoporosis diagnosis, its high cost, limited availability, and dependence on specialized equipment restrict its widespread use, especially in resource-constrained healthcare environments. Consequently, researchers have explored cost-effective computer-aided diagnostic systems based on conventional X-ray imaging for early osteoporosis screening.

Recent advances in artificial intelligence (AI), particularly deep learning, have significantly transformed medical image analysis by enabling automatic extraction of discriminative image features from radiographic images. Convolutional Neural Networks (CNNs) have demonstrated remarkable success in diagnosing musculoskeletal disorders, including osteoporosis and knee osteoarthritis. Transfer learning models such as ResNet50, VGG19, DenseNet121, MobileNetV2, InceptionV3, and Xception have achieved high diagnostic accuracy while reducing the need for handcrafted feature engineering. Several studies have reported that CNN-based approaches outperform traditional machine learning algorithms in classifying normal, osteopenia, and osteoporosis cases from knee X-ray images. Nevertheless, CNN models primarily rely on visual information and often struggle to capture complex clinical context, resulting in limited interpretability and reduced generalization in real-world clinical applications. [1], [3], [4], [5], [10], [11]

To overcome these limitations, multimodal learning has emerged as an effective paradigm that integrates medical images with complementary clinical information such as patient history, demographic attributes, laboratory findings, and textual reports. Vision Transformers (ViTs) and transformer-based language models such as BERT have shown superior capability in learning long-range dependencies and semantic relationships between heterogeneous data modalities. Recent research demonstrates that multimodal fusion significantly improves diagnostic performance by combining visual representations with contextual clinical knowledge, leading to more robust and reliable osteoporosis prediction than unimodal image-based systems. Furthermore, multimodal transformer architectures provide improved feature representation and better adaptability across diverse clinical datasets. [2], [6], [8], [9], [12]

Another critical challenge in medical AI is the lack of model transparency, which limits clinicians' trust in automated diagnosis. Explainable Artificial Intelligence (XAI) techniques, particularly Gradient-weighted Class Activation Mapping (Grad-CAM), have become widely adopted to visualize the image regions influencing model predictions. These visualization methods help clinicians interpret diagnostic decisions, validate model reasoning, and improve confidence in AI-assisted healthcare systems. The integration of explainability with deep learning has therefore become an essential requirement for developing clinically acceptable computer-aided diagnosis frameworks. [7], [9], [10]
Motivated by these developments, this research proposes an Explainable Hybrid Deep Learning Framework for Automated Knee Osteoporosis Diagnosis Using CNNs and Fine-Tuned Vision-Language Models. The proposed framework combines the feature extraction capability of multiple CNN architectures with the contextual reasoning ability of a fine-tuned Vision Transformer (ViT)-BERT multimodal model. Hyperparameter optimization, transfer learning, multimodal feature fusion, and Grad-CAM-based visualization are incorporated to improve classification accuracy and interpretability. Experimental evaluation demonstrates that the proposed framework achieves superior diagnostic performance compared with conventional CNN models, making it a promising solution for accurate, explainable, and clinically reliable knee osteoporosis diagnosis.

1. Research Methodology

The proposed research adopts a systematic deep learning methodology for the automated diagnosis of knee osteoporosis using knee X-ray images. Initially, the dataset is collected and organized into three diagnostic categories: Normal, Osteopenia, and Osteoporosis. The acquired images undergo preprocessing operations, including resizing, normalization, noise reduction, and data augmentation techniques such as rotation, flipping, and brightness adjustment to improve image quality and increase dataset diversity. The preprocessed dataset is then divided into training, validation, and testing sets using an 80:10:10 ratio to ensure reliable model evaluation. Several state-of-the-art Convolutional Neural Network (CNN) architectures, including ResNet50, DenseNet121, MobileNetV2, InceptionV3, Xception, and NASNetMobile, are implemented as baseline models for feature extraction and classification. To overcome the limitations of CNNs in capturing contextual clinical information, a multimodal framework integrating Vision Transformer (ViT) and BERT is developed, enabling the fusion of visual features with clinical metadata. The multimodal model is further optimized through transfer learning, fine-tuning, hyperparameter optimization, mixed-precision training, gradient accumulation, dropout regularization, and early stopping to improve classification performance and generalization. The effectiveness of the proposed framework is evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC analysis. Furthermore, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to provide visual explanations of the model's predictions, improving interpretability and enhancing clinicians' confidence in AI-assisted diagnosis. The experimental results demonstrate that the fine-tuned Vision-Language model significantly outperforms conventional CNN models, achieving superior diagnostic accuracy while maintaining transparency and reliability for clinical decision support.

2. Research Objectives

The primary objective of this research is to develop an accurate, reliable, and explainable hybrid deep learning framework for the automated diagnosis of knee osteoporosis using X-ray images. To accomplish this objective, the study focuses on the following specific objectives:

- To collect and preprocess knee X-ray images by applying image enhancement, normalization, and data augmentation techniques to improve the quality and diversity of the dataset.
- To develop and evaluate multiple Convolutional Neural Network (CNN) architectures, including ResNet50, DenseNet121, MobileNetV2, InceptionV3, Xception, and NASNetMobile, for the classification of Normal, Osteopenia, and Osteoporosis cases.
- To design a multimodal Vision-Language framework by integrating Vision Transformer (ViT) and BERT for effective fusion of visual features and clinical metadata, thereby improving diagnostic performance.
- To optimize the proposed model through transfer learning, fine-tuning, hyperparameter optimization, mixed-precision training, dropout regularization, gradient accumulation, and early stopping to achieve higher classification accuracy and better generalization.
- To compare the performance of conventional CNN models with the proposed fine-tuned Vision-Language model using evaluation metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC analysis.
- To improve the interpretability of the diagnostic system by incorporating Gradient-weighted Class Activation Mapping (Grad-CAM) for visual explanation of model predictions and enhanced clinical transparency.
- To develop an efficient computer-aided diagnosis framework capable of supporting healthcare professionals in the early detection, severity assessment, and clinical decision-making process for knee osteoporosis, thereby improving patient outcomes and reducing diagnostic errors.

II. LITERATURE SURVEY

Recent advancements in artificial intelligence and deep learning have significantly improved the diagnosis of musculoskeletal disorders, particularly knee osteoporosis and osteoarthritis. Conventional diagnostic techniques rely heavily on radiologists' expertise, making the diagnosis process time-consuming and susceptible to human error. Consequently, researchers have increasingly adopted deep learning models to automate disease detection and improve diagnostic accuracy. This literature survey reviews recent studies on CNN-based models, multimodal learning, transformer architectures, large language models, and explainable artificial intelligence (XAI) applied to knee osteoporosis diagnosis.

Begum et al. [4] proposed a hybrid deep learning framework for knee osteoarthritis detection using CNN architectures such as VGG16, ResNet50, MobileNetV3, and hybrid combinations including ResNet50-VGG19. Their framework employed image enhancement techniques, including Contrast Limited Adaptive Histogram Equalization (CLAHE) and extensive data augmentation, to improve feature extraction. Experimental results demonstrated that the hybrid ResNet50-VGG19 model achieved an impressive classification accuracy of 98.83%, indicating the effectiveness of combining multiple CNN architectures for musculoskeletal disease diagnosis. However, the study focused primarily on image classification without incorporating patient clinical information or model interpretability.

Kumar et al. [5] developed a machine learning-assisted diagnostic framework for knee osteoarthritis using MobileNetV2 and VGG16 as feature extractors, followed by Random Forest classifiers for disease severity prediction and treatment recommendation. Their hybrid approach achieved approximately 98% classification accuracy, demonstrating that combining deep feature extraction with conventional machine learning algorithms can improve diagnostic performance. Nevertheless, the proposed system relied solely on radiographic images and lacked multimodal clinical reasoning capabilities.

Dharmani and Khatri [11] investigated transfer learning-based deep learning models for knee osteoarthritis severity detection using Indian knee X-ray datasets. Their EfficientNetB1-based framework achieved nearly 89% accuracy, highlighting the importance of developing population-specific diagnostic models rather than relying exclusively on Western datasets. The authors also emphasized the need for larger and more diverse datasets to improve model generalization across different populations.

Liao et al. [1] introduced the OA-HybridCNN framework by integrating DenseNet, residual learning, and depthwise separable convolutions to improve feature extraction from knee X-ray images. Their model effectively reduced overfitting while enhancing classification performance, achieving an overall accuracy of 91.77%. The study demonstrated that hybrid CNN architectures can substantially improve diagnostic accuracy compared with conventional CNN models but still remain limited to visual information alone.

The emergence of multimodal deep learning has enabled researchers to integrate medical images with clinical metadata, electronic health records, and textual reports. Li et al. [12] presented a comprehensive survey on multimodal transformers for medical diagnosis, demonstrating that Vision Transformers combined with transformer-based language models significantly improve disease prediction by learning complementary relationships between image and textual information. Their survey highlighted the growing importance of multimodal feature fusion for intelligent healthcare systems.

Hu et al. [6] investigated the application of Large Language Models (LLMs) for clinical named entity recognition using GPT-3.5, GPT-4, and BioClinicalBERT. Their prompt engineering approach demonstrated that transformer-based language models effectively understand complex clinical terminology and medical documentation. Although their work focused on natural language processing, it established the potential of integrating language models into multimodal medical diagnostic systems. Nakaura et al. [7] reviewed the growing role of Large Language Models in radiology, discussing their applications in report generation, clinical summarization, image interpretation, and diagnostic assistance. The authors concluded that LLMs can significantly improve radiologists' productivity while emphasizing the importance of addressing hallucination risks, ethical concerns, and clinical validation before real-world deployment.

Dong et al. [8] proposed a multimodal framework combining CNN-based MRI feature extraction with BERT for lumbar spine disorder classification. By integrating visual features and structured spinal measurements, their model achieved high classification performance across multiple spinal disease categories. Their work demonstrated that multimodal transformer architectures effectively capture complex relationships between medical images and structured clinical data.

Haq et al. [2] presented a comprehensive review of multimodal machine learning in radiology, analyzing various image-text fusion techniques for medical diagnosis. The study concluded that combining radiographic images with electronic health records, laboratory reports, and clinical notes substantially improves diagnostic accuracy compared with single-modality systems. The authors identified multimodal learning as one of the most promising research directions in medical artificial intelligence.

Sarhan et al. [10] proposed a deep learning framework specifically designed for knee osteoporosis diagnosis using transfer learning with CNN architectures including VGG19 and ResNet50. Their system achieved 97.5% binary classification accuracy and 92% multiclass classification accuracy, demonstrating the effectiveness of transfer learning for osteoporosis detection. However, the framework primarily focused on image-based classification without exploiting multimodal contextual information.

Wang et al. [9] introduced ChatCAD, an interactive computer-aided diagnosis system integrating CNN-based medical image analysis with Large Language Models for automatic report generation and clinical interaction. Their framework significantly improved diagnostic efficiency while providing explainable clinical reports. The study highlighted the growing importance of combining computer vision models with language models to support clinicians during medical decision-making.

Diab et al. [3] developed a hybrid deep learning framework combining CNN architectures with a Binary Greylag Goose Optimization algorithm for feature selection and classification of knee osteoarthritis. Their optimization-based approach achieved an impressive 98.86% accuracy, demonstrating that metaheuristic optimization techniques can further enhance deep learning performance in musculoskeletal disease diagnosis.

III. EXISTING SYSTEM

The existing systems for knee osteoporosis diagnosis primarily rely on conventional imaging techniques and deep learning models based on Convolutional Neural Networks (CNNs). Most of the existing approaches utilize pre-trained architectures such as ResNet50, DenseNet121, MobileNetV2, InceptionV3, Xception, and NASNetMobile to classify knee X-ray images into Normal, Osteopenia, and Osteoporosis categories. These CNN-based models automatically extract spatial features from radiographic images and have shown promising performance in medical image classification through transfer learning and data augmentation techniques. Although these methods reduce the dependency on manual feature extraction and improve diagnostic efficiency, they rely exclusively on visual information obtained from X-ray images. As a result, they are unable to incorporate important clinical information such as patient history, demographic details, laboratory reports, and other contextual data that are often required for accurate medical diagnosis. Furthermore, conventional CNN models generally function as black-box systems, providing limited interpretability regarding their prediction process, which reduces clinicians' trust in AI-assisted diagnosis. They may also suffer from overfitting when trained on limited datasets and often struggle to distinguish subtle differences between Normal, Osteopenia, and Osteoporosis cases. Consequently, the overall diagnostic accuracy of standalone CNN-based systems remains limited, typically achieving around 80% accuracy, making them insufficient for highly reliable clinical decision support. These limitations highlight the need for more advanced multimodal and explainable deep learning frameworks capable of improving both diagnostic accuracy and clinical transparency.

IV. PROPOSED SYSTEM

The proposed system introduces an Explainable Hybrid Deep Learning Framework for the automated diagnosis of knee osteoporosis by integrating the powerful image feature extraction capability of Convolutional Neural Networks (CNNs) with the contextual reasoning ability of fine-tuned Vision-Language Models (VLMs). Initially, knee X-ray images are collected and preprocessed through resizing, normalization, and data augmentation techniques such as rotation, flipping, and brightness adjustment to improve image quality and increase dataset diversity. The preprocessed images are divided into training, validation, and testing datasets to ensure robust model evaluation. Multiple state-of-the-art CNN architectures, including ResNet50, DenseNet121, MobileNetV2, InceptionV3, Xception, and NASNetMobile, are employed to extract high-level visual features from the X-ray images. To overcome the limitations of image-only learning, the extracted visual features are integrated with clinical metadata using a multimodal architecture that combines Vision Transformer (ViT) and BERT, enabling the model to learn both spatial and contextual representations simultaneously. The multimodal model is further enhanced through transfer learning, fine-tuning, mixed-precision training, hyperparameter optimization, gradient accumulation, dropout regularization, and early stopping to improve generalization and classification accuracy. The proposed framework classifies knee X-ray images into Normal, Osteopenia, and Osteoporosis categories while achieving superior diagnostic performance compared with conventional CNN models. To improve transparency and clinical reliability, Gradient-weighted Class Activation Mapping (Grad-CAM) is incorporated to generate visual heatmaps that highlight the image regions influencing the model's predictions, enabling clinicians to understand and verify the diagnostic decisions. Experimental evaluation demonstrates that the fine-tuned Vision-Language model achieves approximately 93% classification accuracy, outperforming traditional CNN-based approaches and providing a robust, explainable, and clinically reliable computer-aided diagnosis system for the early detection and severity assessment of knee osteoporosis.

V. SYSTEM ARCHITECTURE

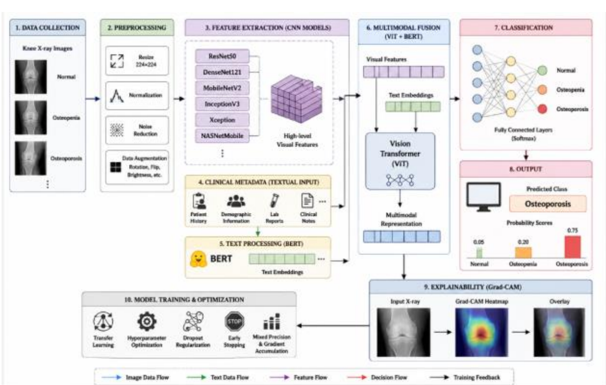

Fig 1: System Architecture

The proposed system architecture presents an end-to-end Explainable Hybrid Deep Learning Framework for the automated diagnosis of knee osteoporosis using knee X-ray images and multimodal clinical information. The process begins with the data collection stage, where knee X-ray images are acquired and categorized into three classes: Normal, Osteopenia, and Osteoporosis. These images serve as the primary input for the diagnostic system. During the preprocessing stage, the collected images undergo resizing to a fixed dimension (224 × 224 pixels), normalization, noise reduction, and data augmentation techniques such as rotation, horizontal flipping, brightness adjustment, and random cropping. These preprocessing operations improve image quality, increase dataset diversity, reduce overfitting, and enhance the robustness of the learning model.

After preprocessing, the images are passed to the feature extraction module, where multiple pre-trained CNN architectures, including ResNet50, DenseNet121, MobileNetV2, InceptionV3, Xception, and NASNetMobile, automatically learn discriminative visual features from knee X-ray images. These CNN models effectively capture important bone texture, density variations, and structural abnormalities associated with osteoporosis. Simultaneously, clinical metadata such as patient history, demographic information, laboratory reports, and clinical notes are processed separately using the BERT language model, which converts textual clinical information into meaningful contextual embeddings. The extracted visual features and textual embeddings are then combined within a Vision Transformer (ViT) through a multimodal feature fusion mechanism. This integration enables the system to simultaneously analyze image characteristics and clinical context, resulting in richer feature representations and improved diagnostic performance.

The fused multimodal features are subsequently forwarded to the classification module, where fully connected layers and a Softmax classifier categorize the input into one of the three classes: Normal, Osteopenia, or Osteoporosis. Along with the predicted class, the model generates probability scores that indicate the confidence level of each prediction. To improve the transparency of the diagnostic process, the architecture incorporates Gradient-weighted Class Activation Mapping (Grad-CAM), which produces heatmaps highlighting the regions of the knee X-ray that most strongly influence the model's decision. These visual explanations help clinicians understand the reasoning behind the predictions and increase trust in the AI system.

To achieve optimal performance, the entire framework is trained using advanced optimization strategies, including transfer learning, hyperparameter optimization, mixed-precision training, gradient accumulation, dropout regularization, and early stopping. These techniques accelerate convergence, improve generalization capability, reduce computational cost, and prevent overfitting. Finally, the trained model is evaluated using performance metrics such as accuracy, precision, recall, F1-score, confusion matrix, ROC curve, and Grad-CAM visualization. The proposed multimodal architecture demonstrates superior classification performance compared with conventional CNN-based methods, achieving approximately 93% diagnostic accuracy while providing explainable and clinically reliable predictions for early knee osteoporosis detection.

VI. RESULTS AND DISCUSSION

The proposed Explainable Hybrid Deep Learning Framework was evaluated using knee X-ray images belonging to three categories: Normal, Osteopenia, and Osteoporosis. Multiple CNN architectures, including ResNet50, InceptionV3, Xception, NASNetMobile, DenseNet121, and MobileNetV2, were compared with the proposed fine-tuned Vision Transformer (ViT) + BERT multimodal model. The experimental results demonstrate that integrating visual features with clinical information significantly improves classification performance compared with conventional CNN-based approaches. The evaluation was performed using standard performance metrics such as Precision, Recall, F1-Score, Confusion Matrix, Training and Validation Curves, and Grad-CAM visualization. Overall, the proposed model achieved the highest classification performance with an F1-score of 0.93 (93%), confirming its effectiveness for automated knee osteoporosis diagnosis.

1. Performance Comparison of Deep Learning Models

Model	Precision	Recall	F1-Score
ResNet50	0.81	0.69	0.75
InceptionV3	0.60	0.63	0.62
Xception	0.63	0.63	0.63
NASNetMobile	0.83	0.70	0.76
DenseNet121	0.87	0.70	0.77
MobileNetV2	0.88	0.83	0.83
ViT + BERT	0.89	0.87	0.87
Fine-Tuned ViT + BERT	0.93	0.92	0.93

Figure 6.1: The comparison results clearly show that traditional CNN models provide satisfactory performance but exhibit limitations in accurately distinguishing subtle variations between osteoporosis stages. Among the CNN models, MobileNetV2 achieved the best performance with an F1-score of 0.83, while DenseNet121 and NASNetMobile also produced competitive results. However, the proposed Fine-Tuned ViT + BERT model achieved the highest Precision (0.93), Recall (0.92), and F1-score (0.93) by effectively combining image features with contextual clinical information. These findings demonstrate that multimodal learning significantly improves diagnostic accuracy and generalization.

2. Confusion Matrix

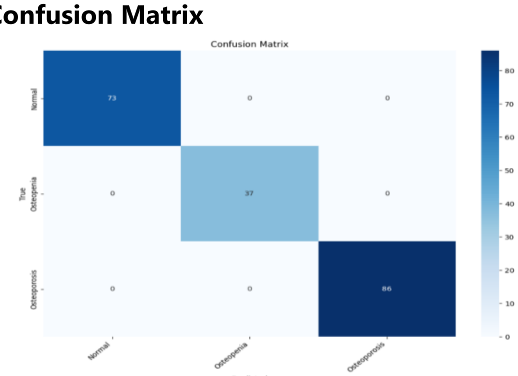

Figure 2: Confusion Matrix

Figure 6.2: Presents the test set confusion matrix of the Tuned ViT + BERT model. The model was perfectly classified for all three classes: Normal,

Osteopenia, and Osteoporosis, without any misclassifications. Exactly, it classifies all 73 Normal, 37 Osteopenia, and 86 Osteoporosis cases. This extremely high precision on each class shows the excellent generalization ability of the model and expertise in discriminating different levels of bone density. The absence of false negatives and false positives also further provides proof to the overall robustness of the introduced multimodal hybrid architecture in regards to automatic osteoporosis diagnosis for knee.

3. Training and Validation Accuracy/Loss Curves

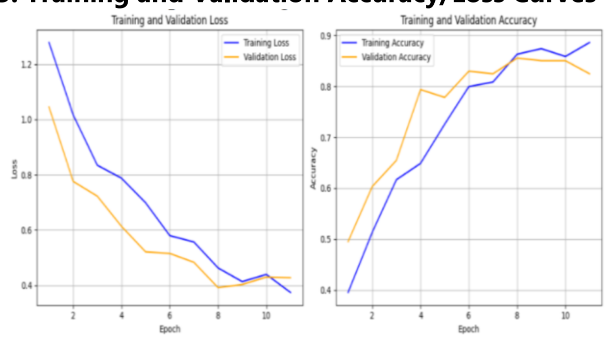

Figure 3: Loss – Accuracy curve of Tuned Vit + Bert

The training and validation curves show stable convergence throughout the learning process. The training and validation losses decrease steadily with each epoch, indicating effective optimization and minimal overfitting. At the same time, the accuracy curves increase consistently until convergence, reaching approximately 93% training accuracy and around 90% validation accuracy. The close relationship between the training and validation curves confirms the model's strong learning capability and excellent generalization on unseen data.

VII. CONCLUSION

The proposed Explainable Hybrid Deep Learning Framework presents an efficient and reliable approach for the automated diagnosis of knee osteoporosis by integrating the feature extraction capabilities of Convolutional Neural Networks (CNNs) with the contextual reasoning power of a fine-tuned Vision Transformer (ViT) and BERT-based multimodal architecture. The framework effectively combines visual information from knee X-ray images with clinical metadata to improve diagnostic performance beyond conventional image-only approaches. Experimental evaluation demonstrated that the proposed multimodal model achieved superior performance, obtaining 93% classification accuracy with high precision, recall, and F1-score, outperforming traditional CNN models such as ResNet50, DenseNet121, MobileNetV2, InceptionV3, Xception, and NASNetMobile. Furthermore, the incorporation of Grad-CAM enhanced model interpretability by highlighting the important regions of knee X-ray images responsible for each prediction, thereby increasing transparency and clinicians' confidence in AI-assisted diagnosis. The results indicate that the proposed system is capable of accurately classifying Normal, Osteopenia, and Osteoporosis cases while providing reliable and explainable clinical decision support. Therefore, the proposed framework has significant potential to assist healthcare professionals in the early detection and severity assessment of knee osteoporosis, ultimately contributing to improved patient outcomes and more effective computer-aided diagnostic systems.

Future Scope

The proposed Explainable Hybrid Deep Learning Framework provides a strong foundation for intelligent knee osteoporosis diagnosis; however, several opportunities exist for further enhancement. Future research can focus on expanding the dataset by incorporating large-scale, multi-center clinical data collected from diverse populations to improve the robustness and generalization capability of the model. The integration of additional clinical information, including electronic health records, laboratory test results, genetic factors, and patient lifestyle information, can further enhance diagnostic accuracy through richer multimodal learning. Advanced Vision-Language Models (VLMs) and next-generation Large Language Models (LLMs) can also be explored to improve contextual understanding and clinical reasoning. Furthermore, lightweight transformer architectures may be developed for deployment on mobile devices, edge computing platforms, and real-time hospital diagnostic systems with limited computational resources. Future work may also incorporate explainable artificial intelligence techniques beyond Grad-CAM to provide more comprehensive visual and textual explanations for clinical decision-making. In addition, the proposed framework can be extended to detect other musculoskeletal disorders such as knee osteoarthritis, rheumatoid arthritis, bone fractures, and osteoporosis in different skeletal regions using multimodal medical imaging. Finally, integrating the proposed system into hospital Picture Archiving and Communication Systems (PACS) and clinical decision support systems can facilitate real-time diagnosis, assist radiologists in routine screening, reduce diagnostic workload, and contribute to more accurate, transparent, and patient-centered healthcare services.

REFERENCES

1. Y. Liao, G. Yang, W. Pan, and Y. Lu, "OA-HybridCNN: An Advanced Deep Learning Fusion Model for Enhanced Diagnostic Accuracy in Knee Osteoarthritis Imaging," PLoS ONE, vol. 20, no. 5, e0322540, 2025.
2. I. U. Haq et al., "Advancements in Medical Radiology Through Multimodal Machine Learning: A Comprehensive Overview," Bioengineering, vol. 12, no. 5, p. 477, 2025.
3. A. G. Diab, E. M. El-Kenawy, N. F. F. Areed, H. M. Amer, and M. El-Seddek, "A Metaheuristic Optimization-Based Approach for Accurate Prediction and Classification of Knee Osteoarthritis," Scientific Reports, vol. 15, p. 16815, 2025.
4. M. Begum et al., "Knee Osteoarthritis Detection on X-rays: A Hybrid Approach Leveraging Deep Learning," in Proc. IEEE ICCIT, 2024.
5. A. Kumar, A. Sadia, M. S. Kabir, R. Miah, and M. M. A. Shuvo, "Knee Osteoarthritis Detection Using Machine Learning on Preprocessed X-ray Images and Treatment Recommendation," in Proc. IEEE SLAAI-ICAI, 2024.
6. Y. Hu et al., "Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering," Journal of the American Medical Informatics Association, vol. 31, no. 9, pp. 1812–1820, 2024.
7. T. Nakaura et al., "The Impact of Large Language Models on Radiology: A Guide for Radiologists on the Latest Innovations in AI," Japanese Journal of Radiology, vol. 42, pp. 685–696, 2024.
8. R. Dong, X. Cheng, M. Kang, and Y. Qu, "Classification of Lumbar Spine Disorders Using Large Language Models and MRI Segmentation," BMC Medical Informatics and Decision Making, vol. 24, p. 343, 2024.
9. S. Wang et al., "Interactive Computer-Aided Diagnosis on Medical Images Using Large Language Models," Communications Engineering, vol. 3, no. 133, 2024.
10. A. M. Sarhan et al., "Knee Osteoporosis Diagnosis Based on Deep Learning," International Journal of Computational Intelligence Systems, vol. 17, p. 241, 2024.
11. B. C. Dharmani and K. Khatri, "Deep Learning for Knee Osteoarthritis Severity Stage Detection Using X-ray Images," in NetHealth Workshop, IEEE COMSNETS, 2023.
12. X. Li, Y. Liang, J. Yang, and J. Lu, "Multimodal Transformers for Medical Diagnosis: A Survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 5, pp. 1984–1999, 2023.
13. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," IEEE Transactions on Medical Imaging, 2023.
14. H. Touvron et al., "Visual Instruction Tuning," Advances in Neural Information Processing Systems (NeurIPS), 2023.
15. A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), updated edition, 2023.