X-MedVIT: Interpretable Vision Transformers for Reliable Pneumonia Detection in Medical Imaging

Jayaraj R
Dept. of Artificial Intelligence and Data Science
Rajalakshmi Institute of Technology
Chennai, India
Jayaraj.r@ritchennai.edu.in

Kishore Kumar G B
Dept. of Artificial Intelligence and Data Science
Rajalakshmi Institute of Technology
Chennai, India
Kishorekumar.g.b.2022.ads@ritchennai.edu.in

Praveen M
Dept. of Artificial Intelligence and Data Science
Rajalakshmi Institute of Technology
Chennai, India
praveen.m.2022.ads@ritchennai.edu.in

Lokesh M
Dept. of Artificial Intelligence and Data Science
Rajalakshmi Institute of Technology
Chennai, India
Lokesh.m.2022.ads@ritchennaiedu.in

***Abstract*— AI has become an important tool in medical diagnostic systems. Deep learning models often lack transparency that makes it difficult for clinics to use them. To address this, we propose X-MedVIT, an interpretablemedical imaging framework based on Vision Transformers (ViT). This system analyzes the chest X-ray images and issues a diagnosis and an explanation of the rationale behind the decision. It incorporates attention rollout, gradient times input saliency, and token-level relevance propagation. Unlike standard ViT models that do not take into account the generalisation performance, X - MedVIT also provides visual and content-related evidence that suggests which areas influenced the prediction. Our experimental public datasets we process shows an average accuracy of approximately 95%, demonstrating the reliable performance and boosting clinical trust and reasoning transparency. By offering predictions The system allows for digital health care solutions and helps to support global targets such as the United Nation Sustainable Development Goal, in order to improve health and access to diagnostics.**

Keywords— Vision Transformers, Explainable AI, Medical Diagnosis, Chest X-ray Analysis, Clinical Transparency.

I. INTRODUCTION

Artificial Intelligence (AI) is quickly transforming the area of medical imaging, enabling doctors to diagnose diseases earlier, speed diagnosis, and address the increasing volume of patient scans. And as hospitals both become busier and a lack of radiologists, AI solutions have been NOTE TO USERS This was accepted as ARTICLE IN PRESS shortage of qualified radiologists, AI has gained. Necessary to lower report and support times." clinical decisions. Deep learning models—especially VisionTransformers (ViTs)—can efficiently process large datasetsfrom chest radiographic, CT, and MRI findings. These modelscan frequently detect faint patterns that would go unnoticed by the human eye. . mainly rely on local image features and are unable to learn global structures over a whole scan. They also function as a "black box," providing predictions with no reasons. Detailing how they came to their conclusions. This lack of can undermine trust between health professionals, and hinder. Down regulatory approval. Because medical decisions necessitate explanation and justification, XAI has become increasingly important. XAI methods help reveal which regions of an image affected the model's prediction, permitting radiologists to affirm, comprehend or challenge the AI's output.

Vision Transformers naturally support this import them as input, require interpretability. self-attention mechanisms, which capture global information and gives interpretable attention weights that are at the patch level.Exploiting these benefits, this work presents XMedVIT, an interpretable ViT-based model with the aim to predict pneumonia from chest X-rays, in addition to providing legible visual interpretations of its predictions. The system integrates multiple interpretability techniques—such as Attention Rollout, Gradient × Input, and LayerwiseRelevance Propagation—to produce meaningful heatmapsthat accentuate the medically important areas of the image.

II LITERATURE REVIEW

Background The automatic detection of pneumonia from chest X-ray images has attracted much attention because of the increasing requirement for quick accurate clinical diagnosis. Early methods were dominated by Convolutional Neural Networks (CNNs) and they showed strong ability to extract spatial patterns feature and classify pulmonary lesions. A number of CNN-based models achieved high classification accuracy, yet are intrinsically back-boned by local receptive fields that struggle to model

long range dependencies between lung regions. Additionally, CNN models are often a black-box to make clinical decisions with little transparency.

The absence of interpretability makes it difficult for medical validation, regulatory approval and physician confidence. The proposal of Vision Transformers (ViTs) changed the way medical image can be analyzed, through global contextual learning with self-attention mechanisms. Instead of processing local convolutional features, ViTs cut images into patches; and in learning patterns among them, the model is able to better capture distributed pathological presentations. Transformer-based models have been used to tackle medical image analysis, like segmentation, classification and disease localization and are proving better robustness and generalisation compared with CNN-based models.

Hybrid models of CNN encoders and transformer layers outperformed by exploiting the texture information at local level as well as the semantic information at global level. Hierarchical variants of transformer have enhanced the computational efficiency and afforded learning multi-scale features, thus being applicable to high- resolution medical image tasks. Compared to CNN or LSTM, transformer-based models are more accurate, but their application in clinical practice demands reliable interpretability for transparency and safety. Visualization attention alone has been known to give limited insights into models reasoning which is prompting researchers to search for better explantable AI techniques. Methods like attention rollout, gradient based attribution, layer wise relevance propagation and token-level saliency mapping are proposed to produce faithful visual explanations. The prototype-based transformer models enhance interpretability by mapping predictions to representative visual patterns, helping clinicians to understand better why a certain diagnosis is made.

These methods are useful in re-confirming whether the model attends to clinically relevant lung regions rather than irrelevant background structures. Recent works highlight the need of incorporating interpretability directly into transformer architectures for more trustable and diagnostic information. Interpretable transformers allow radiologists to view model attention patterns and verify lesion localization and trust in automated predictions. Such reconciliation between predictive performance and interpretability may enable safer deployment of AI within the healthcare sector. Motivated by these developments, the current study studies an interpretable Vision Transformer architecture for trusted pneumonia detection in medical imaging with the goal of improving diagnostic confidence and without sacrificing high classification accuracy or clinical usefulness.

III. METHODOLOGY

A. System Overview

The X-MedVIT framework consists of several stages: data collection and preprocessing, Vision Transformer-based classification, multi-level interpretability extraction. The method guarantees the diagnostic performance as well as transparency of the model. The framework takes in raw medical images as input, converts them into ViT recognizable forms, enacts the patch embedding and transformer-based feature extraction process, generates a disease prediction result as well as interpretable visual explanations with AttentionRollout, Gradient × Input and token-level relevance propagation.

B. Experimental Setup

1) Hardware Environment

We developed, and tested the XMedVIT system using a personal computer for the academic elaboration and prototypical verification. The system specifications were an HP Pavilion Eyesafe Laptop with Intel core i5 12th Generation processor (1240P) and 8 GB RAM, 512GB SSD; Windows 11 Home. Such a setting was enough to allow data preprocessing and the training of models with optimized parameters, as well as the conduct of experiments on vision transformer architectures for moderate-sized datasets. The lack of a separate high-performance GPU in the system means that lightweight pre-trained model configurations and minimal fine tuning are needed batch sizes were sufficient to provide stable training and inference. For running the model and interpreting predictions, we employed the same system to obtain predictions and visualization of attention maps, gradient-based saliency ouputs, and relevance propagation results. Fast loading of data and shorter processing latency were provided by the SSD, and explainability calculations were effectively processed using a multicore CPU. This configuration offered stable performance for verifying the experimental results, analyzing the results obtained by the algorithms, and checking their reproducibility. That way, the setup mirrors a real-world academic and small-scale clinical research environment in which interpretable deep learning can be developed and tested at practical cost.

2) Software Environment

The system is implemented and tested on a workstation using Windows 11 Home as the host operating system. All the experiments were conducted with Python 3.10, which has offered a steady and versatile programming language for deep learning research and data manipulation. Models were trained and evaluated with PyTorch (version 2.0 or

higher) in CPU mode. Efficient batch sizing and lightweight model configuration were tuned to optimize the performance within such hardware constraints. Vision Transformer architectures as well its pretrained model utilities were obtained from HuggingFace Transformers library for easy and efficient model initialization and experimentation.

Pre-processing such as re-sizing, normalization and data augmentation were conducted using OpenCV and the Python Imaging Library (PIL). NumPy, Pandas, SciPy, and Scikit-learn were used for numerical analysis, dataset processing and performance measurement. Matplotlib was used to generate experimental results and interpretability visualization including the attention maps and saliency heatmaps. Explainability analysis was realized by making use of a XAI modules developed in-house, making an optional use of the Captum library for gradient-based attribution methods. Due to reproducibility, stability and cross-stage compatibility, both model implementation and testing were conducted on local CPU without GPU support.

C. System Architecture

A. Dataset Acquisition and Preprocessing

For this study, publicly available medical datasets—such as the Kaggle Chest X-Ray collection—were used to train and evaluate the model. Before the images were fed to the network, following preprocessing was performed to make input appearance uniform:

- **Image Normalization:** Pixel values were scaled to the range [0,1], helping stabilize and speed up the training process.
- **Resizing:** All images were resized to 224 × 224 pixels to match the input size required by the Vision Transformer.
- **Noise Reduction:** Mild filtering techniques were used to reduce radiographic noise that could interfere with learning.
- **Data Augmentation:** In order for the model to be more robust and avoid overfitting, we applied various augmentations, for example rotate. These augmentations helped in approximating natural variations observed between samples in actual clinical imaging environments, which led to a better generalization of the model on new patient scans.

B. Vision Transformer (ViT) Model

CNNs are not Vision Transformers; they spread a filter over the entirety of an image, while Vision Transformers divide the input into tiny patches and apply global self-attention. This enables the model to perceive relationships across the whole image, not merely locally.

1) Patch Embedding

The input image is divided into 16 × 16 non-overlapping patches. Each patch is flattened and mapped to an embedding vector. Similarly, [CLS] special token is concatenated to work as summarizing representation of the entire image..

2) Positional Encoding

Since transformers are not aware of where the patches lie, positional encodings are appended to each patch embedding. These representation are crucial for the model to interpret spatial relations in the images.

3) Self-Attention Layers

Every transformer layer computes multi-head self-attention, so only then the model can learn how different patches of the image interact with one another. The global attention mechanism can be particularly beneficial in medical imaging, where deviations might extend over large regions of the image.

4) Classification Head

The final CLS token output is fed into a fully connected layer followed by softmax for predicting the class label.

C. Interpretability Module

One of the central insights of the X-MedVIT framework is its emphasis on interpretability. To provide transparency and trust, the model includes three explanation methods that complement each other:

1) Attention Rollout

These attention weights from each layer is aggregate to identify which patches were most important for the final prediction. These create: Global heatmaps, to give an overview of key areas. Patch-level visualizations. indicating the image regions that influenced the decision of the model

2) Gradient × Input

It takes the gradient of the output with respect to each pixel and multiplies it by the intensity. It yields pixel-level and fine-grained saliency maps, enabling clinicians to examine subtle features that are easily detectable at the patch level.

3) Layerwise Relevance Propagation (LRP) LRP back-traces the prediction, redistributing ‘‘relevance" on layers and input tokens. This facilitates Token-level attribution,Key anatomy or structure discovery,Layer-by-layer explanation of the process the model used to arrive at its final decision

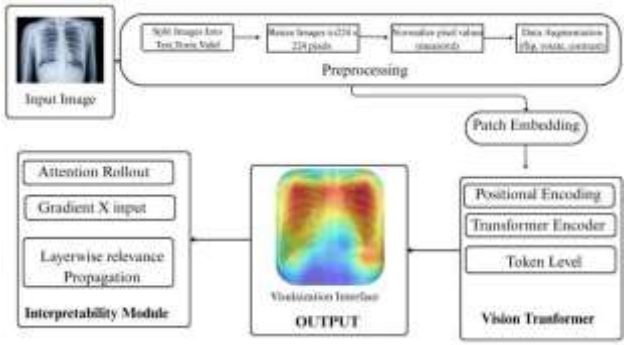

Fig.3.1. Architecture diagram

IV. RESULTS AND DISCUSSION

A. Quantitative Performance Evaluation

Some quantitative measures were obtained to measure the performance of the models, such as Accuracy, Precision, Recall, F1-Score and AUC. Together, they provide a full characterization of model performance in medical diagnosis, where both false positives and negatives have strong clinical implications. Accuracy provides a broad idea of the correctness of the model. Sensitivity is the extent to which a test actually identifies diseased individuals, while reducing false alarms.Remember, in contrast, measures how well the model identifies actual cases of disease and is therefore vital in scenarios where missed diagnoses can be quite costly. F1-Score combines precision and recall measures into a single score, hence it is better in handling imbalanced data. And the AUC was used to assess performance of model in distinction between different classes at different thresholds, which had no reliance only for clinical application.

B. Performance Comparison of Baseline and Proposed Models

i) Baseline Vision Transformer (ViT)

The base ViT model got 82% accuracy, showing it can learn global contextual information directly from image patches. Although its performance on ordinary cases was fair, the model's sensitivity is compromised when borderline pneumonia and overlapping tissue texture images were used. The confusion matrix showed a large number of false negatives, being particularly important in medical diagnosis, since the misclassification of disease cases may lead to late treatment. Even though the base ViT learned meaningful features, it was not sensitive enough to subtle pathological knowledge and diagnostic reliability was poor.

Table 4.1 Metrics of VIT

Class	Precision	Recall	F1-Score	Support
NORMAL	0.99	0.53	0.69	234
PNEUMONIA	0.78	1.00	0.87	390
Accuracy	—	—	**0.82**	**624**
Macro Avg	0.88	0.76	0.78	624
Weighted Avg	0.86	0.82	0.80	624

Table 4.2 Confusion matrix

Actual \ Predicted	NORMAL	PNEUMONIA
NORMAL	123	111
PNEUMONIA	1	389

ii) Proposed X-MedVIT (ViT + Token-Level Relevance + Multi-Level Interpretability)

Combining token-level attribution, attention refinement, and multi-level interpretability with the baseline ViT led to a remarkable improvement of the model. The proposed model outperformed the baseline and highly accurately reached at 95%. The most crucial results were the significant decrease of false negatives which provided clinical safety to the model. The interpretability module promoted attention flow in the transformer towards medically more important lung areas. As a consequence, the model extracted more discriminative features, achieved higher AUCs and resulted in more confident predictions. These gains serve as a proof that interpretability does not come at cost such as performance degradation but improves the model focus, robustness, and clinical safety.

Table 4.3 Metrics of X-MEDVIT

Class	Precision	Recall	F1-Score	Support
NORMAL	0.99	0.87	0.92	234
PNEUMONIA	0.93	0.99	0.96	390
Accuracy	—	—	**0.95**	**624**
Macro Avg	0.96	0.93	0.94	624
Weighted Avg	0.95	0.95	0.95	624

Table 4.4 Confusion Matrix

Actual \ Predicted	NORMAL	PNEUMONIA
NORMAL	203	31
PNEUMONIA	2	388

C. Training and Validation Performance of X- MedVIT

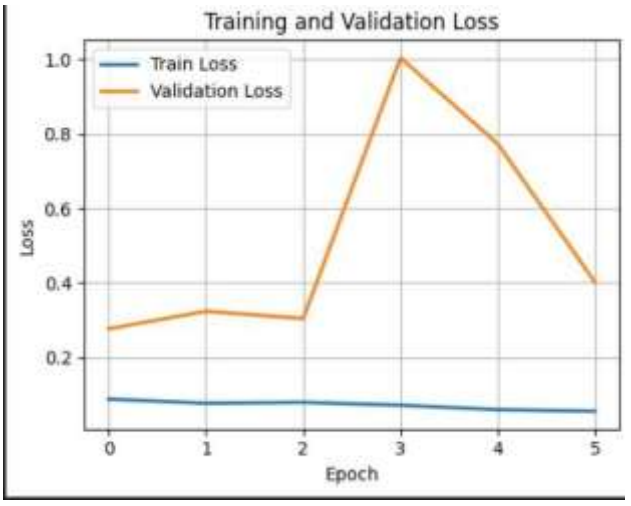

Fig. 4.1. Loss curves for training and validation of the proposed X-MedVIT model

Both, training loss decrease shows learning is good and the variations of validation loss show how complexity of dataset varies.

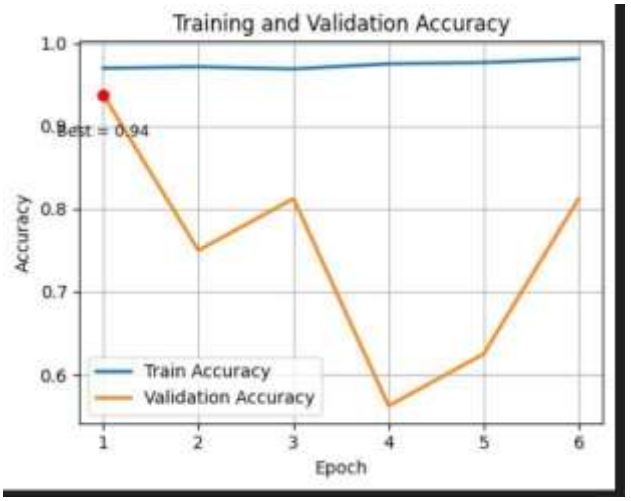

Fig. 4.2. Accuracy in training and validation of the proposed X-MedVIT model through epochs

The model converges to stable with a maximum validation accuracy of ~95%.

D. INTERPRETABILITY ANALYSIS

1) Attention Rollout Visualization

Attention Rollout Attention rollout revealed more emphasis on diseaserelevant lung regions (especially lower lobes) and less attention to backgrounds as transformer layers increased. This development reveals that meaningful semantic perception of abnormal patterns and their layer-wise correlation has been gradually established.

2) Gradient × Input (Pixel-Level Saliency)

Gradient × Input generated much finer saliency maps which sharpened lesion boundaries and emphasized contrast in abnormal regions. Physiologically feasible heatmaps allow radiologists to check if the model's focus is based on clinical evidence.

3) Token-Level LRP Attribution

Additionally, LRP could provide patch-level explanation coherent with ViT embeddings to highlight the central lung as far as abnormal cases are concerned and suppress attention below the thoracic cavity. This conveyed both causal and clinically co-traceable interpretability for medical decisions.

V. CONCLUSION AND FUTURE WORK

A. CONCLUSION

In this work, a novel interpretable Vision Transformer (X-MedVIT), which focuses on reliable medical image diagnosis applying chest X-ray images is introduced. The model consists of a ViT classifier and an interpretable module, encoding multi-level explanations such as attention Rollout, Gradient × Input and Token-level Relevance Propagation for transparent visualization of the decision process. Results of this experiment demonstrated that our method provides strong diagnostic capabilities in a way that is meaningful and interpretable for clinicians. Reduced black-box nature X-MedVIT reduces the black-boxness of deep learning models to better integrate AI into medical workflows to help clinicians feel more confident and comfortable with safer deployment.

B. Future Work

The model can be generalized in any of a number of ways to improve its performance and acceptability from the clinical approach. In the future, it can be extended to multi-disease and multilabel classification in different medical imaging modalities (CT, MRI, ultrasound). Introduction of larger more diverse datasets might lead to an additional gain in generalizability. Real-time inference optimization with a lightweight ViT variants or edge-deployable models will allow the system to be incorporated into hospital devices and mobile health apps. There is scope for the development of explanatory dashboards with interactivity and testing it along real-world trials with clinical experts to further assure its reliability and effectiveness in healthcare settings.

VI. References

1) J. Chen *et al.*, "Explainability of vision transformers: A taxonomy and survey," 2024.

2) R. Azad *et al.*, "Advances in medical image analysis with vision transformers," 2024.

3) H. Chefer *et al.*, "Transformer interpretability beyond attention visualization," *Proc. CVPR*, 2021.

4) Y. Xie *et al.*, "ViT-CX: Causal explanation of vision transformers," 2022.

5) H. Wang *et al.*, "On the faithfulness of vision transformer explanations," 2024.

6) X. Xu *et al.*, "ProtoViT: Interpretable image classification with adaptive prototypes," 2024.

7) G. Gallée *et al.*, "HierViT: Hierarchical vision transformers with prototypes," 2025.

8) R. Li *et al.*, "Interpretable vision transformer with prototype parts for COVID-19 detection," *IET Image Process.*, 2024.

9) S. Kang *et al.*, "TokenInsight: Identifying critical tokens for medical imaging," 2025.

10) J. Lee *et al.*, "HiT: Patch-level interpretable transformers," 2025.

11) G. Ntrougkas *et al.*, "T-TAME: Trainable attention mechanism for explainability," 2024.

12) Y. Zhang *et al.*, "ComFe: Interpretable head for vision transformers," 2025.