

Integrating Visual XAI and LLMs for Interpretable Medical Image Analysis

Xin Hui Lor* and Chern Hong Lim†

Monash University, Malaysia

* E-mail: xlor0001@student.monash.edu

† E-mail: lim.chernhong@monash.edu

Abstract—Despite the growing adoption of deep learning in medical imaging, the limited transparency of model predictions continues to hinder clinical trust and deployment. This work proposes a multimodal explainability framework for medical image captioning that integrates both visual and textual explanations to enhance transparency and interpretability. Using a convolutional neural network trained on the ChestMNIST dataset, our model achieves a binary classification accuracy of 0.9492 across multiple thoracic conditions. We compare three visual XAI methods (SHAP, Saliency, and LIME) and four large language models (Gemini, Ollama, DeepSeek, and Skylark) to generate diagnosis-informed captions. Visual explanations highlight salient image regions influencing predictions, while large language models produce human-readable descriptions grounded in these visual cues. Our findings show that SHAP combined with DeepSeek offers the most balanced performance in terms of explanation fidelity and accessibility. Readability analysis indicates that the captions are suitable for a general audience with college-level proficiency, requiring no medical background. By bridging the gap between model decisions and user understanding, our approach improves the transparency and trustworthiness of AI-driven medical diagnosis systems.

I. INTRODUCTION

The integration of machine learning (ML) and deep learning (DL) into the medical domain has significantly advanced medical imaging analysis [1], [2]. Despite their promising capabilities in identifying complex patterns and correlations within large dataset [2], they often function as “black-boxes” [3]. Hence, the adoption of such systems in high-stakes clinical environments remains limited due to concerns regarding their interpretability and trustworthiness.

The scale of medical imaging further underscores the importance of interpretable AI systems. For example, approximately 46.6 million imaging tests were conducted in the UK between April 2023 and March 2024, with chest X-rays being the most commonly used modality [4]. As clinicians increasingly rely on AI systems to process and interpret this growing volume of data, the need for explainable artificial intelligence (XAI) has become more urgent. XAI seeks to transform opaque models into transparent “white-box” systems by making their predictions understandable to stakeholders [5].

Although XAI methods have been applied to medical imaging, they often require specialized expertise to interpret, offering limited contextual or semantic clarity for clinical decision-making and patient communication. As a result, their practical utility in real-world medical workflows remains limited.

To address these challenges, we propose a multimodal explainability framework that combines accurate multilabel classification of chest X-rays with interpretable outputs in both visual and textual forms. Specifically, the framework integrates visual saliency maps with large language model (LLM) generated explanations to enhance transparency and support a wider range of users.

We center our study on the following research questions:

- 1) Which XAI methods are effective for localizing diagnostically relevant features in medical image classification tasks?
- 2) How readable and semantically coherent are LLM-generated explanations of medical image captions for non-specialists?
- 3) Can the integration of visual saliency and language-based explanations form a unified framework enhance the interpretability of DL models for medical image classification?

To answer these questions, we proposed the following:

- 1) We design a multimodal explainability framework that integrates visual saliency and natural language explanation to enhance the interpretability of medical image classification model.
- 2) Comprehensive experiments is conducted to investigate the integration of XAI techniques into LLM models for generating context-aware, layperson-understandable explanations of deep learning outputs in medical imaging.

The remainder of this paper is organized as follows: Section II surveys prior work on XAI, LLMs, and their integration. Section III introduces the proposed framework, detailing the model architecture and multimodal explanation strategy. Section IV describes the experimental setup and reports results, with a focus on evaluating explanation quality. Section V analyzes the findings in comparison with state-of-the-art approaches and discusses key limitations. Finally, Section VI concludes the paper by addressing the research questions.

II. RELATED WORKS

A. Explainable AI (XAI) in Medical Imaging

Visual explanation techniques are the most prevalent XAI methods used in medical imaging [5]. Class Activation Mapping (CAM) is a foundational method that highlights class-discriminative regions by leveraging spatial information in convolutional feature maps, which is effectively used in medical

context. For instance, [6] applied it to ResNet-based model for histopathological feature localization, while [7] utilized it on 3D CNNs to identify Parkinson’s disease markers in the substantia nigra.

Grad-CAM [8] generalized CAM to a broader range of CNN architectures without requiring architectural modifications. Guided Grad-CAM, when combined with VGG16, has further enabled the identification of histomorphological patterns relevant to disease classification [9].

SHAP offers a theoretically grounded, model-agnostic framework based on cooperative game theory to quantify each feature’s contribution. Its deep learning extension, DeepSHAP, integrates DeepLIFT with Shapley values for enhanced interpretability, though at a higher computational cost [10].

Perturbation-based methods explore input-output sensitivity by modifying inputs and observing resultant outputs. Techniques like occlusion sensitivity and LIME are widely used. The Prediction Difference Analysis (PDA) method assess regional importance by simulating feature absence. [11] proposed multi-resolution PDA variants to offer fine-grained and hierarchical insights into model decision-making. [12] introduced a hybrid method combining DenseNet-121, MULAN, and Feature Pyramid Networks to refine feature resolution and improve anatomical localization.

B. Large Language Models (LLMs)

LLMs are broadly categorized as encoder-only, decoder-only and encoder-decoder architectures [13]. Decoder-only models such as ChatGPT have gained significant traction, with approximately 18.1% of studies applying them to medical or healthcare tasks [13]. However, general-domain LLMs often lack nuanced contextual understanding, especially in rare or complex-clinical cases, which may result in shallow or misleading outputs [13].

To address domain-specific shortcomings, several specialized medical LLMs have been developed. HEAL [14], a 13-billion parameter model, was trained to generate physician-validated clinical notes from doctor-patient dialogues. [15] uses GPT-4 Vision Preview as base model to support structured multimodal interactions, integrating domain-specific prompts for enhanced clinical reasoning. [16] further improved model reasoning by fine-tuning ChatGLM2-6B using QLoRA for instruction tuning.

C. Integration of XAI with LLMs

Efforts to integrate XAI techniques with LLMs aim to enhance the interpretability of multimodal AI systems. [17] demonstrated that combining SHAP with GPT-4 can yield context-aware and semantically rich explanations. [18] suggested two LLM-XAI integration pipelines for disease diagnosis: one decoupling ML model and LLM with post-hoc LIME explanations, and another employing LIME before enriching outputs through LLM-generated narratives.

[19] introduced XpertAi, a hybrid framework that fuses traditional XAI methods like SHAP and LIME with LLMs including GPT and Claude, to contextualize chemical data

predictions using domain-specific knowledge. Similarly, [20] employed SHAP with LLaMa v3 to construct a generalizable mathematical framework for interference analysis in 6G networks. [21] proposed augmenting LLM outputs with structured domain-specific knowledge graphs, improving the interpretability of clinical decision support systems.

The integration of XAI with LLMs remains in its early stages but represents a promising and rapidly evolving research direction. While LLMs can enhance interpretability by generating natural language explanations, existing XAI methods often lack clarity and clinical relevance, particularly in medical imaging. Critically, there is no systematic evaluation of which XAI-LLM combinations are most effective. This gap motivates the proposed methodology, which aims to systematically explore and benchmark the integration of LLMs with XAI techniques in the medical context, as detailed in Section III.

III. METHODOLOGY

Figure 1 illustrates the overall architecture of our proposed system. The approach integrates XAI techniques with LLMs to generate transparent and accessible captions for medical images. Our framework is trained and evaluated on the ChestMNIST dataset [22], which comprises 112,120 grayscale chest X-ray images labeled across 14 thoracic diseases. Each image can exhibit multiple co-occurring conditions or none at all. The dataset is licensed under the CC BY 4.0 license [23].

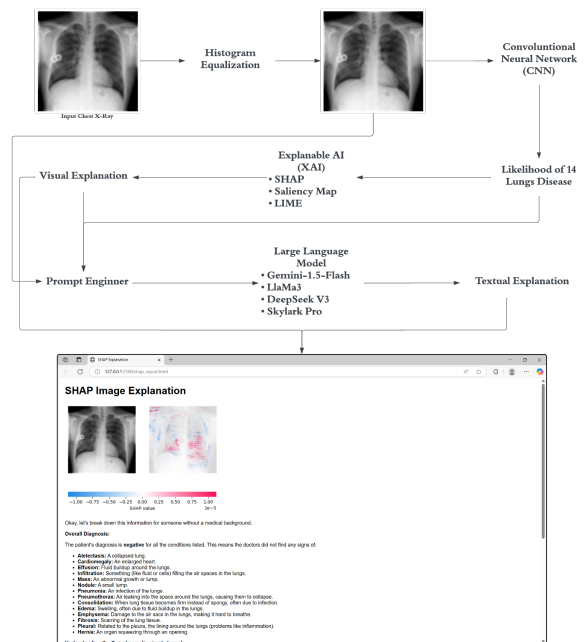


Fig. 1. Overview of proposed multimodal explainability framework

To perform disease classification, we developed a convolutional neural network (CNN) consisting of five convolutional blocks. The network accepts 128×128 grayscale images and outputs a 14-dimensional sigmoid-activated vector, each dimension indicating the predicted probability of a specific

disease being present. The model is trained using the Adam optimizer with a binary cross-entropy loss function, evaluated using binary accuracy. Training was conducted on 78,468 preprocessed samples over five epochs, with a batch size of 64 and a validation split of 0.3.

For post hoc interpretability, we employed three widely-used XAI methods: SHAP, Saliency Maps, and LIME. SHAP explanations were computed using a DeepExplainer instantiated with 200 randomly selected background images from the training set, enabling pixel-wise attribution across all disease classes. Saliency maps were generated by computing the gradient of each class output with respect to the final convolutional layer, thereby highlighting image regions with maximal impact on the classification. LIME was applied by generating 1,000 perturbed samples per instance to construct a local surrogate model, binary masks were then overlaid to indicate regions most influential to the model’s predictions [24].

Following visual explanation, we designed a structured prompt to guide the LLMs in caption generation. Each prompt enumerates all 14 diagnostic labels, explicitly marking unpredicted diseases as “not” present to prevent false inferences of a healthy diagnosis. The prompt includes the original chest X-ray alongside visual artifacts from SHAP, saliency, and LIME. The chest X-ray and visual explanation are resized to 64×64 or 28×28 depending on model context limits. Additionally, each prompt instructs the model to “keep the explanation clear and non-technical for a general audience”.

To evaluate model robustness, we employed four high-performance LLMs: Gemini 1.5 Flash, LLaMa 3, DeepSeek v3, and Skylark Pro. Gemini 1.5 Flash supports multi-million token context windows and efficient token streaming [25]; LLaMa 3 is trained on over 15 trillion multilingual tokens [26]; DeepSeek v3 uses 14.8 trillion tokens for general-purpose reasoning [27]; and Skylark Pro, developed by BytePlus, is optimized for reliable contextual response generation in high-stakes domains [28]. Each model generates a natural language explanation based on both the diagnostic output and the accompanying visual rationale.

The final output is a synchronized visual-textual interface that pairs saliency-based visualizations with LLM-generated explanatory captions. This multi-modal representation enhances transparency by grounding textual outputs in interpretable visual evidence, aiming to support both clinical expert interpretation and non-expert understanding in alignment with the goals of explainable AI in healthcare.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

The primary objective of our experiments is to evaluate the interpretability and utility of combining different XAI techniques with LLMs for generating accurate and accessible image captions, particularly for non-expert users in medical contexts. We investigate the explanatory capacity of twelve combinations, derived by pairing each of the three XAI

methods: SHAP, Saliency Maps, and LIME, with four LLMs: Gemini 1.5 Flash, LLaMa 3, DeepSeek v3, and Skylark Pro.

A random sample of 150 chest X-ray images was selected from the ChestMNIST dataset for evaluation. For each image, the trained multi-label classification model produced disease predictions, which were then explained via the three XAI methods. These visual outputs were integrated into structured prompts and passed to each of the four LLMs, resulting in a total of 1,800 generated captions (150 images × 3 XAI × 4 LLMs).

B. Evaluation Metrics

Given the lack of ground truth, we focus on two key aspects for evaluation: readability and informativeness. Readability was quantified using standard metrics computed via the textstat Python library: Flesch Reading Ease Score (FRES), SMOG Index, Flesch-Kincaid Grade Level (FK Grade), Dale-Chall Readability Score, and difficult word count. FRES indicates overall ease of comprehension, with higher scores denoting simpler text, computed as proposed by [29]:

$$FRES = 206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

The SMOG Index estimates the number of years of formal education required to understand a text, and is widely used to evaluate the readability of health-related documents [30]:

$$SMOG = 1.0430 \sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291$$

The FK Grade similarly estimates the U.S. school grade level necessary for comprehension [31]:

$$FKGrade = 0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59$$

The Dale-Chall Readability Score evaluates lexical complexity based on a curated list of 3,000 common words understood by fourth-grade students, computed as proposed by [32]:

$$Dale - Chall = 0.1579 \left(\frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0.0496 \left(\frac{\text{words}}{\text{sentences}} \right)$$

, where “difficult words” are those not found in the Dale-Chall lexicon. In addition, the total count of such difficult words is reported to reflect the lexical burden imposed on readers. An interpretation of FRES score and Dale-Chall score is provided in Table I.

Informativeness was evaluated using word count and disease mention frequency. Word count captures the verbosity of the explanation and the extent of elaboration. Disease mention frequency quantifies how many of the 14 diagnostic terms from ChestMNIST were explicitly referenced in each caption. This metric assesses the degree to which the generated explanation aligns with the underlying model predictions and supports clinical traceability. Together, these metrics allow us to quantify both the accessibility and specificity of each caption.

TABLE I
INTERPRETATION OF READABILITY SCORE

FRES[29]	School Level	Readability
90 - 100	5th grade	Very easy
80 - 89	6th grade	Easy
70 - 79	7th grade	Fairly easy
60 - 69	8th and 9th grade	Plain English
50 - 59	10th to 12th grade (high school)	Fairly Difficult
30 - 49	college	Difficult
0 - 29	college graduate	Very difficult

Dale-Chall[32]	School Level
4.9 and below	4th grade and below
5.0 - 5.9	5th and 6th grade
6.0 - 6.9	7th and 8th grade
7.0 - 7.9	9th and 10th grade
8.0 - 8.9	11th and 12th grade
9.0 - 9.9	13th to 15th grade (college)
10.0 and above	16th grade and above (college graduate)

C. Results

a) *Qualitative Analysis:* Figure 2 presents the visual explanations generated by SHAP, Saliency, and LIME respectively, applied to a randomly selected chest X-ray image from the ChestMNIST dataset. All three methods highlight the lower right lung, indicating a consistent attribution of this region to the model prediction.

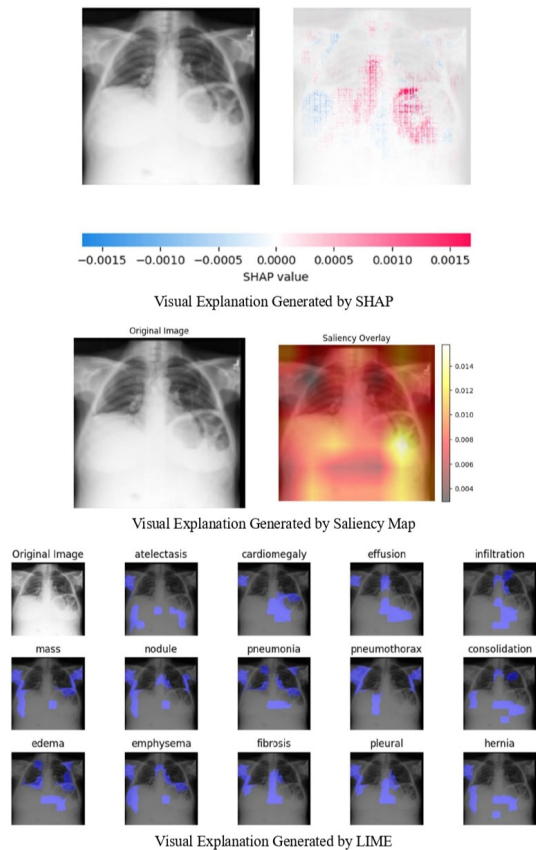


Fig. 2. Visual Explanation Generated by SHAP, Saliency and LIME (best view in color)

SHAP and Saliency both offer a global perspective by assigning pixel-wise importance scores, capturing general model behavior without tailoring the attribution to specific output classes. While this enables a high-level interpretation of influential image regions, it limits the granularity needed for understanding class-specific rationales, which is a crucial aspect in multi-label medical diagnosis tasks. In contrast, LIME provides class-specific insights by identifying which regions contribute to each predicted label. However, it lacks a quantifiable measure of regional importance, making it less interpretable in terms of relative influence across the image.

To further enhance interpretability, textual explanations were generated using four LLMs: Gemini, Ollama, DeepSeek, and Skylark. Each LLM was prompted under two distinct conditions: (1) without any visual explanation guidance, and (2) with visual explanations generated by SHAP, Saliency, or LIME, provided as auxiliary input in the form of NumPy arrays. Figure 3 presents a sample result for the same image used in the visual explanation. Under the first condition, LLMs generated generalized diagnostic narratives, describing typical radiological manifestations associated with each disease class. These descriptions tended to rely on textbook knowledge rather than the specific features of the input image.

XAI	DeepSeek Explanation
Without XAI	<p>Based on the lung X-ray image represented by the numpy array and the given diagnosis, here's an explanation in simple terms:</p> <ol style="list-style-type: none"> 1) Atelectasis: This means there is a partial collapse or deflation of the lung tissue, likely seen as a denser (whiter) area in the middle or lower parts of the lungs in the X-ray. This could be due to blocked airways or other issues preventing the lung from fully expanding. 2) Not Cardiomegaly: The heart size appears normal in the X-ray, with no enlargement. The central region (where the heart is) doesn't show unusual shadows or expansion. 3) Effusion: There is fluid buildup around the lungs (pleural effusion), which might appear as a cloudy or white area at the bottom of the lungs in the X-ray. This can make breathing difficult. 4) Not Infiltration / Not Mass / Not Nodule: No abnormal spots, growths, or dense areas (like tumors or infections) are seen in the lungs. The lung tissue looks mostly clear except for the atelectasis and effusion. 5) Not Pneumonia / Not Pneumothorax / Not Consolidation: No signs of lung infection (pneumonia), collapsed lung (pneumothorax), or fluid-filled lung tissue (consolidation). The darker areas (normal air-filled lungs) dominate most of the image. 6) Not Edema / Not Emphysema / Not Fibrosis: No fluid in the lung air sacs (edema), no damage to lung air sacs (emphysema), and no scarring (fibrosis). The lung texture appears smooth without excessive whiteness or cobweb-like patterns. 7) Not Pleural Thickening / Not Hernia: No thickening of the lung lining or abnormal protrusion of organs into the chest cavity. The edges of the lungs and diaphragm look normal. <p>Relating to the Image:</p> <ul style="list-style-type: none"> Atelectasis: Likely visible as a patchy white area in the middle or lower lung regions (e.g., around the values 80-150 in the array, indicating denser tissue). Effusion: The bottom corners of the lungs (e.g., high values like 200+ in the lower rows) may show whiteness due to fluid. Normal Heart: Central values (middle columns) are not excessively bright, suggesting no heart enlargement. <p>In summary, the X-ray shows a partially collapsed lung (atelectasis) and fluid around the lungs (effusion), but no other major abnormalities like infections, tumors, or heart issues. The rest of the lungs appear healthy.</p>
SHAP	<p>Based on the diagnosis list and the SHAP values provided, here's a simple explanation of the findings in the lung X-ray:</p> <ol style="list-style-type: none"> 1. Atelectasis (Present) <ul style="list-style-type: none"> What it is: Partial or complete collapse of a lung or part of a lung. Why it's detected: The X-ray shows areas with lower density (darker regions), particularly in the lower lung zones, which suggests collapsed air sacs. Possible causes: Blocked airways (e.g., mucus plug), pressure from fluid, or weak breathing after surgery. Effects: Can cause shortness of breath and reduced oxygen levels. 2. Effusion (Present) <ul style="list-style-type: none"> What it is: Fluid buildup around the lungs (pleural effusion). Why it's detected: The X-ray shows white, cloudy areas at the lung bases, indicating fluid accumulation. Possible causes: Infection, heart failure, or inflammation. Effects: Can compress the lungs, making breathing difficult. 3. No Cardiomegaly (Heart is normal-sized) <ul style="list-style-type: none"> Why not detected: The heart shadow in the X-ray appears within normal size limits, and the SHAP values do not strongly indicate an enlarged heart. 4. No Infiltration, Mass, Nodule, Pneumonia, or Pneumothorax <ul style="list-style-type: none"> Why not detected: <ul style="list-style-type: none"> Infiltration: No cloudy patches suggesting infection or inflammation. Mass/Nodule: No abnormal round or dense spots in the lungs. Pneumonia: No widespread white patches indicating infection. Pneumothorax: No visible dark air pockets outside the lungs (no lung collapse due to air leakage). 5. No Other Significant Findings (Edema, Emphysema, Fibrosis, Hernia, etc.) <ul style="list-style-type: none"> Why not detected: <ul style="list-style-type: none"> Edema: No "puffy" white patterns suggesting fluid in lung tissue. Emphysema/Fibrosis: No signs of lung tissue damage or scarring. Hernia: No abnormal protrusions in the chest cavity. <p>Summary: The X-ray shows atelectasis (lung collapse) and pleural effusion (fluid around the lungs), but no signs of other major lung diseases. The dark areas suggest collapsed lung tissue, while the white patches at the bottom indicate fluid buildup. The heart appears normal, and there are no tumors, infections, or air leaks. If this were a real patient, doctors might investigate further to determine the cause of the fluid and lung collapse (e.g., infection, heart issues, or trauma). Would you like any part explained in more detail?</p>

Fig. 3. Textual Explanation Generated by DeepSeek With and Without XAI Integration

In contrast, when guided by visual explanations, the LLMs produced more diagnostic and image-grounded narratives. The reports typically began with an overall assessment aligned with the model's prediction, followed by detailed references to image regions highlighted by the corresponding XAI method. The models also occasionally reasoned about absent conditions,

citing both the visual absence of expected features and the lack of attribution in the XAI maps. Additionally, the generated explanations often included potential causes and effects of the identified conditions, offering a more comprehensive and clinically meaningful interpretation.

b) *Quantitative Analysis*: The generated textual explanations were evaluated on readability and informativeness using the metrics in Section IV-B, with average scores for each XAI-LLM pairing shown in Table II.

Readability scores exhibit notable variability between systems. FRES values range from 28 to 68, corresponding to the levels of textual complexity from 'very difficult' to 'plain English'. SMOG indices span from 10 to 16, while FK Grade levels range from 7 to 15, mapping approximately to U.S. 7th grade through college-level reading proficiency. Dale-Chall scores are between 9 and 11, further indicating a baseline requirement of high school to early college-level literacy. While these results imply the need for general reading proficiency, the explanation for domain-specific terminology ensures that medical knowledge is not a prerequisite for comprehension, improving accessibility for non-expert users.

The informativeness metrics, disease mentions and disease mention ratio, generally appear low due to the nature of LLM generated text, which often places the disease name in a section heading for readability rather than repeating it throughout the explanatory paragraph.

TABLE II
AVERAGE VALUES OF READABILITY AND INFORMATIVENESS METRICS
ACROSS ALL SYSTEMS (BEST VIEW IN COLOUR)

LLM	XAI	FRES (↑)	SMOG (↓)	FK Grade (↓)	Dale- Chall (↓)
Gemini	SHAP	40.835	14.554	12.509	9.414
Gemini	Saliency	28.073	15.741	14.781	11.045
Gemini	LIME	40.975	14.009	12.869	10.017
Ollama	SHAP	35.207	14.761	13.275	9.180
Ollama	Saliency	34.343	15.195	13.405	9.800
Ollama	LIME	35.636	15.055	13.155	9.257
DeepSeek	SHAP	47.810	12.867	10.579	10.178
DeepSeek	Saliency	46.483	12.495	10.311	10.325
DeepSeek	LIME	44.570	13.007	10.921	10.071
Skylark	SHAP	63.895	10.634	8.250	8.804
Skylark	Saliency	67.659	10.237	7.378	8.779
Skylark	LIME	61.078	11.275	8.776	9.017
LLM	XAI	Difficult Words (↓)	Total Words (↑)	Disease Men- tions (↑)	Disease Men- tions Ratio (↑)
Gemini	SHAP	87.753	2421.153	20.107	0.008
Gemini	Saliency	45.200	961.433	11.007	0.013
Gemini	LIME	47.727	1224.387	11.353	0.010
Ollama	SHAP	77.993	2147.360	12.033	0.005
Ollama	Saliency	70.873	1678.840	5.047	0.003
Ollama	LIME	71.907	1957.073	12.813	0.006
DeepSeek	SHAP	68.967	1691.513	14.567	0.008
DeepSeek	Saliency	72.513	1791.987	10.153	0.006
DeepSeek	LIME	72.260	1779.287	10.920	0.006
Skylark	SHAP	46.127	1576.227	8.247	0.005
Skylark	Saliency	45.213	1554.787	6.167	0.004
Skylark	LIME	48.620	1570.933	8.113	0.005

Note: Green cells show the most favorable values; red cells show the least favorable.

Among the twelve XAI-LLM combinations, Skylark and Saliency ranked first across all readability metrics and effectively minimized difficult word usage, but performed poorly

on informativeness metrics. In contrast, Gemini and SHAP achieved top performance in informativeness but consistently ranked among the lowest in readability.

The DeepSeek and SHAP configuration demonstrated the most balanced performance, ranking within the top 50% in all evaluated dimensions, indicating a strong trade-off between clarity and content density. These results suggest that the DeepSeek and SHAP combination offers a promising balance, effectively improving transparency and accessibility in medical image captioning through integrated visual and textual explanations.

V. DISCUSSION

On the MedMNIST benchmark, ChestMNIST classification accuracy ranges from 0.779 to 0.948[23]. Our CNN model achieved 0.9492, slightly surpassing existing benchmarked methods. Beyond accuracy, we evaluated explanation effectiveness through readability and informativeness. Our findings suggest the most effective configuration combined SHAP for visual attribution with DeepSeek for textual explanation. SHAP highlights class-relevant regions by quantifying local feature importance, while DeepSeek generates layperson-friendly descriptions. Readability assessments indicate that captions generated typically require a college-level reading proficiency (approximately 13 years of formal education), but medical or technical knowledge is not required.

Despite these promising results, several limitations were identified in this study. Most XAI methods are post hoc, preventing integration into the training process and thereby limiting their applicability in end-to-end learning frameworks. Additionally, LLMs often produce templated opening phrases; for example, DeepSeek frequently begins with "Based on the provided diagnosis list, the patient...". This is likely a byproduct of prompt standardization, which, while promoting consistency and output quality [16], may constrain the diversity and nuance of generated explanations. Hallucination in LLMs also remains a concern, as they may produce inaccurate information that could mislead users.

Technical challenges also arise in embedding visual explanations and image features into LLM prompts. Inputs must be downsampled to meet token limitations, potentially omitting subtle but clinically significant details. Furthermore, different LLMs impose varying token limits, necessitating distinct compression strategies and leading to inconsistencies in the granularity of information conveyed. Moreover, LLM performance can fluctuate under various conditions, such as during high query volumes, with outputs exhibiting inconsistency across repeated runs. This unpredictability can undermine reliability and pose significant challenges for deployment in real-world clinical settings.

VI. CONCLUSION

This study introduces a novel pipeline that integrates XAI techniques with LLMs to improve the interpretability of ML-based medical image diagnosis. While visual methods expose

the “where” of model attention, textual explanations augmented by XAI articulate the “why”, bridging the gap between model predictions and clinically meaningful reasoning.

In addressing the research questions, SHAP, Saliency, and LIME were effective in identifying key diagnostic regions. SHAP and Saliency offer global visualizations with quantified importance scores, while LIME provides class-specific explanations without importance quantification, highlighting complementary strengths. Captions generated by LLMs, when paired with these XAI outputs, were generally readable at the high school to college graduate level. Importantly, they required no prior medical or technical expertise, aligning with our goal of improving accessibility for general audiences.

The integration of visual and textual explanations within a unified framework significantly enhances model interpretability. The generated narratives not only describe model predictions but contextualize them using visual attributions and medically relevant information. This multimodal approach facilitates patient understanding of diagnostic outcomes while supporting clinicians in rapidly identifying critical image regions, thereby contributing to more transparent and trustworthy AI-assisted medical diagnostics.

ACKNOWLEDGEMENTS

The authors would like to thank ByteDance for providing access to their Skylark LLM.

REFERENCES

- [1] M. Shehab, L. Abualigah, Q. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. I. Alsalibi, and A. H. Gandomi, “Machine learning in medical applications: A review of state-of-the-art methods,” *Computers in Biology and Medicine*, vol. 145, p. 105458, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482522002505>
- [2] I. D. Apostolopoulos and P. P. Groumpos, “Fuzzy cognitive maps: their role in explainable artificial intelligence,” *Applied Sciences*, vol. 13, no. 6, p. 3412, 2023.
- [3] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, “Explainable ai in medical imaging: An overview for clinical practitioners—beyond saliency-based xai approaches,” *European journal of radiology*, vol. 162, p. 110786, 2023.
- [4] N. England and N. Improvement, “Diagnostic imaging dataset statistical release,” *London: Department of Health*, Jul 2024.
- [5] B. H. Van der Velden, H. J. Kuijff, K. G. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (xai) in deep learning-based medical image analysis,” *Medical Image Analysis*, vol. 79, p. 102470, 2022.
- [6] A.-C. Woerl, M. Eckstein, J. Geiger, D. C. Wagner, T. Daher, P. Stenzel, A. Fernandez, A. Hartmann, M. Wand, W. Roth *et al.*, “Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides,” *European urology*, vol. 78, no. 2, pp. 256–264, 2020.
- [7] S. Chakraborty, S. Aich, and H.-C. Kim, “Detection of parkinson’s disease from 3t t1 weighted mri scans using 3d convolutional neural network,” *Diagnostics*, vol. 10, no. 6, p. 402, 2020.
- [8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: visual explanations from deep networks via gradient-based localization,” *International journal of computer vision*, vol. 128, pp. 336–359, 2020.
- [9] J. Kubach, A. Muhlechner-Fahrngruber, F. Soylemezoglu, H. Miyata, P. Niehusmann, M. Honavar, F. Rogerio, S.-H. Kim, E. Aronica, R. Garbelli *et al.*, “Same same but different: A web-based deep learning application revealed classifying features for the histopathologic distinction of cortical malformations,” *Epilepsia*, vol. 61, no. 3, pp. 421–432, 2020.
- [10] S. Walia, K. Kumar, S. Agarwal, and H. Kim, “Using xai for deep learning-based image manipulation detection with shapley additive explanation,” *Symmetry*, vol. 14, no. 8, p. 1611, 2022.
- [11] D. Seo, K. Oh, and I.-S. Oh, “Regional multi-scale approach for visually pleasing explanations of deep neural networks,” *IEEE Access*, vol. 8, pp. 8572–8582, 2019.
- [12] K. Sahatova and K. Balabaeva, “An overview and comparison of xai methods for object detection in computer tomography,” *Procedia Computer Science*, vol. 212, pp. 209–219, 2022.
- [13] S. Shool, S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, and M. Tara, “A systematic review of large language model (llm) evaluations in clinical medicine,” *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, p. 117, 2025.
- [14] D. Yuan, E. Rastogi, G. Naik, S. P. Rajagopal, S. Goyal, F. Zhao, B. Chintagunta, and J. Ward, “A continued pretrained llm approach for automatic medical note generation,” *arXiv preprint arXiv:2403.09057*, 2024.
- [15] D. P. Panagoulas, M. Virvou, and G. A. Tsihrintzis, “Evaluating llm-generated multimodal diagnosis from medical images and symptom analysis,” *arXiv preprint arXiv:2402.01730*, 2024.
- [16] C. Wu, Z. Lin, W. Fang, and Y. Huang, “A medical diagnostic assistant based on llm,” in *China Health Information Processing Conference*. Springer, 2023, pp. 135–147.
- [17] A. Zytek, S. Pidò, and K. Veeramachaneni, “Llms for xai: Future directions for explaining explanations,” *arXiv.org*, 2024.
- [18] K. Attai, M. Ekpenyong, C. Amannah, D. Asuquo, P. Ajuga, O. Obot, E. Johnson, A. John, O. Maduka, C. Akwaowo, and F.-M. Uzoka, “Enhancing the interpretability of malaria and typhoid diagnosis with explainable ai and large language models,” *Tropical medicine and infectious disease*, vol. 9, no. 9, pp. 1–23, 2024.
- [19] G. P. Wellawatte and P. Schwaller, “Extracting human interpretable structure-property relationships in chemistry using xai and large language models,” *arXiv.org*, 2023.
- [20] H. A. Tahir, W. Alayed, W. U. Hassan, and A. Haider, “Proposed explainable interference control technique in 6g networks using large language models (llms),” *Electronics (Basel)*, vol. 13, no. 22, pp. 4375–, 2024.
- [21] C. Park, H. Lee, S. Lee, and O. Jeong, “Synergistic joint model of knowledge graph and llm for enhancing xai-based clinical decision support systems,” *Mathematics (Basel)*, vol. 13, no. 6, pp. 949–, 2025.
- [22] X. Wang, Y. Peng *et al.*, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *CVPR*, 2017, pp. 3462–3471.
- [23] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, “Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification,” *Scientific Data*, vol. 10, no. 1, p. 41, 2023.
- [24] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should I trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.
- [25] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” 2024.
- [26] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, “The llama 3 herd of models,” 2024.
- [27] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, “Deepseek-v3 technical report,” 2024.
- [28] “Model releases—modelark,” Jan 2025. [Online]. Available: <https://docs.byteplus.com/en/docs/ModelArk/1159178>
- [29] R. Flesch, “How to write plain english,” *University of Canterbury*. Available at http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml. [Retrieved 5 February 2016], 1979.
- [30] A. S. Hedman, “Using the smog formula to revise a health-related document,” *American Journal of Health Education*, vol. 39, no. 1, pp. 61–64, 2008.
- [31] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, “Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel,” Institute for Simulation and Training, University of Central Florida, Tech. Rep., 1975.
- [32] E. Dale and J. S. Chall, “A formula for predicting readability: Instructions,” *Educational research bulletin*, pp. 37–54, 1948.