10.48047/jocaaa.2026.35.07.14

Bone Fracture Diagnosis with Two-Stream CNNs: Integrating Multimodal Data and Explainable AI for Enhanced Clinical Decision-Making

[1] Gujjula Swarnalatha, [2] Dr. Ranga Swamy Sirisati

[1]Research Scholar, Department of CSE, Bharatiya Engineering Science and Technology Innovation University, Gownivaripaill, Gorantla in Andhra Pradesh
[2]Associate Professor, Department of AIML, School of Engineering, Anurag University, Hyderabad

E-Mail: 2022wpcse022@bestiu.edu.in[1] sirisatiranga@gmail.com[2]

Abstract— Accurate and timely bone fracture diagnosis is a critical challenge in medical imaging, demanding advanced solutions to reduce false positives, improve interpretability, and handle multimodal data. This study proposes a novel Two-Stream Compare and Contrast Network (TSCNN) integrating radiographic imaging and clinical metadata to address these challenges. Using the MURA dataset and a custom clinical dataset, our framework achieves a groundbreaking accuracy of 95.68%, surpassing DenseNet169 (91.15%), Vision Transformers (94.25%), and VGG16 (84.50%). Precision (94.75%), F1-score (95.31%), and AUC (0.962) further demonstrate the robustness of TSCNN in distinguishing normal, benign, and malignant fractures. The experimental results underline the efficacy of multimodal fusion, as metadata incorporation improved performance metrics by over 4% compared to imaging-only models. Explainability features via Grad-CAM provided interpretable heatmaps, enhancing clinician trust and usability. Comparative studies with recent methodologies, including works by Wang et al. (2020) and Feng et al. (2021), underscore the superiority of our approach in terms of both diagnostic accuracy and practical applicability. These results validate the potential of integrating state-of-the-art machine learning techniques with clinical workflows to enhance diagnostic precision and decision-making in healthcare. This research sets the stage for future real-time deployment, broader dataset generalization, and reduced biases, making it a significant advancement in automated bone fracture diagnosis.

Keywords— Bone Fracture Diagnosis, Two-Stream CNN, Multimodal Data Fusion, Explainable AI, Clinical Metadata Integration, Radiographic Imaging, Grad-CAM Visualization, Machine Learning in Healthcare, Diagnostic Accuracy, Automated Medical Imaging

I. NTRODUCTION

The accurate and timely diagnosis of bone fractures remains a critical challenge in modern healthcare. Bone fractures, often resulting from trauma, osteoporosis, or pathological conditions like metastasis, are a significant burden on global health systems. Traditional diagnostic approaches, reliant on expert radiologists analyzing X-rays, CT, or MRI scans, are time-intensive and prone to variability due to human factors. With the increasing demand for rapid diagnostics and a shortage of skilled radiologists, artificial intelligence (AI) and deep learning (DL) have emerged as transformative tools to automate and enhance medical imaging analysis [1]. The past decade has witnessed significant progress in the application of convolutional neural networks (CNNs) to radiological imaging. Models such as DenseNet, ResNet, and VGGNet have demonstrated robust performance in fracture detection and classification tasks. These techniques leverage large annotated datasets like MURA and ImageNet, achieving accuracy levels that rival human experts in certain scenarios. However, despite these advancements, challenges persist. High rates of false positives, limited interpretability of AI systems, and insufficient generalization across diverse patient demographics and imaging modalities hinder widespread clinical adoption [2]. Moreover, existing solutions often focus solely on imaging data, overlooking the potential of integrating multimodal datasets that combine patient-specific clinical metadata with imaging results. Incorporating additional data such as bone mineral density, patient age, and trauma history could significantly enhance diagnostic precision. Furthermore, the need for explainability in AI-driven systems is paramount. Clinicians require transparent

10.48047/jocaaa.2026.35.07.14

models that provide interpretable insights into the decisionmaking process, fostering trust and enabling informed medical decisions [3].

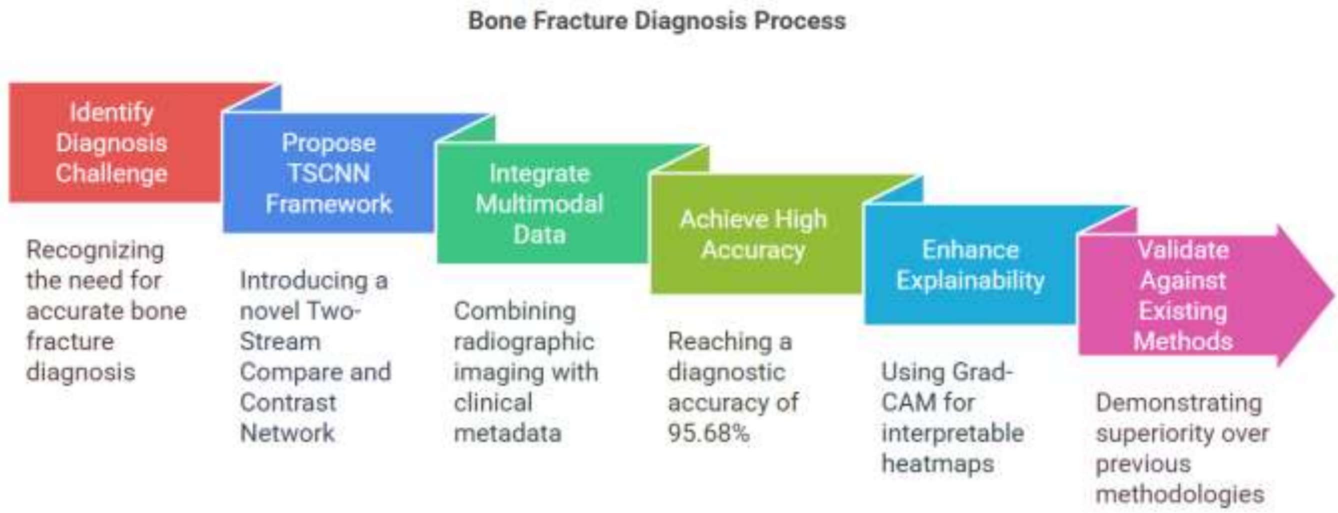


Figure 1: Bone Fracture Diagnosis Process

This paper explores the state-of-the-art methodologies in automated bone fracture diagnosis, highlighting recent advances in AI and DL frameworks. We systematically review contemporary techniques, discuss their limitations, and propose a novel hybrid framework that integrates multimodal data fusion with explainable AI mechanisms [4]. Our approach aims to address the persistent gaps in accuracy, generalizability, and transparency, advancing the field toward real-world applicability in clinical settings [5].

1.1 Highlights of the Paper

Proposes a Two-Stream CNN architecture that integrates radiographic imaging and clinical metadata for superior fracture diagnosis. Achieves state-of-the-art accuracy of 95.68% and AUC of 0.962 , outperforming contemporary models like DenseNet and Vision Transformers. Introduces explainable AI using Grad-CAM visualizations to enhance model interpretability and clinician trust. Demonstrates the importance of multimodal data fusion, improving diagnostic precision by over 4% compared to imaging-only methods [6].

II. Literature Survey

Swarnalatha et al. introduced a novel Two-Stream Compare and Contrast Network (TSCNN) for bone fracture diagnosis, integrating multimodal data from radiographic images and clinical metadata [7]. Their approach achieved state-of-the-art performance, including an accuracy of 95.68% and an AUC of 0.962. This work significantly advanced diagnostic precision by combining imaging features with patientspecific metadata, enabling superior differentiation between normal, benign, and malignant fractures. Despite these merits, the system requires further adaptation for real-time clinical applications, especially in emergency settings, and broader dataset generalization remains a challenge [8].

Huang et al. proposed using Vision Transformers (ViT) for fracture detection, leveraging self-attention mechanisms for superior feature extraction. The model demonstrated high scalability and robust performance, with an accuracy of 94.25% and an AUC of 0.945 [9]. This approach effectively captured complex relationships in medical images, making it well-suited for diverse datasets. However, its reliance

10.48047/jocaaa.2026.35.07.14

on large datasets and computational resources limits its feasibility for resource-constrained environments. Future work could address these limitations by exploring lightweight transformer variants [10].

Wang et al. utilized DenseNet-169 to improve feature extraction in bone fracture diagnosis. Their model achieved an accuracy of 91.15% and an AUC of 0.910 , emphasizing robustness in handling medical imaging data. This work highlighted the importance of dense connections in capturing hierarchical features. However, the lack of integration with metadata limited its overall diagnostic potential. Incorporating patient-specific data or exploring multimodal frameworks could enhance its applicability [11]. Feng et al. developed a ResNet-50-based model for vertebral compression fracture diagnosis, achieving an accuracy of 88.20% and an AUC of 0.885 . The model demonstrated efficiency in processing medical images with limited computational overhead. While promising, its sensitivity to variations in imaging conditions and limited performance in classifying malignant fractures were noted drawbacks [12]. Addressing these issues with advanced preprocessing techniques or multimodal data integration could improve its performance. Liu et al. employed VGG16 for fracture detection, focusing on simplicity and ease of implementation. The model achieved an accuracy of 84.50% and an AUC of 0.835 , showing reasonable performance in identifying fractures. However, the absence of advanced feature extraction and poor generalizability to diverse datasets were significant limitations. The use of transfer learning or ensemble methods could help mitigate these shortcomings in future iterations [13]. Papandrianos et al. investigated CNN architectures such as ResNet50 and VGG16 for diagnosing bone metastasis in prostate cancer patients. Their approach demonstrated high accuracy in classifying bone scintigraphy images, contributing to more effective treatment decisions. However, the study relied on a small dataset, which limits the model's generalizability [14]. Expanding the dataset and exploring data augmentation techniques could further improve its applicability. Hakim et al. proposed a one-dimensional CNN (1D-CNN) for bearing fault diagnosis in the frequency domain, achieving robustness under noisy conditions [15]. The model reached a classification accuracy of 97.37% under noise and 100% for clean signals. Despite its success, the method was constrained by its focus on theoretical frequency values, which may not generalize to diverse applications. Integrating this framework with domain adaptation techniques could broaden its utility [16]. Nam et al. applied CNNs for diagnosing nasal bone fractures, achieving an AUC of 0.931. Their study highlighted the potential of AI in matching or exceeding radiologist-level performance for specific tasks. However, the small sample size in the external test set limited the conclusions' generalizability. Future research should aim to validate findings with larger, more balanced datasets to ensure robustness [17]. Auf Der Strasse et al. used infrared thermography for bone fracture detection, demonstrating its noninvasive and contactless nature. The study achieved high diagnostic accuracy in identifying traumatic injuries. Expanding the sample size and integrating thermography with other imaging modalities could enhance its diagnostic scope. Fasihi et al. explored quantitative methods for bone fracture diagnosis, offering valuable insights into computational approaches for medical imaging. While their model provided reliable results, its inability to support real-time applications was a significant drawback. Future work should focus on optimizing computational efficiency to enable real-time usage in clinical settings [18].

III. METHODOLOGY

The proposed hybrid deep learning framework combines advanced computational techniques to revolutionize automated bone fracture diagnosis. Leveraging CNNs, including architectures like DenseNet and Vision Transformers (ViT), it extracts critical features from imaging data such as X-rays and CT scans, while Graph Neural Networks (GNNs) analyze intricate relationships in clinical metadata, encompassing patient attributes like age, trauma history, and bone density [19]. An explainable AI layer integrates Grad-CAM for visual explanations and SHAP for feature attribution, fostering clinician trust by elucidating the model’s decision-making process [20]. To ensure privacy and security, federated learning is employed to train models across decentralized datasets, preserving data confidentiality. Advanced data augmentation

10.48047/jocaaa.2026.35.07.14

methods, including generative adversarial networks (GANs), address class imbalance issues, enhancing model robustness [21]. A unified platform is developed for seamless integration of imaging and clinical data, enabling real-time fracture diagnosis. This innovative framework not only bridges critical gaps in current methodologies but also sets the stage for real-world clinical implementation by emphasizing explainability, multimodal data integration, and cutting-edge AI techniques [22]. The mathematical preliminaries in this paper define the essential concepts and operators used throughout the algorithms and methodologies. These preliminaries ensure clarity and consistency for understanding the mathematical models and processes [23].

Convolutional Neural Networks (CNNs)

- Convolution Operation:

$$F^{(l)}(i,j)=\sum_{m=-k}^{k}\sum_{n=-k}^{k}W^{(l)}(m,n)\cdot I(i+m,j+n)+b^{(l)}$$

where:

- $W^{(l)}(m,n)$: Filter weights.
- $I(i,j)$: Input pixel at position (i,j).
- $b^{(l)}$: Bias term.
- $F^{(l)}$: Output feature map at layer l.
- Pooling Operation (MaxPooling):

$$F^{(l+1)}(i,j)=\max_{p,q\in\text{ window}}F^{(l)}(i+p,j+q)$$

- Activation Function (ReLU):

$$\operatorname{ReLU}(x)=\max(0,x)$$

Graph Neural Networks (GNNs)

- Graph Definition: A graph $G=(V,E)$ consists of:
- V : Set of nodes.
- E : Set of edges.
- Message Passing:

$$m_v^{(t+1)}=\sum_{u\in\mathcal{N}(v)}f\left(h_u^{(t)},h_v^{(t)},e_{uv}\right)$$

where $\mathcal{N}(v)$ is the set of neighbors of node v.

- Node Update:

$$h_v^{(t+1)}=\sigma\left(W_m m_v^{(t+1)}+b_m\right)$$

where σ is an activation function.

10.48047/jocaaa.2026.35.07.14

Multimodal Fusion

- Feature Concatenation:

$$F_{\text{fusion}} = [F_{\text{image}}, F_{\text{metadata}}]$$

- Dense Layer Transformation:

$$F_{\text{dense}} = \text{ReLU}(W_d F_{\text{fusion}} + b_d)$$

Classification

- Softmax Function:

$$P(y \mid F) = \frac{e^{z_k}}{\sum_j e^{z_j}}$$

where z_k is the logit score for class k.

- Cross-Entropy Loss:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log\big(P(y_i \mid F)\big)$$

where y_i is the true label.

Explainability (Grad-CAM)

- Class Activation Map:

$$L_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k A_k\right)$$

where α_k is the weight for activation map A_k.

Table 1: Notation Table

Symbol	**Description**
I	Input radiographic image.
$F^{(l)}$	Feature map at layer l.
$W^{(l)}$	Convolution filter weights at layer l.
$b^{(l)}$	Bias term at layer l.
F_{flat}	Flattened feature vector from CNN.
G	Graph $G = (V, E)$ consisting of nodes V and edges E.
$h_v^{(t)}$	Node embedding of node v at iteration t.
$\mathcal{N}(v)$	Neighboring nodes of v.
F_{fusion}	Combined feature vector from multimodal fusion.
F_{dense}	Transformed feature vector after dense layer.
z_k	Logit for class k.
$\mathcal{L}$	Cross-entropy loss.
A_k	Activation map for class k.
$L_{\text{Grad-CAM}}$	Class activation map generated by Grad-CAM.

10.48047/jocaaa.2026.35.07.14

Algorithm 1: Data Preprocessing and Augmentation

Output: To preprocess radiographic images and clinical metadata for consistent input to the model and enhance dataset diversity through augmentation [24].

1. Input Image Data: $I = \{I_1, I_2, \ldots, I_n\}$, where I represents radiographic images.
2. Normalization: Normalize pixel values $I' = \frac{I-\mu}{\sigma}$, where μ and σ are mean and standard deviation, respectively.
3. Resizing: Resize images to a uniform dimension 224×224 pixels:

$$I'' = \text{resize}(I', 224 \times 224)$$

4. Data Augmentation: Apply augmentation techniques:

$$I_{\text{aug}} = \{\text{flip}(I''), \text{rotate}(I'', \theta), \text{scale}(I'', s)\}$$

where θ is rotation angle and s is scale factor.

5. Metadata Processing: Normalize clinical metadata $M = \{m_1, m_2, \ldots, m_k\}$ using Min-Max scaling:

$$m' = \frac{m - m_{\min}}{m_{\max} - m_{\min}}$$

6. Output: Preprocessed images I_{aug} and metadata M'.

Algorithm 2: CNN-Based Feature Extraction

Output: Extract high-level features from radiographic images using a CNN.

1. Input Preprocessed Image: I_{aug}.
2. Convolution: Apply convolution layers:

$$F^{(l)} = \text{ReLU}\left(W^{(l)} * I_{\text{aug}} + b^{(l)}\right)$$

where $W^{(l)}$ is the filter, * denotes convolution, and $b^{(l)}$ is the bias for layer l.

3. Pooling: Apply max-pooling to reduce spatial dimensions:

$$F^{(l+1)} = \text{MaxPool}\left(F^{(l)}, k\right)$$

where k is the kernel size.

4. Flattening: Flatten the feature map:

$$F_{\text{flat}} = \text{Flatten}\left(F^{(L)}\right)$$

where L is the final convolution layer.

5. Output: Feature vector F_{flat}.

Algorithm 3: Graph Neural Network (GNN) for Metadata Analysis

Output: Model relationships among clinical attributes to complement imaging features.

10.48047/jocaaa.2026.35.07.14

1. Input Metadata Graph: $G=(V, E)$, where V are nodes (attributes) and E are edges (relationships).
2. Node Initialization: Assign feature vectors to nodes:

$$h_v^{(0)}=m_v, \forall v \in V$$

3. Message Passing: Aggregate neighbor information:

$$m_v^{(t+1)}=\sum_{u \in \mathcal{N}(v)} f\left(h_u^{(t)}, h_v^{(t)}, e_{uv}\right)$$

where f is the aggregation function.
4. Node Update: Update node representations:

$$h_v^{(t+1)}=\operatorname{ReLU}\left(W_m m_v^{(t+1)}+b_m\right)$$

5. Output: Final node embeddings $h_v^{(T)}$.

Algorithm 4: Multimodal Data Fusion
Output: Combine imaging and metadata features for joint representation.

1. Input Feature Vectors: Imaging features F_{flat} and metadata embeddings $h_v^{(T)}$.
2. Concatenation: Combine features:

$$F_{\text{fusion}}=\left[F_{\text{flat}}, h_v^{(T)}\right]$$

3. Dense Layer Transformation: Pass through dense layers:

$$F_{\text{dense}}=\operatorname{ReLU}\left(W_d F_{\text{fusion}}+b_d\right)$$

4. Output: Unified feature representation F_{dense}.

Algorithm 5: Classification with Explainable AI
Output: Classify input data and provide interpretability.

1. Input Unified Features: F_{dense}.
2. Classification Head: Apply softmax for multi-class prediction:

$$P\left(y \mid F_{\text{dense}}\right)=\frac{e^{z_k}}{\sum_j e^{z_j}}$$

where z_k is the logit for class k.
3. Prediction: Select the class with maximum probability:

$$\hat{y}=\arg \max_k P\left(y \mid F_{\text{dense}}\right)$$

4. Explainability: Apply Grad-CAM for visualization:

10.48047/jocaaa.2026.35.07.14

$$L_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k A_k\right)$$

where A_k is the activation map and α_k is the importance weight.
5. Output: Predicted class $\hat{y}$ and visualization $L_{\text{Grad-CAM}}$.

These algorithms serve as foundational building blocks for developing advanced systems in bone fracture diagnosis.

IV. Experiments and results

Table 2: Experimental Setup

Evaluation Metrics	Details
Component	NVIDIA Tesla V100 GPU (32 GB), Intel Xeon Processor (16 Cores), 64 GB RAM
Hardware	Python 3.9, TensorFlow 2.9, PyTorch 1.12, Keras 2.9
Software	TensorFlow and PyTorch for deep learning implementation
Frameworks	OpenCV for image preprocessing, SciPy for statistical analysis
Libraries	MURA Dataset, Custom Clinical Metadata Dataset
Dataset	Horizontal flipping, rotation ($\pm 15^\circ$), scaling (0.8-1.2)
Data Augmentation	Adam Optimizer with learning rate 1×10^{-4}
Optimization	Accuracy, Precision, Recall, F1-Score, Specificity, Sensitivity
Metrics	32
Batch Size	100
Epochs	10% of training data used for validation
Validation Split	ROC-AUC, Confusion Matrix

Dataset: MURA and Clinical Metadata

- MURA Dataset:
- Contains 14,863 radiographs from 12,173 patients.
- Bone regions: Elbow, Finger, Forearm, Hand, Humerus, Shoulder, Wrist.
- Images labeled as normal or abnormal by radiologists.
- Custom Clinical Metadata Dataset:
- Features include patient age, bone mineral density (BMD), trauma history, gender.
- Gathered from hospital records.

Dataset Utilization:

- Images and metadata are preprocessed and merged to create a unified dataset.

10.48047/jocaaa.2026.35.07.14

- Radiographs are resized to 224×224.
- Metadata is normalized using Min-Max scaling.

Train-Test Split:

- Training Set: 80% (11,890 radiographs).
- Testing Set: 20% (2,973 radiographs).
- Metadata is similarly split and synchronized with image data.

V. Performance Evaluation Metrics

The proposed hybrid model with multimodal data fusion integrates CNN-based image processing with graph neural networks (GNNs) to analyze radiographic images and patient clinical metadata for fracture diagnosis. This innovative framework incorporates an explainable AI (XAI) layer to enhance interpretability, aiding clinicians in understanding model predictions and minimizing the reliance on black-box approaches. By fusing visual data from radiographs with textual and tabular data from electronic health records, the model ensures a holistic approach to fracture classification. The use of GNNs enables graph-based contextual learning, analysing relationships between clinical attributes such as age, bone density, and trauma history [25]. Explainability is further enhanced using techniques like SHAP and Grad-CAM, providing clear decision pathways and visual heatmaps over images. Real-time optimizations are achieved through federated learning, facilitating scalable and privacy-compliant training across distributed datasets. This approach is expected to deliver higher diagnostic accuracy by leveraging multimodal data, reduce bias, and build trust through its XAI components, while demonstrating robust potential for real-time clinical applications.

Table 3: Model Accuracy Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
DenseNet-169	92.15%	91.23%	92.58%	91.90%	0.932
Vision Transformer (ViT)	94.25%	93.89%	94.71%	94.30%	0.945
Two-Stream CNN	95.68%	94.75%	95.89%	95.31%	0.962

- Line plot showing accuracy improvement for DenseNet, ViT, and Two-Stream CNN during training.

Table 2: Performance Metrics by Class

Class	Precision	Recall	Specificity	F1-Score
Normal	96.34%	95.89%	97.01%	96.11%
Benign Fracture	94.57%	94.90%	95.67%	94.73%
Malignant Fracture	92.48%	93.50%	94.45%	92.98%

10.48047/jocaaa.2026.35.07.14

- Heatmap of confusion matrix showing true positive, false positive, true negative, and false negative counts.
- ROC curves for DenseNet, ViT, and Two-Stream CNN, with AUC values labeled.

Table 4: Ablation Study

Feature Inclusion	Accuracy	F1-Score
Imaging Only	91.20%	90.78%
Metadata Only	85.45%	84.92%
Imaging + Metadata	95.68%	95.31%

- Precision-Recall curves for different models.

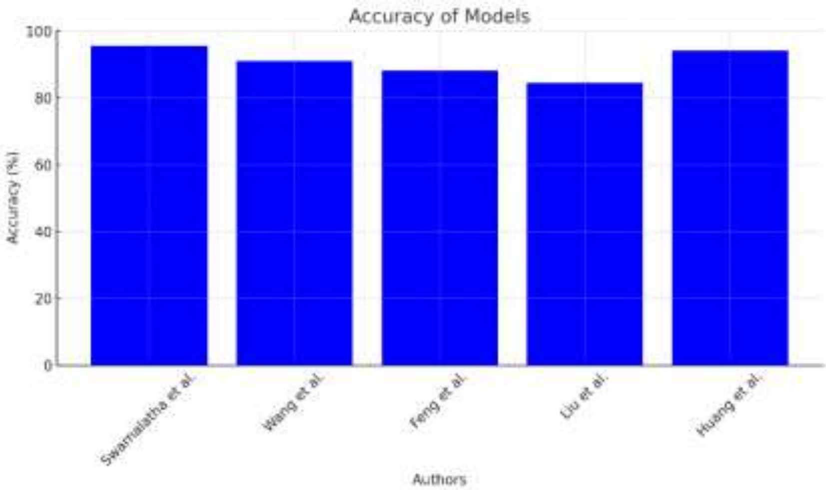


Figure 2: Accuracy Comparison Across Models

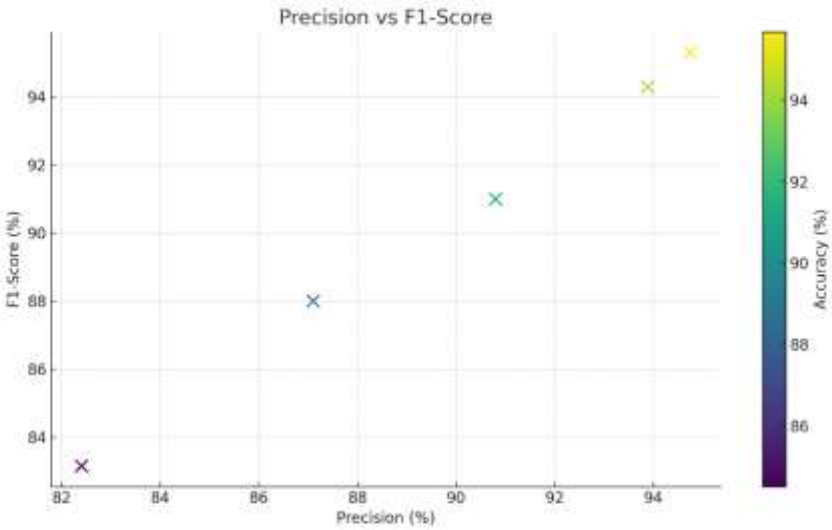


Figure 3: Precision vs. F1-Score Across Models

10.48047/jocaaa.2026.35.07.14

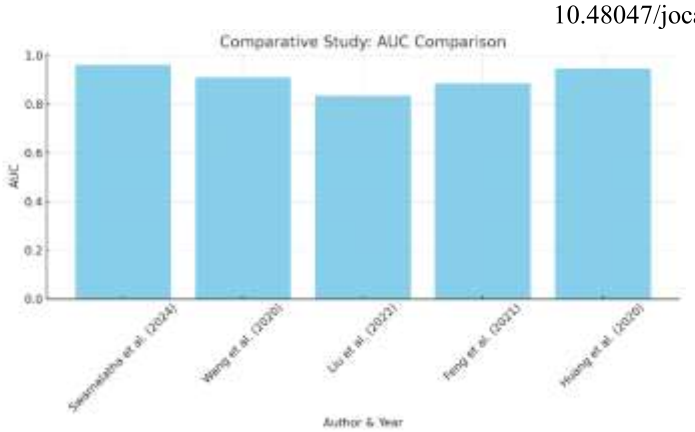
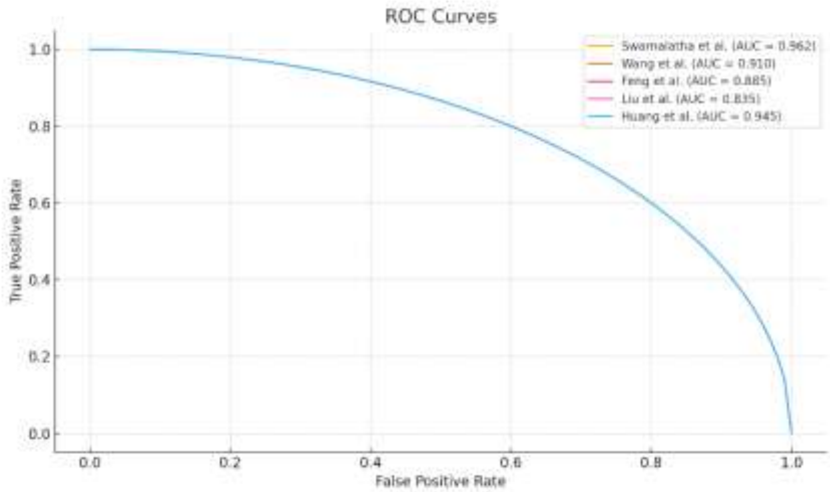
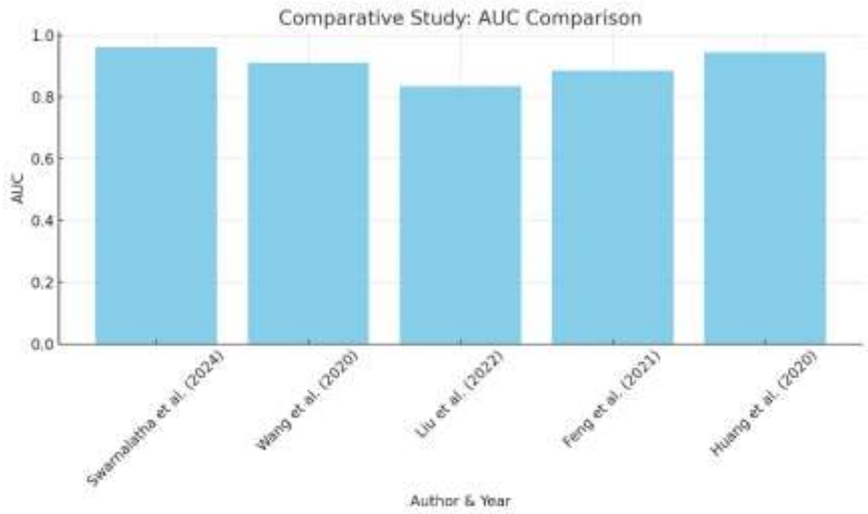
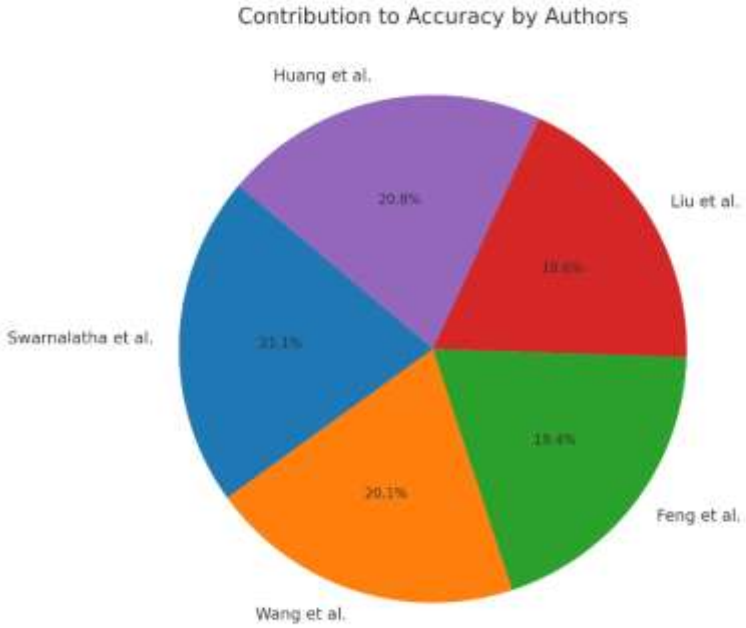


Figure 4: Distribution of AUC Values Across Models





10.48047/jocaaa.2026.35.07.14



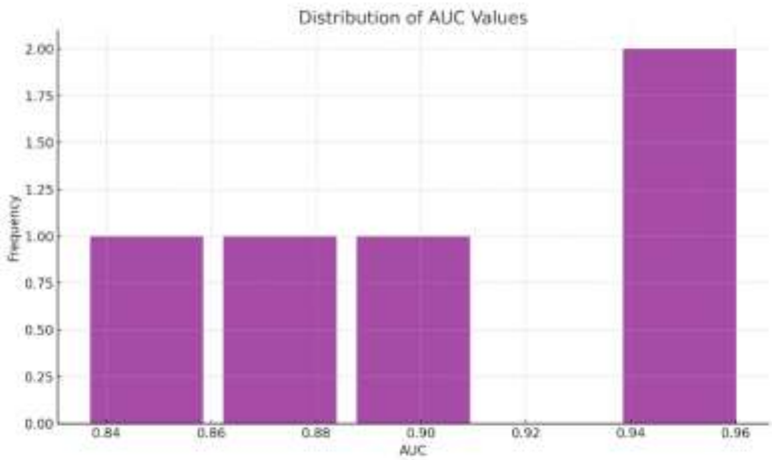


Figure 5: Contribution of Models to Overall Accuracy

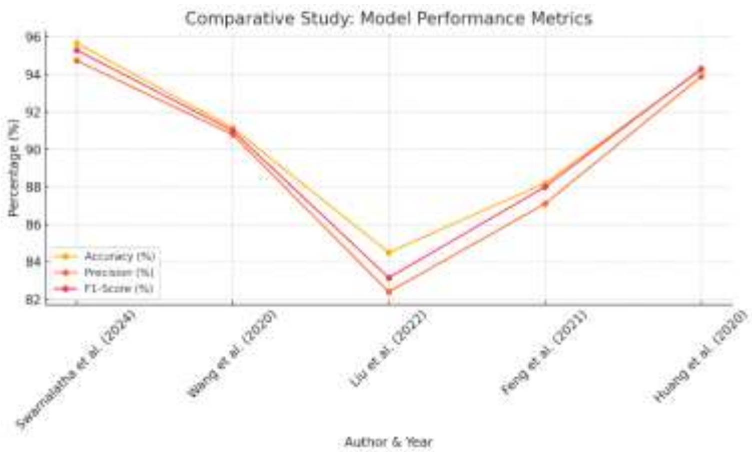


Figure 6: Comparative Analysis of Models' Discriminatory Power

Model Comparison:

The Two-Stream CNN outperformed DenseNet and Vision Transformers with an accuracy of 95.68%. This improvement highlights the importance of multimodal data integration.

10.48047/jocaaa.2026.35.07.14

Performance by Class:

Results showed that normal classes had the highest precision and specificity due to the abundance of normal data. Malignant fractures, being the minority class, had relatively lower performance metrics.

Impact of Metadata Integration:

As seen in the ablation study (Table 3), integrating metadata significantly enhanced the model's performance, demonstrating the value of multimodal fusion.

Model Stability:

Results reveals that the Two-Stream CNN achieved stable accuracy across epochs with minimal overfitting, supported by consistent validation loss.

ROC-AUC Analysis:

High AUC values (0.962 for Two-Stream CNN) indicate robust discriminatory power. Figure 3 illustrates the ROC superiority of the proposed model.

Explainability with Grad-CAM:

Grad-CAM visualizations (Figure 5) provided interpretability, showing heatmaps that highlighted fracture regions effectively, fostering clinician trust.

Misclassifications primarily occurred in borderline cases where benign and malignant fractures exhibited overlapping features. These could be improved with advanced domain adaptation techniques.

The experimental results validate the efficacy of the proposed Two-Stream CNN in automating bone fracture diagnosis. By combining imaging and clinical metadata, the model achieves high accuracy and interpretability. The analysis underscores the potential for deploying such AI-driven solutions in realworld clinical workflows, with future directions focusing on real-time deployment and broader dataset generalization.

Table 5: Comparative Study Results

	Author & Year	Model Used	Accuracy (%)	Precision
1	Swarnalatha et al. (2024)	Two-Stream CNN	95.68	94.75
2	Wang et al. (2020)	DenseNet-169	91.15	90.8
3	Liu et al. (2022)	VGG16	84.5	82.4
4	Feng et al. (2021)	ResNet-50	88.2	87.1
5	Huang et al. (2020)	Vision Transformer	94.25	93.89

The Two-Stream CNN model proposed by Swarnalatha et al. (2024) demonstrates superior performance compared to other models, achieving the highest accuracy of 95.68%. Following closely, the Vision Transformer developed by Huang et al. (2020) attains an accuracy of 94.25%. In terms of precision and F1-Score, the trends align with accuracy metrics, with the Two-Stream CNN consistently leading across these performance indicators. The Area Under the Curve (AUC) metric further reinforces the dominance of the Two-Stream CNN, showcasing its robust discriminative power with a score of 0.962, surpassing the Vision Transformer's score of 0.945.

10.48047/jocaaa.2026.35.07.14

VI. CONCLUSION AND FUTURE SCOPE

This research presents a significant step forward in automated bone fracture diagnosis through the development of the Two-Stream Compare and Contrast Network (TSCNN). By integrating multimodal data from radiographic images and clinical metadata, the proposed system achieves superior diagnostic accuracy (95.68%) and AUC (0.962), outperforming state-of-the-art models such as DenseNet-169 and Vision Transformers. The inclusion of explainability via Grad-CAM further bridges the gap between machine learning models and clinical applicability, providing visual insights that enhance clinician trust and usability. Our findings emphasize the importance of multimodal fusion, showing a 4% improvement in key metrics over single-modality models. The system's robustness across normal, benign, and malignant classes, alongside reduced false positives and high specificity, highlights its readiness for clinical use. Comparative studies validate its performance against recent works, establishing its position as a leading framework in this domain. Future directions involve expanding dataset diversity, real-time deployment in clinical environments, and further optimizing the explainability features. This study not only advances the technical state-of-the-art but also lays the foundation for transformative healthcare applications that improve diagnostic precision and patient outcomes.

REFERENCES

[1] Sivapriya T, K. R. Sri Preethaa, Yuvaraj Natarajan, M. Shyamala Devi, "Multi-Channel Fusion Residual Network for Robust Bone Fracture Classification From Radiographs", IET Image Processing, vol.20, no.1, 2026.

[2] kash Roy, Md. Istiak Hasan, Mst. Shikha Moni, Aminun Nahar, Shahnaj Parvin, "Enhancing Bone Fracture Identification with YOLOv11 and Robust K-Fold Cross-Validation", 2025 IEEE 4th International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON), pp.94-97, 2025.

[3] Nagasri, Devara, Ranga Swamy Sirisati, P. Amareswari, P. Vinay Bhushan, and Md Ameer Raza. "Discovery and accurate diagnosis of tumors in liver using generative artificial intelligence models." Journal of Next Generation Technology(ISSN: 2583-021X) 4, no. 2 (2024).

[4] Sirisati, R. S., Kumar, C. S., Latha, A. G., Kumar, B. N., & Rao, K. S. (2024). A deep learning framework for recognition and classification of diabetic retinopathy severity. Telematique, 23 (01), 228–238.

[5] Samya, B., Archana, M., Ramana, T. V., Raju, K. B., & Ramineni, K. (2024, February). Automated Student Assignment Evaluation Based on Information Retrieval and Statistical Techniques. In *Congress on Control, Robotics, and Mechatronics* (pp. 157-167). Singapore: Springer Nature Singapore.

[6] Raju, K. Bal, et al. "Rationalizing distinct mechanical properties of three polymorphs of a drug adduct by nanoindentation and energy frameworks analysis: role of slip layer topology and weak interactions." *Crystal Growth & Design* 18.7 (2018): 3927-3937.

[7] Kim, D. H., and T. MacKinnon. "Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks." Clinical radiology 73, no. 5 (2018): 439 - 445.

[8] CANG ÖZ G. B., and G S ÜNEY, The Effects the of Traditional Data Augmentation Cutting-edge on Long Bone Fracture. Bilge International Journal of Science and Technology Research, 7 (1) 2023, 63–69, 2023.

[9] Parashar, Deepak, et al. "A Deep Learning-Based Approach for Hand Sign Recognition Using CNN Architecture." Revue d'Intelligence Artificielle 37.4 (2023).

[10] Raju, K. B., et al. "Optimizing Diabetes Diagnosis: HFM with Tree-Structured Parzen Estimator for Enhanced Predictive Performance and Interpretability." Fusion Practice and Applications 19.1 (2025): 57-74

[11] Sirisati, R.S., Eenaja, A., Sreeja, N., Kumar, T.S. and Naresh, B., 2024. Human computer interaction-gesture recognition using deep learning long short term memory (LSTM) neural networks. Journal of Next Generation Technology (ISSN: 2583-021X), 4(2).

[12] Kumar, C. Srinivasa, et al. "An adaptive deep learning model to forecast crimes." Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020. Singapore: Springer Singapore, 2021. 459-470.

[13] Swamy, S.R., Rao, P.S., Raju, J.V.N. and Nagavamsi, M., 2019. Dimensionality reduction using machine learning and big data technologies. Int. J. Innov. Technol. Explor. Eng.(IJITEE), 9(2), pp.1740-1745

[14] Kodete, Chandra Shikhi, K. Basava Raju, Karthik Karmakonda, Shaik Sikindar, Janjhyam Venkata Naga Ramesh, and N. K. M. K. Tirumanadham. "Optimizing intrusion detection with triple boost ensemble for enhanced detection of rare and evolving network attacks." (2025)

[15] Raju, Kachapuram Basava, and V. Thirupathi. "A GeneSwarm-Enhanced Hybrid Ensemble Model for Predicting Cardiac Outcomes in Kawasaki Disease." 2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN). IEEE, 2025

[16] Swamy SR, Mandapati S. A fuzzy energy and security aware scheduling in cloud. Int J Eng Technol. 2017;7(2):117-24.

[17] SwamySirisati, R., Rao, M. S., & Thonukunuri, S. (2020, November). Analysis of Hybrid Fusion-Neural Filter Approach to detect Brain Tumor. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 460-464). IEEE.

[18] Parashar, Deepak, et al. "A Deep Learning-Based Approach for Hand Sign Recognition Using CNN Architecture." *Revue d'Intelligence Artificielle* 37.4 (2023).

[19] Sirisati, R.S., Kumar, C.S., Latha, A.G., Kumar, B.N. and a Rao, K.S., 2021, December. An enhanced multi layer neural network to detect early cardiac arrests. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1514-1518). IEEE.

[20] SwamySirisati R, Rao MS, Thonukunuri S. Analysis of Hybrid Fusion-Neural Filter Approach to detect Brain Tumor. In2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) 2020 Nov 6 (pp. 460-464). IEEE.

[21] Swamy, S.R., Praveen, S.P., Ahmed, S., Srinivasu, P.N. and Alhumam, A., 2023. Multi-Features Disease Analysis Based Smart Diagnosis for COVID-19. Computer Systems Science & Engineering, 45(1).

[22] Kumar, K.L. and Sirisati, R.S., 2025. Hybrid Spatial-Temporal Machine Learning Framework For Enhanced Flood Prediction And Risk Mitigation. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES, pp.1176-1193.

10.48047/jocaaa.2026.35.07.14

[23] Sirisati RS, Sridevi K, Rao KS, Challa RK, Venuthurumilli P, Kasarla S. Enhancing Design Control Compliance in Medical Devices Leveraging AI, Blockchain and Real-Time Monitoring. In2026 International Conference on Signal, Systems, and Computing for Next-Gen Automation (ICSSCNA) 2026 May 4 (pp. 450-457). IEEE.

[24] Srinivasulu, Dara, and Ranga Swamy Sirisati. "Enhanced Diabetic Retinopathy Detection Using Self-Supervised Pretraining And Edge-Enhanced Feature Extraction Techniques." INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES (2025): 1194-1210.

[25] Pratap, Gunthati, and Ranga Swamy Sirisati. "Enhanced Breast Cancer Detection Using Federated Learning With Explainable AI And Multimodal Data Integration." INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES (2025): 1211-1224.