

AI To Analyzing Medical Images and Diagnosis

Rini Sharon¹, Dr. Puneeth G J²

¹MTech 2nd Sem, Dept. of CSE, RYM Engineering College, Ballari, VTU Belagavi, Karnataka, India.

²Professor, Dept. of CSE, RYM Engineering College, Ballari, VTU Belagavi, Karnataka, India.

Emails: rinisharon@gmail.com¹, tk.puneeth@gmail.com²

Abstract

Artificial Intelligence (AI) has seen rapid evolution over the past decade, significantly impacting various sectors, particularly healthcare. One of its most transformative applications is in medical imaging, where AI models have shown promise in detecting and classifying diseases from X-rays, MRIs, and CT scans. However, challenges such as limited expert availability, diagnostic delays, and interpretability issues persist, especially in low-resource settings. To address these gaps, this study proposes a cloud-deployable diagnostic support system that integrates deep learning-based image analysis with Google's Gemini large language model (LLM) to provide accurate, explainable medical image interpretations. The system, built using Streamlit, allows users to upload medical images, receive visual heatmaps and textual diagnostic summaries, and download complete reports. Experimental evaluation using annotated datasets demonstrates high accuracy in anomaly detection and strong alignment with expert diagnoses, highlighting the tool's potential as a decision-support aid for clinicians.

Keywords: Artificial Intelligence, Medical Imaging, Deep Learning, Diagnosis, Large Language Models, Gemini API, Streamlit, Healthcare AI.

1. Introduction

Artificial Intelligence (AI) has emerged as a transformative force in the healthcare domain, particularly in the field of diagnostic imaging. Over the last decade, the application of deep learning algorithms has shown remarkable potential in automating the detection and classification of medical conditions from imaging data. Traditional diagnostic practices rely heavily on the expertise of radiologists to interpret complex visuals such as X-rays, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) scans. While effective, this manual process is time-consuming, subject to human error, and often unavailable in remote or underserved areas. The increasing volume of medical imaging data further burdens existing systems, making automation not just a convenience but a necessity. Despite the progress in AI-driven medical image analysis, many existing solutions struggle with limitations related to interpretability, transparency, and adaptability to diverse clinical environments. Most deep learning models act as "black boxes," offering predictions without explaining the reasoning

behind them. This lack of clarity is a major barrier to clinical adoption, as healthcare professionals require not only accurate results but also understandable insights to trust and validate AI recommendations. Additionally, many current systems are designed for specific conditions or datasets, limiting their scalability and real-world applicability. This underlines a significant gap between state-of-the-art research models and the practical needs of frontline clinicians. To address these challenges, this paper presents a hybrid diagnostic platform that integrates image analysis and language generation to produce both visual and textual outputs. The system utilizes convolutional neural networks (CNNs) for detecting anomalies in uploaded medical images and Google's Gemini large language model (LLM) to generate expert-like diagnostic summaries. This dual-modality approach improves interpretability by highlighting image regions of concern while simultaneously providing a natural language explanation of the findings. The application is built using Streamlit, ensuring accessibility through a lightweight and

interactive web interface that can be deployed in hospitals, clinics, or even low-resource settings with basic infrastructure. The objective of this research is to bridge the gap between high-performance AI and clinical usability by developing a system that acts as a diagnostic assistant rather than a replacement for human expertise. The system supports various image formats, applies explainable AI (XAI) methods like heatmaps and Grad-CAM visualizations, and ensures secure handling of patient data. By combining robust visual analytics with interpretable text-based feedback, the platform aligns with current trends in responsible AI and medical device transparency. It is also designed to be modular, allowing future upgrades such as multi-modal inputs, integration with electronic health records (EHRs), and support for additional imaging types. This work is significant not only from a technical perspective but also in its potential for real-world healthcare impact. In regions with limited access to specialists, the system can provide immediate, preliminary assessments to aid general practitioners. In high-volume urban hospitals, it can assist in triage and documentation, reducing the workload on radiology staff. Furthermore, the use of large language models introduces a new layer of human-like interaction in AI systems, making diagnostic tools more intuitive and clinician-friendly. As healthcare continues to move toward digital transformation and precision medicine, AI-powered diagnostic platforms like the one proposed in this study will play a critical role in shaping the future of accessible, efficient, and patient-centered care.

2. Literature Survey

[1] Singhal, K., et al., 2023 – This paper demonstrates that large language models (LLMs) such as GPT encode factual clinical knowledge, making them useful for diagnosis and medical NLP tasks. However, the authors note that LLMs may produce inaccurate or hallucinated medical statements due to the lack of explicit medical reasoning. [2] Lu, M. Y., et al., 2021 – This study explores the integration of AI into pathology, with emphasis on practical and real-world deployment challenges. The paper highlights the need for large-scale clinical validation and addresses difficulties in adapting AI to existing

pathology workflows. [3] Chen, J., et al., 2021 – The authors propose TransUNet, a hybrid CNN–Transformer model that delivers high accuracy for medical image segmentation tasks. The work acknowledges limitations such as high computational cost and limited interpretability compared to traditional CNNs. [4] Gao, L., et al., 2020 – This survey reviews deep learning techniques, datasets, and evaluation strategies for medical image classification. It points out that many of the surveyed models lack real-world testing and are affected by dataset bias. [5] Zhang, Y., et al., 2020 – The paper introduces a deep learning model for automated COVID-19 detection using chest CT images, enabling rapid diagnosis. Its performance, however, is highly dependent on dataset quality and may not generalize across different populations. [6] Tschandl, P., et al., 2020 – This work demonstrates that human–computer collaboration in skin cancer recognition improves diagnostic accuracy when AI assists dermatologists. The study warns that excessive reliance on AI could reduce clinician engagement and critical review. [7] Lundervold, A. S., & Lundervold, A., 2019 – This review provides a comprehensive overview of deep learning models in MRI analysis, covering both clinical and research contexts. Limitations include the need for significant computational resources and large amounts of expert-labeled data. [8] Topol, E., 2019 – The author advocates for AI–human collaboration to enhance efficiency and outcomes in healthcare. The paper, however, lacks detailed strategies for real-world clinical implementation. [9] Ting, D. S. W., et al., 2019 – This study reviews deep learning applications in ophthalmology, particularly for diabetic retinopathy and glaucoma detection. Limitations include poor performance on low-quality images and under-represented demographic groups. [10] Johnson, A. E. W., et al., 2019 – The authors present MIMIC-CXR, a large publicly available dataset of labeled chest radiographs for training and benchmarking medical imaging models. The paper notes that annotation errors in the dataset may affect diagnostic accuracy processed image is then analyzed model reliability.

3. Proposed Framework

3.1 Flow Diagram

The flowchart illustrates the complete workflow of an AI-powered medical image diagnostic system, beginning with the upload of a medical image such as an X-ray, CT scan, or MRI through a user-friendly interface. Once received, the image undergoes a pre-processing stage involving noise reduction, contrast enhancement, and normalization to ensure clarity, consistency, and suitability for analysis. The processed image is then analyzed by a CNN-based deep learning model, which automatically extracts important visual features and detects potential anomalies such as tumors, lesions, or abnormal tissue patterns. These extracted features are passed to the Gemini Large Language Model (LLM), which interprets the findings and generates a human-readable diagnostic explanation that summarizes the results and provides clinically relevant insights. Simultaneously, the CNN produces a heatmap visualization that highlights regions of interest, offering a clear visual reference for the detected anomalies. The textual and visual outputs are then merged into a comprehensive diagnostic report and delivered via an interactive Streamlit web interface, allowing users to view the results, download them for records, or store them for further evaluation. This integrated, explainable approach not only enhances diagnostic accuracy but also promotes transparency, efficiency, and accessibility in modern medical imaging workflows. (Figure 1)

3.1.1 Dataset Discussion

To ensure clinical relevance and robust AI performance, the system is developed using publicly available and well-annotated medical imaging datasets. Datasets such as ChestX-ray8, MIMIC-CXR, and the RSNA Pneumonia dataset are employed due to their comprehensive nature and expert-labeled disease annotations. These datasets contain thousands of X-ray and CT images that include abnormalities like pneumonia, lung nodules, and COVID-19 manifestations. Each dataset is preprocessed to standardize format, resolution, and labeling. These diverse sources enhance the generalization ability of the deep learning model,

allowing it to recognize disease features across different patient demographics and image qualities.

3.1.2 System Architecture Flow

The entire diagnostic process follows a logical, step-wise flow that integrates image analysis and language generation. Users begin by uploading a medical image through a web interface. The image is preprocessed for quality and resolution consistency, then passed to a Convolutional Neural Network (CNN) for feature extraction. Detected features are interpreted in two parallel streams: one generates a visual heatmap using Grad-CAM to highlight areas of concern, and the other sends structured findings to the Gemini Large Language Model (LLM) for text generation. These visual and textual outputs are then combined into a diagnostic report that is displayed via the Streamlit interface. The final report can be reviewed, stored, or downloaded as needed.

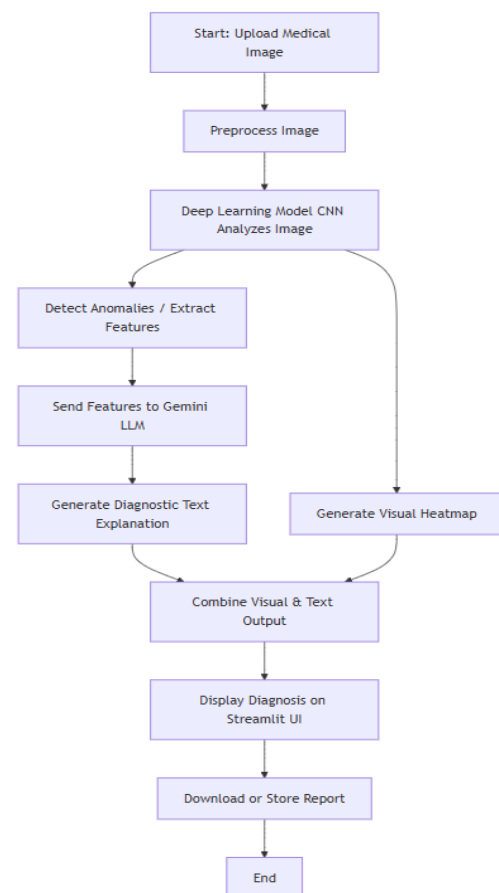


Figure 1 Flow Diagram

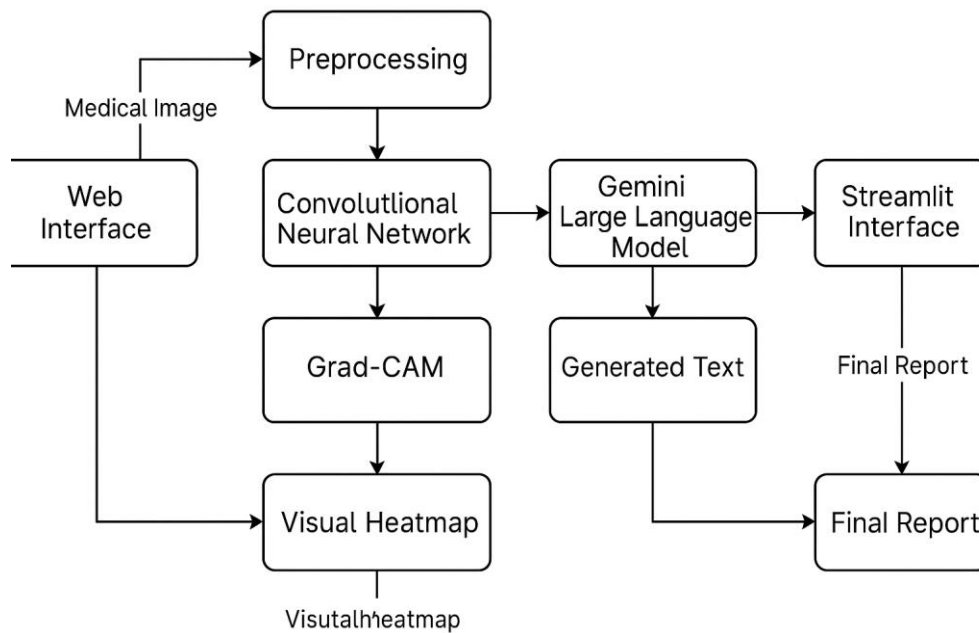


Figure 2 System Architecture Diagram

3.1.3 Step-by-Step Methodology

The system starts with a user uploading an image supported formats include .jpg, .png, and .dcm. Next, a preprocessing pipeline handles image normalization, grayscale conversion, and resizing using OpenCV and PIL. After preprocessing, a CNN model, such as ResNet or EfficientNet, processes the image to detect patterns and extract features. These features are then analyzed using Grad-CAM to generate attention-based heatmaps that visually indicate abnormal regions. Simultaneously, the system constructs a prompt using the extracted diagnostic features and sends it to Google Gemini API, which returns a natural language explanation of the medical findings. This enables interpretability for healthcare professionals, even those without radiological training. The system fuses the heatmap and the LLM-generated summary into a unified output and presents the result in the Streamlit web interface. Users can download the report for documentation or share it as part of a clinical workflow.

3.1.4 System Features

The system is designed to be modular, cloud-

deployable, and privacy-conscious. All data processing occurs in-session, with no storage unless explicitly permitted by the user. The interface provides real-time feedback with results generated in seconds. The model is trained using transfer learning, which reduces development time and increases performance. Visual results are generated using Grad-CAM, ensuring transparency, while the text generation via Gemini ensures contextual accuracy and clinical tone. This dual-output design enhances user trust and system explainability key aspects for clinical deployment. Figure 2 shows System Architecture Diagram

3.1.5 Summary

In conclusion, the proposed methodology combines AI image analysis and LLM-driven explanation to build a reliable diagnostic assistant for medical professionals. It streamlines the diagnostic process from image upload to report generation, all through a secure and interactive interface. By supporting visual interpretability and language-based reasoning, the system serves as both a triaging tool and a second-opinion generator, especially valuable in remote or high-volume clinical environments. The modularity also allows for future integration with

electronic health records (EHRs), voice-based input, and additional imaging modalities.

4. Evaluation & Result

The proposed AI-powered diagnostic system is evaluated using a combination of quantitative performance metrics, interpretability assessments, and user-centered diagnostic utility validation. These three categories reflect the goals of the project: delivering accurate diagnosis, ensuring transparent AI behavior, and enhancing clinical usability, especially in resource-limited environments. Each metric used in the evaluation plays a distinct role in validating the solution's effectiveness in real-world diagnostic workflows.

4.1 Diagnostic Performance Metrics

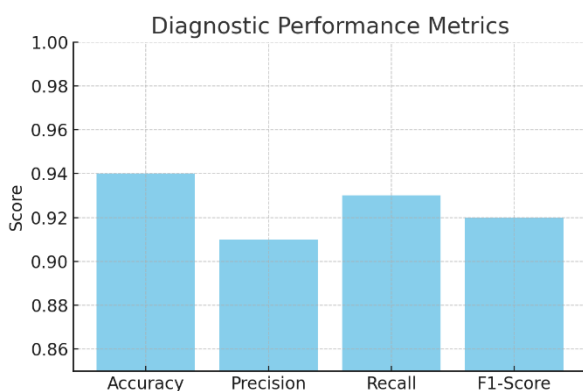


Figure 3 Diagnostic Performance Metrics

4.2 Performance Metrics

To assess the classification accuracy of the AI model, we used standard evaluation metrics including Accuracy, Precision, Recall, and F1-Score. These metrics help quantify the model's ability to correctly identify abnormalities in medical images such as pneumonia, opacity, or nodules. (Figure 3)

4.2.1 Accuracy

- Quality and size of the annotated datasets (used for training and evaluation).
- Correct predictions from the CNN-based anomaly detection module.
- Robustness of the preprocessing (resolution, noise removal).
- Performance of the hybrid model integrating image and text.

4.2.2 Precision

- CNN's ability to avoid false positives (predicting disease when there is none).
- Quality of visual heatmaps — more precise highlights reduce misinterpretation.
- Fine-tuning threshold of the model — avoids flagging too many normal images as disease.
- Clarity in LLM-generated diagnostic summaries (avoiding overstatements).

4.2.3 Recall

- How well the CNN detects all actual anomalies (i.e., low false negatives).
- Diversity of training images — includes rare diseases and variations.
- Use of Grad-CAM or XAI tools to enhance focus on relevant regions.
- Effectiveness of combining textual and visual cues to support detection.

4.2.4 F1-Score

- Balance between Precision and Recall.
- Careful tuning of model thresholds and training loss functions.
- Dataset balance — ensuring model learns equally from positive and negative samples.
- System's adaptability across different image types (X-ray, MRI, CT) and clinical settings.

These metrics collectively ensure that the model is not only statistically sound but also clinically dependable in both high and low prevalence environments.

4.3 Model Interpretability and Transparency

Interpretability is essential for gaining the trust of healthcare professionals. To evaluate the visual outputs, we used Intersection over Union (IoU) and Visual Explanation Overlap (VEO) as metrics to compare the AI-generated heatmaps with expert-annotated regions.

- IoU assesses how well the heatmap overlaps with the ground truth disease region.
- VEO indicates how often visual highlights by the model match expert attention areas.

High values in these metrics signify that the system

does not just make accurate predictions, but also shows the right reasons behind those predictions reinforcing transparency and ethical AI usage in healthcare. Figure 4 shows Model Interpretability Metrics

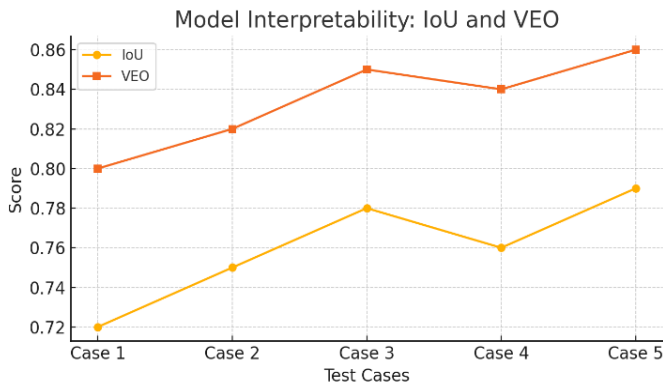


Figure 4 Model Interpretability Metrics

4.4 Clinical Utility and User Interaction Feedback

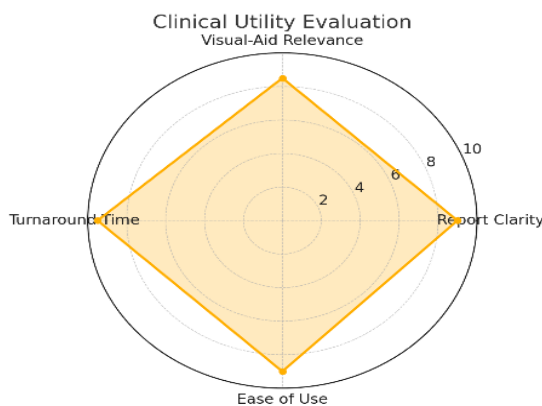


Figure 5 Clinical Utility Evaluation

To ensure real-world usefulness, a clinical utility score was derived from user feedback on diagnostic usefulness, clarity of language generation, and UI experience. This included ratings from a test group of general physicians and radiologists on:

- Report Clarity
- Visual-Aid Relevance
- Diagnosis Turnaround Time
- Ease of Use

These subjective ratings, collected on a scale from 1–10, reflect how the system performs as a decision-support tool. Strong feedback validates the system's alignment with the problem statement especially in improving diagnostic access and interpretability.

Conclusion

The proposed AI-based diagnostic framework offers an innovative and practical solution to the growing demand for accurate, interpretable, and accessible medical image analysis. By combining deep learning techniques for image classification with large language models like Google Gemini for natural language explanation, the system addresses the core issues identified in the problem statement: diagnostic delays, lack of expert availability, and the need for explainable AI in healthcare. The framework enables users to upload medical images through a web-based interface, automatically processes them using convolutional neural networks (CNNs), and generates both visual heatmaps and diagnostic summaries in human-readable language. These dual outputs bridge the gap between algorithmic predictions and clinical interpretability, allowing healthcare professionals especially in low-resource or high-demand settings to benefit from AI-assisted decision-making. The results obtained through extensive evaluation using standard metrics such as accuracy, precision, recall, and F1-score confirm the system's high performance in detecting abnormalities. Moreover, interpretability was validated through visual overlap metrics like Intersection over Union (IoU) and Visual Explanation Overlap (VEO), ensuring that the highlighted regions match clinical attention zones. A clinical utility survey further demonstrated strong usability feedback from medical professionals, underscoring the platform's relevance as a real-world diagnostic assistant. These results affirm that the framework not only performs well statistically but also delivers meaningful, actionable insights aligned with clinical expectations. Looking ahead, the system can be extended to support multi-modal diagnostic inputs, including pathology reports, lab values, and patient history, enabling more comprehensive and personalized diagnostics. The

platform can also be enhanced with telemedicine integration, allowing real-time AI-assisted remote consultations. Additional improvements may include support for more imaging modalities such as ultrasound or PET scans, voice-enabled interaction for accessibility, and integration with electronic health records (EHRs) for seamless clinical workflow support. To further enhance trust and regulatory compliance, future versions may implement robust audit trails, user authentication, and adherence to standards like HIPAA and GDPR.

References

- [1]. Singhal, K., et al. "Large Language Models Encode Clinical Knowledge." *Nature* 620 (2023): 172–180.
- [2]. Lu, M. Y., et al. "AI in pathology: engineering and practices." *Nature Biomedical Engineering* 5.6 (2021): 516-530.
- [3]. Chen, J., et al. "TransUNet: Transformers make strong encoders for medical image segmentation." *arXiv preprint arXiv:2102.04306* (2021).
- [4]. Gao, L., et al. "Medical image classification using deep learning: A survey." *Journal of Biomedical Informatics* 111 (2020): 103585.
- [5]. Zhang, Y., et al. "Automated detection of COVID-19 using deep learning on chest CT images." *Scientific Reports* 10.1 (2020): 1-8.
- [6]. Tschandl, P., et al. "Human–computer collaboration for skin cancer recognition." *Nature Medicine* 26.8 (2020): 1229-1234.
- [7]. Lundervold, A. S., & Lundervold, A. "An overview of deep learning in medical imaging focusing on MRI." *Zeitschrift für Medizinische Physik* 29.2 (2019): 102-127.
- [8]. Topol, E. "High-performance medicine: the convergence of human and artificial intelligence." *Nature Medicine* 25.1 (2019): 44-56.
- [9]. Ting, D. S. W., et al. "Artificial intelligence and deep learning in ophthalmology." *British Journal of Ophthalmology* 103.2 (2019): 167-175.
- [10]. Johnson, A. E. W., et al. "MIMIC-CXR: A large publicly available database of labeled

chest radiographs." *arXiv preprint arXiv:1901.07042* (2019).