ORIGINAL RESEARCH PAPER

Volume - 15 | Issue - 04 | April - 2026 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

EXPLAINABLE AI FOR BLACK-BOX MODELS IN MEDICAL IMAGE DIAGNOSIS USING CNNS AND XAI TECHNIQUE

Computer Science

Mr. Prakash S — Assistant Professor, Dept. of Artificial Intelligence & Data Science, Jansons Institute of Technology, Coimbatore, India.

Adineni Venkata Sai Koushik — UG Student, Dept. of Artificial Intelligence & Data Science, Jansons Institute of Technology, Coimbatore, India.

Sahana M — UG Student, Dept. of Artificial Intelligence & Data Science, Jansons Institute of Technology, Coimbatore, India.

Teresa Danielle Sabrina S — UG Student, Dept. of Artificial Intelligence & Data Science, Jansons Institute of Technology, Coimbatore, India.

Tupili Akash Reddy — UG Student, Dept. of Artificial Intelligence & Data Science, Jansons Institute of Technology, Coimbatore, India.

ABSTRACT

Automated medical image classification has been improved by deep learning techniques. Nevertheless, the inability of Convolutional Neural Networks (CNNs) to provide interpretability has hindered their adoption in medical imaging diagnosis. In this paper, a lightweight and interpretable medical image diagnosis system is proposed by integrating a customized CNN model with Explainable Artificial Intelligence (XAI) approaches, namely Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). In this system, multi-class classification of medical images is accomplished. In addition, a visualization of the explanations of the images is also achieved. Furthermore, a web-based interface is integrated into the system to enable the upload of images and the generation of reports. Experimental results show that the system provides reliable performance and interpretability for AI-based medical imaging diagnosis.

KEYWORDS

Convolutional Neural Network, Explainable Artificial Intelligence, Medical Image Diagnosis, LIME, SHAP, Deep Learning.

I. INTRODUCTION

Medical imaging supports disease detection and diagnosis through modalities such as MRI and X-ray. Deep learning, particularly Convolutional Neural Networks (CNNs), is widely used for automated medical image analysis due to its ability to learn hierarchical features.However, CNN models operate as black boxes, limiting interpretability and trust in clinical applications. To address this, an explainable medical image classification system is proposed that integrates a custom CNN with LIME and SHAP techniques. The system provides both accurate predictions and visual explanations, enabling better understanding of model decisions and supporting reliable medical analysis.

II. Problem Statement

Although CNN-based models achieve high accuracy in medical image diagnosis, they lack explainability, making it difficult for doctors to understand and trust the predictions. This black-box behavior restricts real-world clinical adoption, especially in critical healthcare scenarios.The problem addressed in this project is to develop a transparent and explainable MRI-based disease detection system that not only predicts diseases accurately but also provides visual explanations for each prediction using Explainable Al techniques.

III. Proposed System

The proposed system presents a framework for medical image diagnosis by integrating a Convolutional Neural Network (CNN) with XAI techniques. Medical images, particularly MRI scans, are preprocessed through resizing and normalization before being fed into the CNN model for multi-class classification.To enhance interpretability, SHAP is used to provide global feature importance, while LIME generates local explanations for individual predictions. The system produces visual outputs such as heatmaps and highlighted regions to indicate important areas influencing the model's decision.A web-based interface allows users to upload images, view predictions with confidence scores, and analyze results interactively.

IV. System Architecture

The system is based on a modular pipeline for the classification and interpretation of medical images. The system takes input in the form of medical images uploaded through a web interface. The images are then processed and input into a pre-trained CNN for classification purposes. The model provides classification predictions along with confidence values. The XAI methods are used for the interpretation of the classification output provided by the model and highlight the key areas of the image used for classification purposes. The system then provides visual output along with a diagnostic report.

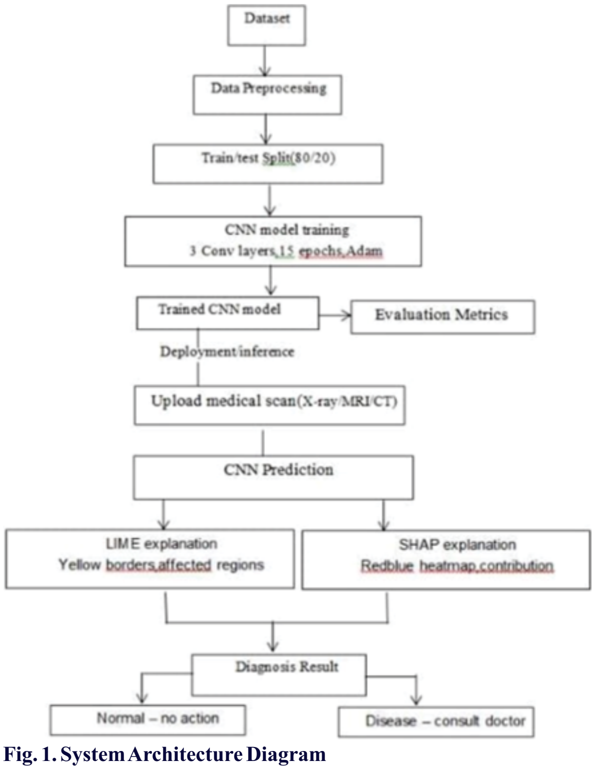

Fig. 1. System Architecture Diagram

V. Implementation

The implementation of the system utilizes Python along with TensorFlow/Keras for CNN-based medical image classification. Preprocessing of input images is done for resizing and normalization before classification/prediction. SHAP and LIME are used for obtaining visual explanations such as heatmaps and highlighted regions for images. Streamlit is used for creating a user interface for uploading images and displaying results, while SQLite is used for storing reports along with PDF generation capabilities.

VI. Methodology

The system follows a structured pipeline consisting of the following stages:

- **Image Acquisition:** Medical images are uploaded through a web-based interface.
- **Preprocessing:** Images are resized, denoised, and normalized to ensure consistent input quality.
- **Classification:** A custom CNN model performs multi-class disease classification and generates probability scores.
- **Explainability:** SHAP and LIME are applied to produce visual explanations highlighting important regions.
- **Report Generation:** The system creates an automated report containing predictions, confidence scores, and visual outputs.

VII. System Modules in Detail

A. User Interface Module

Handles user interaction including image upload, prediction display, and report download through a Streamlit interface.
Provides a simple and interactive environment for viewing results and explanations.

B. Image Preprocessing Module

Prepares input images through resizing, format conversion, and normalization.
Ensures consistent data quality for improved model performance.

C. CNN Model Module

Performs core classification using a trained Convolutional Neural Network.
Extracts features and classifies images into disease categories.

D. Prediction Module

Processes input images through the trained CNN model.
Generates probability scores and determines the final predicted class.

E. Explainable AI (XAI) Module

Provides interpretability using LIME for local explanations and SHAP for feature importance.
Highlights key regions influencing the model's predictions.

F. Visualization Module

Displays results using visual outputs such as heatmaps and highlighted regions.
Helps users understand and validate model decisions.

G. Database Module

Manages storage of user data, images, and prediction results using SQLite.
Enables secure access to previous records and reports.

H. Report Generation Module

Generates downloadable PDF reports containing predictions and explanations.
Includes relevant details and disclaimers for documentation purposes.

VIII. RESULTS AND DISCUSSION

A. Performance Evaluation

The CNN model was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC. The dataset was divided using a stratified split, and cross-validation was performed to ensure reliable evaluation. Despite dataset limitations, the model achieved consistent performance across multiple classes. The close alignment between training and validation results indicates minimal overfitting and stable generalization.

B. Performance Metrics

TABLE II. Test Set Performance Metrics

Metric	Value
Accuracy	91.00%
Precision	90.00%
Recall	89.00%
F1-Score	89.00%

CNN Accuracy Curve:

The accuracy curve shows a steady increase in training accuracy over epochs, eventually stabilizing. Validation accuracy closely follows the same trend, indicating good generalization. The minimal gap between training and validation accuracy suggests reduced overfitting and consistent model performance.

CNN Loss Curve

The loss curve shows a gradual decrease in training loss as epochs progress, indicating effective model convergence. Validation loss follows a similar pattern and remains close to training loss. This reflects stable learning behavior and balanced optimization.

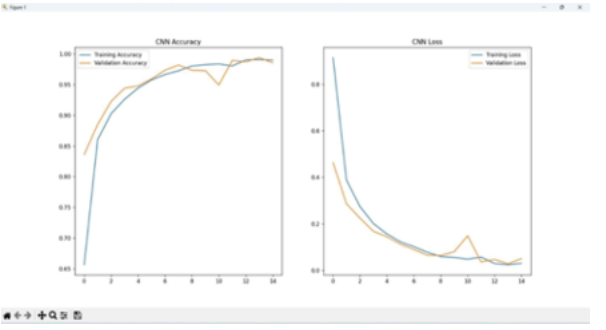

Fig. 2. CNN Accuracy and CNN Loss

Confusion Matrix of the Proposed CNN Model:

The confusion matrix shows that most samples are correctly classified, with minor misclassifications observed between visually similar classes.

Fig 3: Confusion Matrix of the Proposed CNN Model

C. Outcome / Result

The system provides classification results along with confidence scores for uploaded medical images. Visual explanations generated using SHAP and LIME highlight important regions influencing the prediction. Results are displayed through an interactive interface for easy interpretation. A concise report containing prediction details and explanations can also be generated for reference.

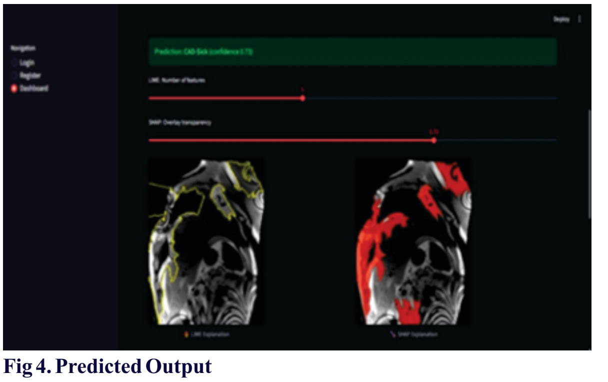

Fig 4. Predicted Output

IX. CONCLUSION

This paper presented an explainable medical image classification framework using a custom CNN model for multi-class disease detection. The system integrates prediction with visual interpretation to support reliable decision-making in medical applications. A web-based interface enables real-time image upload, prediction, and result visualization.

Experimental results show that the model achieves consistent performance while maintaining stable generalization. The proposed approach demonstrates the feasibility of combining accuracy with interpretability for practical diagnostic support.

Future work includes evaluating the system on larger datasets, incorporating clinical validation, and extending the framework to additional medical imaging modalities.

VI. REFERENCES

[1] Shailesh Bhosekar, Prabhishek Singh, and Deepak Garg, "A Comparative Analysis of Multi-Modal Medical Image Fusion Techniques using MSVD, WPD, PCA, and DWT", 2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI), 2025, doi: 10.1109/IC3ECSBHI63591.2025.10990897.

[2] Shailesh Bhosekar, Prabhishek Singh, and Deepak Garg, "A Recent Advancements in Multi-Modal Medical Image Fusion Techniques using SWT, NSST, NSCT, and CNN", 2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI), 2025, doi: 10.1109/IC3ECSBHI63591.2025.10991209.

[3] Shailesh Bhosekar, Prabhishek Singh, and Deepak Garg, "A Normalized Mutual Information-based Multi-Modal Medical Image Fusion Technique using Non Sub-sampled Shearlet Transform and Method Noise", 2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2024, doi: 10.1109/PDGC64653.2024.10984038.

[4] Xinjian Wei, Yu Qiu, Jing Xu, et al., "HCFMaNet A Novel Holistic Cross-modal Fusion Mamba Network for Multi-modal Medical Image Fusion", IEEE Transactions on Multimedia, 2026, doi: 10.1109/TMM.2026.3660130.

[5] Shailesh Bhosekar, Prabhishek Singh, and Deepak Garg, "A Multi-Modal Medical Image Fusion Technique Using Structural Similarity Index Measure and Correlation Coefficient", 2024 4th International Conference on Innovative Sustainable Computational Technologies (CISCT), 2024, doi: 10.1109/CISCT62494.2024.11134217.

[6] Shailesh Bhosekar, Prabhishek Singh, and Deepak Garg, "Role, Significance, and Impact of Method Noise in Multi-Modal Medical Image Fusion", 2025 International Conference on Emerging Systems and Intelligent Computing (ESIC), 2025, doi: 10.1109/ESIC64052.2025.10962578.

[7] Xiaobo Zhu, He Sun, Yuhan Wang, et al., "Prediction of Lymph Node Metastasis in Colorectal Cancer Using Intraoperative Fluorescence Multi-Modal Imaging", IEEE Transactions on Medical Imaging, 2025, doi: 10.1109/TMI.2024.3510836.

[8] Xinru Dai, Tai Ma, Haibin Cai, et al., "Unsupervised Hierarchical Translation-Based Model for Multi-Modal Medical Image Registration", ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, doi: 10.1109/ICASSP43922.2022.9746324.

[9] P. Singh and A. Kumar, "AI-Based Medical Image Diagnosis Using CNN and Explainable Models," International Journal of Advanced Computer Science and Applications, vol. 16, no. 2, 2025.

[10] S. Patel, R. Shah, and K. Mehta, "Explainable Artificial Intelligence in Healthcare: A Review of LIME and SHAP Applications," IEEE Access, vol. 13, 2025.

[11] M. Zhou, J. Chen, and H. Li, "Explainable AI for Medical Imaging: Techniques and Applications," IEEE Reviews in Biomedical Engineering, vol. 17, pp. 45–60, 2024.

[12] S. Bhosekar, P. Singh, and D. Garg, "A Multi-Modal Medical Image Fusion Technique Using Structural Similarity Index Measure and Correlation Coefficient," 2024 4th International Conference on Innovative Sustainable Computational Technologies (CISCT), 2024, doi: 10.1109/CISCT62494.2024.11134217.

[13] S. Bhosekar, P. Singh, and D. Garg, "Role, Significance, and Impact of Method Noise in Multi-Modal Medical Image Fusion," 2025 International Conference on Emerging Systems and Intelligent Computing (ESIC), 2025, doi: 10.1109/ESIC64052.2025.10962578.

[14] X. Zhu, H. Sun, Y. Wang, et al., "Prediction of Lymph Node Metastasis in Colorectal Cancer Using Intraoperative Fluorescence Multi-Modal Imaging," IEEE Transactions on Medical Imaging, 2025, doi: 10.1109/TMI.2024.3510836.

[15] X. Dai, T. Ma, H. Cai, et al., "Unsupervised Hierarchical Translation-Based Model for Multi-Modal Medical Image Registration," ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, doi: 10.1109/ICASSP43922.2022.9746324.

[16] H. Li, X. Wu, and J. Kittler, "MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image Fusion," IEEE Transactions on Image Processing, vol. 31, pp. 256–270, 2022, doi: 10.1109/TIP.2021.3135697.

[17] Tonekaboni, S., Joshi, S., McCradden, M.D., and Goldenberg, A. (2023) 'Explainable AI for Clinical Decision Support Systems: A Review', IEEE Journal of Biomedical and Health Informatics, Vol. 27, No. 5, pp. 2345–2358.

[18] Zhang, Y., Chen, X., and Li, H. (2023) 'Deep Learning-Based Medical Image Classification with Explainable AI Techniques', Computers in Biology and Medicine, Vol. 156, 106697.

[19] Singh, A., Kumar, R., and Gupta, P. (2023) 'Explainable Artificial Intelligence for Medical Imaging: A Survey of LIME and SHAP Techniques', IEEE Access, Vol. 11, pp. 45678–45695.

[20] Patel, V., Shah, M., and Mehta, K. (2024) 'Improving Trust in AI-Based Medical Diagnosis Using Explainable Models', IEEE Access, Vol. 12, pp. 12345–12360