The Research of Medical Science Review
ISSN: 3007-1208 & 3007-1216 Volume 3, Issue 4, 2025

LAYER-WISE RELEVANCE PROPAGATION IN LARGE-SCALE NEURAL NETWORKS FOR MEDICAL DIAGNOSIS

Jawad Ahmad[1], Muhammad Ihsan Ur Rehman[2], Muhammad Shaheer ul Islam[3], Areeba Rashid[*4], Muhammad Zukhruf Khalid[5], Areesha Rashid[6]

[1]*MBBS Scholar (4th year), Sahiwal Medical College, Sahiwal, 57000, Punjab, Pakistan,*
[2] *MBBS, Sahiwal Medical College, Sahiwal, 57000, Punjab, Pakistan,*
[3]*MBBS Scholar (Final year), Faisalabad Medical University (FMU), Faisalabad, 38000, Punjab, Pakistan,*
[*4]*MBBS Scholar (Final year), D.G. Khan Medical College, Dera Ghazi Khan, 32200, Punjab, Pakistan,*
[5]*MBBS, Quetta Institute of Medical Sciences, 87300, National University of Medical Sciences (NUMS), Pakistan,*
[6]*M.Phil, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, 45320, Pakistan*

[1]jawadsharafat@gmail.com, [2]ahsanurrehman468@gmail.com, [3]shaheernasir2002@gmail.com,
[*4]areebar@dgkmc.edu.pk, [5]zukhrufswag1122@gmail.com, [6]areesha.rashid@bs.qau.edu.pk

[*4]https://orcid.org/0009-0002-3888-1125, [6]https://orcid.org/0009-0006-0401-2174

DOI: https://doi.org/10.5281/zenodo.15152153

Keywords
Artificial Intelligence (AI), Disease, Layer-wise Relevance Propagation (LRP), Medical diagnosis, Neural Networks, Vision Transformers (ViTs).

Article History
Received on 25 February 2025
Accepted on 25 March 2025
Published on 05 April 2025

Corresponding Author: *

Areeba Rashid
E-mail: areebar@dgkmc.edu.pk

Abstract

Large-scale neural networks have recently transformed medical diagnosis with exceptional accuracy across various imaging tasks with high accuracy and efficiency. It is true that as one relies on artificial intelligence (AI) for clinical settings, the necessity of interpretability and transparency becomes more and more critical. In this review, we focus on Layer-wise Relevance Propagation (LRP), a technique that enables us to enhanced interpretability of neural networks by identifying on what regions the model is relying the most to its decision. Additionally, it demonstrates neural networks in the medical field of radiology, pathology, cardiology, neurology, stating where advanced learning algorithms are utilized for such tasks as tumor detection, image segmentation, and disease classification, and also outlines their findings. LRP creates clinical trust and genuine collaboration between health care team and machine systems to resolve key transparency issues. Relevant current challenges, including scalability and computational demands, that must be addressed via further research are discussed, in order to further refine LRP for complex models and to integrate it into clinical workflows. Despite these challenges, LRP shows great promise in generating robust chains of divisions as clinical applications across all of these imaging modalities (X-ray, MRI, CT scan).

INTRODUCTION

In recent years, large-scale neural networks particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have revolutionized medical diagnosis by achieving remarkable accuracy across diverse imaging tasks. In medical imaging analysis, which has become the domain of CNNs and ViTs, tasks ranging from tuberculosis detection in chest X-rays (97.1% sensitivity with EfficientNet-B4) (Pordeli Shahreki et al., 2024) to load classifying Alzheimer's disease using Magnetic Resonance

Imaging (MRI) have reached state-of-the-art achievements. Due to their ability to identify and reduce the number of diagnostic errors and the associated resource burden (Jia et al., 2024), these models are very best at picking up and highlighting small patterns in large datasets. Despite high computational costs, reliance on carefully labeled datasets, and the inherent 'black box' decisions of how to intervene however, their deployment has not been widely attempted (Tulsani et al., 2023).

Need for Interpretability in Medical AI Systems

Machine learning in medical diagnosis enhances precision and decision-making by supporting clinical expertise. Given that surgeon skill and experience bear a direct relationship with the reduction in the risk of diseases (A. Rashid et al., 2024), AI-driven models can further aid in minimizing errors and improving patient outcomes by providing data-driven insights and real-time assistance.

It is important to be able to interpret the model decisions such that they are compatible with known pathological markers like LRP heatmaps showing correspondence with Alzheimer's brain regions that meet the standards for a transparent AI system for health applications (Lim et al., 2024). AI involves designing systems that mimic human intelligence (Shafiq et al., 2025), including explainability maps that help radiologists interpret and validate AI-generated results. This enhances trust in AI-assisted diagnoses, particularly in critical applications like COVID-19 detection, ensuring accuracy and reliability in clinical workflows. While current studies imply as high as 10.3% of AI/medical research' code is shareable on reproducible code, and less than 7% of AI/medical research implements robust interpretability frameworks, this critical gap is evident in clinical translation (Gupta et al., 2024).

Layer-Wise Relevance Propagation (LRP)

LRP is a post hoc explanation method to quantify each input feature contribution to model predictions by backward propagating relevance scores at the layers of a network. The key characteristics are that total relevance is preserved across layers so that the mathematical consistency is still preserved (Otsuki et al., 2024). Pixel level heatmaps of regions influences diagnoses like lung lesions in X-rays (Gupta et al., 2024), and hippocampal atrophy in Alzheimer's MRIs (Böhle et al., 2019). To make it workable for modern large scale networks, the techniques like 'relevance splitting' was extended to Residual Network (ResNet) and ViTs through skip connections (Otsuki et al., 2024). AI instruments produce better solutions to complex problems and the enhanced understanding of the concepts (Shafiq et al., 2025). In the end, LRP has the potential to significantly transition medical AI from being a less accurate, less reliable, less transparent diagnostic partner to a more reliable, more accurate, more transparent diagnostic partner.

Objective and Scope of the LRP

We have reviewed LRP's approach in bridging the interpretability gap for medical AI systems. It analyzes methodological improvements in LRP for transformer and CNN architectures (Ranjan et al., 2024). Challenges in obtaining these models include managing computational overhead, and domain specific validation in LRP clinical applications across imaging modalities (Xray, MRI, CT) are promising (Tulsani et al., 2023), but there remain questions in producing actual clinical and results the analysis focuses on practical implementations of LRP that can be applied to improve diagnostic transparency without trading off accuracy, such as being integrated into these high accuracy models (Gupta et al., 2024). Essential as neural networks are to medical diagnosis, however, their intermediary processes of making decisions are complex and therefore must be intelligible to clinical standards; methods such as LRP serve well to provide such interpretations. In this section, we analyze their architectural foundations and an enhancement of transparency provided by LRP.

Fundamentals of Neural Networks in Medical Diagnosis

Various neural network architectures have been employed in medical diagnosis, each offering distinct advantages in image analysis and disease detection. For histopathology analysis of breast cancer, CNNs are dominantly used with 96-99 % accuracy in breast cancer classification and for tuberculosis in detecting

tuberculosis from chest X-rays with 97.1 % sensitivity, CNNs provide hierarchical feature extraction to identify localized patterns like tumor margins or microcalcifications (Zeynali et al., 2024). The ViTs capture long range dependencies in images but fail to outperform CNNs in stroke lesion segmentation (nnU-Net model outperformed ViT based model on ISLES2022 dataset) (Zafari-Ghadim et al., 2024). Xception Transformer combines CNNs feature extraction in local and ViTs in global and reaches 91-100% accuracy in breast cancer classification (Zeynali et al., 2024). RNN Byproducts such as RATCHET and other Architectures that use Recurrent Neural Network (RNNs) along with CNNs to generate radiology reports are ways that sequential data processing helps clinical work flows (Hou et al., 2021).

Concept and Mechanism

Layer wise Relevance Propagation (LRP) is an algorithm proposed by Bach et al. to understand and explain the decision made by a complex machine learning model, especially neural network. Basically, it traces back those relevance scores as they come from the output layer to the input layer and what had most relevance in a given prediction. To this end, many of the machines learning models are "black boxes" that produce accurate predictions but fail to explain why the predictions were made. Using this approach, the goal of LRP is to explain classification decisions pixel wise such as produced as heatmaps (Bach et al., 2015). Conservation property is a key principle in LRP, since the relevance conserved as it propagates backward through the layers of the network. By recent developments, LRP has been extended to handle more complex architectures such as ViTs and ResNets, tailored techniques are required to handle skip connections and attention mechanisms (Achtibat et al., 2024). Commonly, these extensions involve the use of relevance splitting or modified propagation rules to maintain the conservation property and provide an accurate explanation (Ranjan et al., 2024). While LRP is beneficial, applying this technique to high resolution images and large models is computationally expensive and so they require a method like mixed precision quantization or a sift hybrid approach which incorporates faster techniques such as Grad-CAM to improve the efficiency. **Table 1** shows Comparative Analysis with Other XAI Methods highlighting medical imaging performance and their limitations.

Table 1: Comparative Analysis with Other XAI Methods

Method	Mechanism	Medical Imaging Performance	Limitations
LRP	Relevance redistribution	92% alignment with radiologist annotations in chest X-rays (Alam et al., 2023)	High memory overhead for 3D scans
Grad-CAM	Gradient-weighted pooling	Best quantitative scores in multi-label pathology prediction (Alam et al., 2023)	Fails with non-CNN architectures
SHAP	Perturbation analysis	Effective for feature importance ranking (Taiyeb Khosroshahi et al., 2025)	Computationally prohibitive for high-res images
LIME	Local surrogate models	Highest medical significance scores (Alam et al., 2023)	Instability across similar inputs

Although LRP is only one answer to much-needed medical AI, which addresses clinical needs in an anatomically grounded way, one to one must follow with the balance between computational costs and interpretative value. However, emerging as optimal solutions for deployment in the real world, hybrid methods combining Grad-CAM's efficiency (83% faster computation) with LRP's granularity are coming into development (Alam et al., 2023).

Applications of LRP in Medical Diagnosis

LRP has already been very useful in multiple domains of medical diagnosis towards improving model interpretability and clinical trust. Applications of LRP have been common in the task of interpreting deep learning models on chest X-ray, such as pneumonia detection. LRP heatmaps show important regions such as lung opacities in matching with the radiologist annotations. For instance, the

balance between accuracy (91%) and interpretability (Mean Relevance Score of 0.85) made such a ResNet50 based model a good candidate for clinical integration (Colin et al., 2025). LRP has been applied to explain predictions in mass like lesions and Alzheimer's disease diagnosis in brain lesion analysis. LRP centers its predictions on key brain regions (such as the hippocampus) or lesion sites such that sign predictions are consistent with well-known pathological markers (Ferles et al., 2023). More recent works shows the applicability of LRP in feature selection of EEG-based (Electroencephalogram) motor imagery classification and in the graph convolutional neural networks for deciphering BRCAness phenotype relating to cancer (Nam et al., 2023 and Yang et al., 2024). This application shows that LRP can provide with insights into the extreme neural network decisions in ways that make them more interpretable and reliable. Furthermore, LRP's ability to show contributions on the pixels enables it as a useful tool for proof by human experts once automated image classification systems have been proven to reason as they subconsciously should.

Pathology and Histopathology

LRP promotes the development of pixel-level heatmaps that indicate tumor regions, which can be very useful in histopathology where cancer is detected thanks to high resolution images. This has increased the transparency of CNN models in analyzing breast cancer such as by allowing pathologists to verify AI driven decisions (Abuhantash et al., 2024). For example, LRP has also been utilized in machine vision models to process mammographic images, to increase the diagnostic accuracy and provide interpretability of the model's reasoning (Manuela et al., 2024). Furthermore, LRP also aids in cancer classification, in gene expression analysis, where it reduces the complexity of the gene sets by pinpointing genes that make a significant contribution, at the condition of good accuracy. LRP is shown in these applications to be a promising alternative to current diagnostic methods for cancer detection (Sheng-Yi Hsu, 2024).

Disease Prediction and Risk Assessment Models

The risk scores that we derive from patient data have been explained by the LRP within disease prediction frameworks. LRP has been used for example, in Alzheimer's disease prediction using Graph Convolutional Networks (GCNs) and by identifying the critical connections in the brain network based on Functional Magnetic Resonance Imaging (fMRI) data, with a high precision (91%) yet maintaining interpretability (Ango et al., 2024). At the task of prediction based on fMRI data, LRP enabled interpretable explanation with high precision in learning critical connections in brain networks through Graph Convolutional Networks (GCNs). Studies applying GCNs to the Alzheimer 's disease Neuroimaging Initiative (ADNI) database have reported promising predictions on the basis of cognitive status, from Normal Cognition (NC), Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD) (Tekkesinoglu et al., 2024). To illustrate, the study of (Ozdemir et al., 2025) introduces DyEPAD, a dynamic deep learning model based on GCNs and tensor algebraic operations, to predict MCI subjects' progression to AD from EHR. In support of this, another study proposed a regional brain fusion graph convolutional network (RBF-GCN) where structural MRI, diffusion weighted MRI, and amyloid PET is integrated to highlight unique affinities of the AD burden to different regions of the brain. Comorbidity based frameworks modeled through Graph Neural Networks have also shown high classification accuracy in multi class classification at different stages of AD, providing a low cost unconstrained method of early AD prediction (Abuhantash et al., 2024).

In particular, LRP significantly outperforms other explainability methods such as LIME and Deep Taylor Decomposition in robustness metrics on lung disease classification (i.e. COVID19). (Pitroda et al., 2021). One example is U-Net combined with attention mechanism and ViTs for lung disease segmentation and classification of chest X ray images; the details of the study include combining with LRP to identify the key areas feeding into model decisions while achieving high segmentation accuracy. Now not only it increases the diagnostic accuracy but also gives the clinicians the interpretability, which in turn

builds the trust and aids in making a more informed decision in the management of lung diseases (Pitroda et al., 2021).

LRP has already been shown to be useful for personalized medicine and treatment recommendation by explaining model holds and extracting patient specific features predicting outputs. For instance, LRP can mark efficacy biomarkers or safety biomarkers with which clinicians can personalize interventions based on an individual's needs. In the realm of cancer therapy, LRP makes learning outcomes attributable to single genes so that they can among other things explain their importance and pinpoint targetable genes for individualized therapies (Böhle et al., 2019). This capability increases the precision of a precise treatment strategy based on knowledge now deeper on a patient's unique characteristics. Furthermore, we also integrated Genetic Algorithm-Enhanced Convolutional Neural Networks with LRP to improve the understanding of clinicians for CNN based diagnostic decision by pinpointing influential regions in medical image for oral cancer detection (Khanna et al., 2024). As well, LRP is useful for understanding how single pixels contribute to classifications for multiple images datasets thereby generating visual heatmaps that help human experts check decision and direct subsequent analysis on candidate regions of interest (Bach et al. 2015). LRP provides these incites in a form that is detailed, interpretable, and bridges the gap between so complex AI models and ultimately how a clinical decision is to be made.

Evaluation of LRP in Large-Scale Neural Networks

This analysis, however, is also central to the use of LRP in these large scale networks where it can be deployed as trade between accuracy and interpretability. While LRP increased interpretability (Mean Relevance Score of 0.85), it did not hurt the diagnostic accuracy (91%) (Colin et al., 2025). Very well, the use of LRP in Alzheimer's prediction models maintained high sensitivity and specificity while ensuring relevance alignment with clinical biomarkers (Ango et al., 2024). The general success of achieving such a balance often comes with the consideration of the model complexity where simpler models can perform as well or even better as more complex ones, yet providing more transparency. In particular, simpler, intrinsically interpretable neural networks can achieve similar predictive performance as deep convolutional neural nets (CNNs) and better determine key patterns in the data1. In doing so, it points out the need to have both considered when choosing an accurate model and explanation model as we try to balance accuracy and interpretability (Lovo et al., 2024).

Case Studies of LRP in Large Datasets

Datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) have been widely utilized in fMRI-based Graph Convolutional Network (GCN) models, demonstrating their potential to accurately differentiate Alzheimer's patients from healthy individuals while offering interpretable insights into brain connectivity (Ango et al., 2024, Colin et al., 2025). In a dataset for pneumonia detection, LRP heatmaps consistently showed brain regions that are also relevant for diagnosis, while improving clinical trust without sacrificing AUC-ROC performance measures (Colin et al., 2025). Using chest X-rays, lung disease segmentation and classification was studied with U-Net (including attention) and ViTs. To explain, the model decisions were explained using LRP, we identified the crucial areas that helped the model make decisions (Gupta et al. 2024). EEG sleep stage classification is interpreted based on LRP. In the LRP method, the contribution of each frequency pixel in the input time-frequency image to the model prediction is evaluated under the condition that aligns with sleep scoring guidelines. For the input used, it used the MSSENet method which consists of the MSCNN module and the residual squeeze and excitation (R-SE) block based CNN (Zhou et al., 2024).

The Universal Local Adversarial Network (ULAN) is an example of a semi white box attack network that utilizes LRP to generate universal adversarial perturbations (UAP) of the target regions in the SAR (Synthetic Aperture Radar) images. Through LRP, we calculated the model's attention heatmaps and thus the target regions of SAR images that receive high relevance for recognition results (Du et al.,

2022). For AD based on MRI, convolutional neural network decisions were visualized using LRP. The LRP method is patient specific with high inter patient variability and the actual high relevance correlated well with what is known from literature (Böhle et al., 2019).

Large scale medical imaging is complex model and computationally challenging problem that presents significant hurdles for Layer-wise Relevance Propagation (LRP) The fine-grained relevance mapping needed by LRP increases computation overhead when processing such high resolution medical images as 3D MRI scans (Zakaria et al., 2024), and modern architectures such as ViTs further add difficulties to be overcome by designing focused relevance redistribution techniques that are capable to work with skip connections and attention mechanism (Yan et al., 2024). These issues must be tackled utilizing hybrid solutions, i.e., the combination of LRP with the faster methods such as Grad-CAM keeping the computational costs in balance and interpretability. In addition, optimization strategies such as mixed precision quantization can reduce demands for computation and therefore enable LRP analysis of ViTs. However, from a manufacturing standpoint, these advancements are extremely important because they are making LRP an appealing technique in complicated diagnostic medical scenarios (Tan et al., 2024).

Optimization Techniques

Hybrid approaches combining LRP with faster methods like Grad-CAM have been proposed to reduce computational costs while preserving interpretability.(Pitroda et al., 2021) To address the computational demands of LRP, particularly in large-scale medical imaging, hybrid approaches have emerged that combine LRP with computationally efficient methods like Grad-CAM.(Gupta et al., 2024)These hybrid techniques aim to strike a balance between detailed interpretability and practical feasibility.(Bach et al., 2015)By leveraging Grad-CAM for initial feature localization and then refining the analysis with LRP, these methods can reduce computational overhead while preserving key interpretative insights. Such strategies are necessary to facilitate the wider uptake of such AI interpretable in clinical settings, where timely and trusted diagnostics are both necessary and conclude (Lutz et al., 2023).

LRP is a valuable tool for establishing the connection between performance and transparency of medical AI systems. While the computational challenges it presents are easier than many algorithms, its use in large scale neural networks still relies on solving it. **Table 2** provides an overview of the applications of LRP in medical diagnosis. The table is structured into four columns: 'Medical field' specifies the area of healthcare (such as radiology, oncology, or cardiology) where LRP has been applied; 'Ref' cites the corresponding studies or references; 'Application' describes how LRP was utilized within that field (e.g., image interpretation, disease classification, or risk prediction); and 'Outcomes' summarizes the key findings or benefits achieved, such as improved model transparency, enhanced diagnostic accuracy, or better clinical decision support.

Table 2 Applications of LRP in Medical Diagnosis

Medical field	References	Application	Outcomes
Radiology	(Santhi et al., 2025)	Brain tumor detection using advanced learning algorithms	It discusses the use of advanced learning algorithms like CNNs, SVMs, and hybrid models for brain tumor detection from MRI images, enhancing extraction, segmentation, and classification.
	(Azeez et al., 2024)	Brain Tumor Detection using machine learning algorithms	The review discusses machine learning techniques for brain tumor detection from MRI images, including SVM and CNNs, and emphasizes noise reduction, intensity normalization, texture analysis, and diverse datasets for improved diagnostic outcomes.

	(Farooqui et al., 2024)	Brain Tumor Detection using machine learning algorithms	It evaluates machine learning techniques for brain tumor detection from MRI images, highlighting convolutional neural networks' superior accuracy, sensitivity, and specificity. It emphasizes traditional methods' effectiveness in limited dataset scenarios and data augmentation techniques.
	(Junaid et al, 2024)	Brain Tumor Detection using machine learning algorithms	This paper reviews Machine Learning techniques for brain tumor detection using MRI images, highlighting their efficacy, reliability, and computational complexity. It emphasizes early diagnosis, cancer grade classification, and advancements in segmentation.
	(Berghout, 2024)	Brain Tumor Detection using deep learning techniques	It discusses deep learning techniques for brain tumor detection, focusing on Convolutional Neural Networks, GANs, Autoencoders, and Recurrent Neural Networks. It emphasizes transfer learning and explainable AI's limited adoption in medical diagnostics.
Pathology	(Haffner et al., 2024)	Prostate cancer biopsies	It presents a classification framework for metastatic castration-resistant prostate cancer biopsies, highlighting the significance of AR, NKX3.1, INSM1, synaptophysin, and Ki-67 in clinical trial design and practice.
	(Vibishan et al, 2023)	Metastatic castration-resistant prostate cancer biopsies	It discusses resource dynamics and intra-tumor interactions affecting metastatic castration-resistant prostate cancer biopsies growth and progression, rather than providing a specific framework for pathology workup methodologies.
	(Ku et al, 2019)	Metastatic biopsy programs and genomic testing	It highlights the significance of metastatic biopsy programs and genomic testing in identifying actionable targets, DNA repair aberrations, and guiding therapeutic strategies and clinical trial eligibility in advanced prostate cancer.
	(Trigos et al., 2023)	Biomarker expression in metastatic castration-resistant prostate cancer biopsies	It explores biomarker expression in metastatic castration-resistant prostate cancer biopsies, highlighting their implications for treatment and patient stratification, without providing a specific framework.
Cardiology	(McKinn et al., 2024)	Heart Health Yarning Tool	The Heart Health Yarning Tool is a tool designed to promote shared decision-making in cardiovascular disease prevention, especially for Aboriginal and Torres Strait Islander people, enhancing clinician communication and risk assessment.
	(Carlton et al., 2024)	Risk assessment tools	The study highlights inaccuracies in risk assessment tools for atherosclerotic cardiovascular disease prevention in patients with raised lipoprotein (a),

			suggesting the need for tailored approaches.
	(Karwa et al., 2024)	Risk assessment tools	Risk assessment tools like Framingham, atherosclerotic CVD calculator, QRISK, and Reynolds aid in primary prevention by evaluating risk factors, enabling early identification and intervention to reduce cardiovascular disease incidence.
	(van Daalen, 2024)	Risk Assessment	Cardiovascular disease risk scores are crucial for identifying high-risk individuals and guiding preventive interventions, but data source differences can lead to inaccurate estimations.
Neurology	(Bhattacharya, 2024)	Brain lesions identification from unannotated MRI and EEG	It introduces a self-supervised method using contrastive learning to identify brain lesions from unannotated MRI and EEG data, bypassing the need for human intervention.
	(Alaverdyan, 2019)	Epilepsy lesion detection using MRI	It discusses unsupervised representation learning for epilepsy lesion detection using MRI data, focusing on T1-weighted and FLAIR sequences, voxel-level outlier detection, and multimodal data integration.
	(Chen et al., 2018)	Brain MRI lesion detection using constrained adversarial auto-encoders	It discusses unsupervised brain MRI lesion detection using constrained adversarial auto-encoders, focusing on learning healthy brain MRI distributions to enhance detection without labeled datasets.
	(Guo et al., 2015)	lesion detection using T1-weighted MRIs	It discusses automated lesion detection using T1-weighted MRIs, combining unsupervised and supervised methods, but does not specifically address MRI-EEG instance discrimination for brain lesion identification.

Benefits of LRP

This adds reliability of the model in clinical contexts due to its interpretable insights into the model's logic provided by LRP. LRP explains the model's classification by highlighting image regions most relevant to its decision, thus empowering AI driven diagnoses to build trust (Manuela et al., 2024). In particular, LRP helps in nursing acceptance by identifying important parts that affect model decisions. This allows clinicians to validate AI generated diagnoses and can be effectively integrated in clinical workflow (Gupta et al., 2024). LRP closes the gap between complicated AI models and the clinical decisions by offering transparency, trust and practical implementation in health care setting.

Limitations of LRP

While the relevance mapping in LRP can be utilized for high resolution medical images (e.g. 3D MRI scans), high computational overhead arises in applying LRP to this type of images. The applications of LRP are likely to scale as long as the dataset or model size is not very large. However, since LRP is based off of the model's learned features, any bias that the model is trained on will be reflected and subsequently amplified in the relevance maps (Gupta et al., 2024). As a result, while LRP aids interpretability, it must be deployed within the computational limits and constrained so as to prevent LRP bias to ensure its reliable and scalable deployment in clinical settings.

Future Directions

To alleviate the computational cost without sacrificing the interpretability, hybrid approaches consisting of LRP with faster methods such as Grad-CAM were presented (Otsuki et al., 2024). Reducing the computational and memory requirements of ViTs can be done using Mixed-precision quantization (MPQ). In this case, LRP can assign a mixed-precision bit allocation to different layers according to its importance in classification (Ranjan et al., 2024). The potential of such hybrid strategies and optimization techniques for making LRP based interpretations more efficient and practical to deploy in real world medical AI systems, motivates the inquiry into the following questions.

Consequently, LRP needs to be extended to deal with attention layers to enable faithful attributions for the entire black-box transformer model, while being computationally efficient (Achtibat et al., 2024). In the case of ViTs, relevance redistribution techniques need to be tailored to the skip connections and attention mechanisms. The conservation property is guaranteed during the whole process by this formulation, thus they maintain the integrity of the explanations generated by it (Achtibat et al., 2024). Overall, these extensions ensure that LRP is flexible and robust across these types of architectures such as transformers and ResNets, allowing for improved and more trusted model explanations.

Ethical Issues and Regulatory Consequences

Since their complexity, large scale neural network's decision making processes are often opaque, and their clinical acceptance is therefore difficult. For this reason, LRP provides a transparent mapping from input data to model prediction that allows regulators to evaluate the reliability of a model. In settings such as healthcare, understanding the reasoning behind a diagnosis is crucial, as it can directly influence treatment decisions and strengthen patient trust (Samek et al., 2017).

As AI based medical diagnostics continue increasing, it becomes necessary to identify those who are at fault when something goes wrong. As LRP helps, it makes it easier to determine what were the important elements that helped the model choose that path and to identify errors that may have been caused by specific data input or layers of the model itself (Montavon et al., 2018), since the FDA's proposed regulatory Framework for AI/ML based Software as a Medical Device (SaMD), strongly underlines on maintaining continuous monitoring and transparency (Brown et al., 2021).

CONCLUSION

The review explores the ways in which neural networks have redefined medical diagnosis as a mission of high accuracy and efficiency for analyzing complex electronic data. But with more and more reliance on AI they have realized the importance of this. This review provides the context of how the neural network makes a decision by pointing out Layer-wise Relevance Propagation (LRP) as a helpful method that aids in building trust, ensuring transparency, and facilitating the clinical integration process in healthcare settings. Future research should focus on refining interpretability methods like LRP and developing standardized frameworks to ensure that neural networks in healthcare remain transparent, trustworthy, and seamlessly integrated into clinical practice.

Acknowledgments: None.
Funding: None
Ethics Statement: Not applicable
Conflicts of interest: The authors declare no conflict of interest.

REFERENCES:

Abuhantash, F., Abu Hantash, M. K., & AlShehhi, A. (2024). Comorbidity-based framework for Alzheimer's disease classification using graph neural networks. Scientific Reports, 14(1), 21061.

Achtibat, R., Hatefi, S. M. V., Dreyer, M., Jain, A., Wiegand, T., Lapuschkin, S., & Samek, W. (2024). Attnlrp: attention-aware layer-wise relevance propagation for transformers. arXiv preprint arXiv:2402.05602.

Alam, M. U., Hollmén, J., Baldvinsson, J. R., & Rahmani, R. (2023). SHAMSUL: Systematic Holistic Analysis to investigate Medical Significance Utilizing Local interpretability methods in deep learning for chest

radiography pathology prediction. arXiv preprint arXiv:2307.08003.

Alaverdyan, Z. (2019). Unsupervised representation learning for anomaly detection on neuroimaging. Application to epilepsy lesion detection on brain MRI (Doctoral dissertation, Université de Lyon).

Ango, R., Fatima, S., & Nag, A. (2024). Brain Connectivity Analysis in Alzheimer's disease using Graph Convolutional Network. Paper presented at the 2024 4th International Conference on Soft Computing for Security Applications (ICSCSA).

Azeez O, A. M. A. (2024). Classification of Brain Tumor based on Machine Learning Algorithms: A Review.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), e0130140.

Berghout, T. (2024). The Neural Frontier of Future Medical Imaging: A Review of Deep Learning for Brain Tumor Detection.

Bhattacharya, S. (2024). MRI-EEG Instance Discrimination for Brain Lesion Identification

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Frontiers in aging neuroscience, 11, 194.

Brown, N. A., Carey, C. H., & Gerry, E. I. (2021). FDA Releases Action Plan for Artificial Intelligence/Machine Learning-Enabled Software as a Medical Device. The Journal of Robotics, Artificial Intelligence & Law, 4.

Carlton, H. C., Banerjee, A., Burnett, D., & Shipman, K. E. (2024). Comparison of cardiovascular disease risk assessment tools in primary prevention patients with raised lipoprotein (a). European Heart Journal, 45(Supplement_1), ehae666-2709.

Carolina Gomes Alexandre, Tracy Jones, Jessica L. Hicks, John T. Isaacs, Anuj Gupta, Alyza Skaist, Laura Sena, Jennifer Meyers, Emmanuel Antonarakis, Mark Markowski, Samuel Denmeade, Srinivasan Yegnasubramanian, Angelo Michael De Marzo; Abstract 653: Molecular pathology of metastatic prostatic adenocarcinoma treated with bipolar androgen therapy (BAT) reveals a correlation between MYC mRNA and protein. Cancer Res 15 June 2022; 82 (12_Supplement): 653. https://doi.org/10.1158/1538-7445.AM2022-653

Chen, X., & Konukoglu, E. (2018). Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972.

Colin, J., & Surantha, N. (2025). Interpretable Deep Learning for Pneumonia Detection Using Chest X-Ray Images. Information, 16(1), 53.

Du, M., Bi, D., Du, M., Xu, X., & Wu, Z. (2022). ULAN: A universal local adversarial network for SAR target recognition based on layer-wise relevance propagation. Remote Sensing, 15(1), 21.

Farooqui, M. A. (2024). Machine Learning in Brain Tumor Diagnosis: Assessing the Efficacy of Diverse Methods. https://powertechjournal.com/index.php/journal/article/view/1272

Ferles, A., & Barkhof, F. (2023). Computerised emergency work-up for mass-like brain MRI lesions: can explainable AI support radiologists? European Radiology, 33(8), 5856-5858.

Guo, D., Fridriksson, J., Fillmore, P., Rorden, C., Yu, H., Zheng, K., & Wang, S. (2015). Automated lesion detection on MRI scans using combined unsupervised and supervised methods. BMC medical imaging, 15, 1-21.

Gupta, S., Dubey, A. K., Singh, R., Kalra, M. K., Abraham, A., Kumari, V., . . . Singh, I. (2024). Four transformer-based deep learning classifiers embedded with an attention U-Net-based lung segmenter and layer-wise relevance propagation-based heatmaps for COVID-19 X-ray scans. Diagnostics, 14(14), 1534.

Haffner, M. C., Morris, M. J., Ding, C. K. C., Sayar, E., Mehra, R., Robinson, B., ... & Beltran, H. (2025). Framework for the pathology workup of metastatic castration-resistant prostate cancer biopsies. Clinical Cancer Research, 31(3), 466-478.

Hou, B., Kaissis, G., Summers, R. M., & Kainz, B. (2021). Ratchet: Medical transformer for chest x-ray diagnosis and reporting. Paper presented at the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24.

Jia, H., Zhang, J., Ma, K., Qiao, X., Ren, L., & Shi, X. (2024). Application of convolutional neural networks in medical images: a bibliometric analysis. Quantitative Imaging in Medicine and Surgery, 14(5), 3501.

Junaid, L., Noor, J., Bibi, A., Rahman, J. S. U., Akram, F., Khan, S. H., & Selvaperumal, S. K. (2024). A systematic review on a comparative study of AI techniques for the classification of brain tumour. Asian Journal of Science, Engineering and Technology (AJSET), 3(1), 115-134.

Karwa, V., Wanjari, A., Kumar, S., Dhondge, R. H., Patil, R., & Kothari, M. (2024). Optimizing Cardiovascular Health: A Comprehensive Review of Risk Assessment Strategies for Primary Prevention. Cureus, 16(8).

Khanna, S. T., Khatri, S. K., & Sharma, N. K. (2024). GACNNXAI: Employing Genetic Algorithm-Enhanced Convolutional Neural Networks and Explainable Artificial Intelligence and its Applications. Paper presented at the 2024 IEEE Third International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES).

Ku, S. Y., Gleave, M. E., & Beltran, H. (2019). Towards precision oncology in advanced prostate cancer. Nature Reviews Urology, 16(11), 645-654.

Lim, C. Y., Sohn, B., Seong, M., Kim, E. Y., Kim, S. T., & Won, S. Y. (2024). Need for Transparency and Clinical Interpretability in Hemorrhagic Stroke Artificial Intelligence Research: Promoting Effective Clinical Application. Yonsei Medical Journal, 65(10), 611.

Lovo, A., Lancelin, A., Herbert, C., & Bouchet, F. (2024). Tackling the Accuracy-Interpretability Trade-off in a Hierarchy of Machine Learning Models for the Prediction of Extreme Heatwaves. arXiv preprint arXiv:2410.00984.

Lutz, S., Auer, F., Hartmann, D., Chereda, H., Beißbarth, T., & Kramer, F. (2023). Adaptation of Graph Convolutional Neural Networks and Graph Layer-wise Relevance Propagation to the Spektral library with application to gene expression data of Colorectal Cancer patients. bioRxiv, 2023.2001. 2026.525010.

Manuela, U. M., Nakasi, R., Jjingo, D., Hellen, N., Ngobye, M., & Marvin, G. (2024). Machine Vision Intelligence Using Layer-Wise Relevance Backward Propagation For Breast Cancer Diagnosis. Paper presented at the 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN).

McKinn, S., Parnham, J., Follent, D., Tracy, M., Wyber, R., Freeman, N., ... & Bonner, C. (2024). The Heart Health Yarning Tool: co-designing a shared decision-making tool for cardiovascular disease prevention and risk management. medRxiv, 2024-11.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital signal processing, 73, 1-15.

Moritz Böhle, F. E., Martin Weygandt, Kerstin Ritter. Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification.

Nam, H., Kim, J.-M., & Kam, T.-E. (2023). Feature selection based on layer-wise relevance propagation for EEG-based mi classification. Paper presented at the 2023 11th International Winter Conference on Brain-Computer Interface (BCI).

Otsuki, S., Iida, T., Doublet, F., Hirakawa, T., Yamashita, T., Fujiyoshi, H., & Sugiura, K. (2024). Layer-Wise Relevance Propagation with Conservation Property for ResNet. Paper presented at the European Conference on Computer Vision.

Ozdemir, C., Al Olaimat, M., Bozdag, S., & Initiative, A. s. D. N. (2025). A Dynamic Model for Early Prediction of Alzheimer's Disease by Leveraging Graph Convolutional Networks and Tensor Algebra. Paper

presented at the Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.

Pitroda, V., Fouda, M. M., & Fadlullah, Z. M. (2021). An explainable AI model for interpretable lung disease classification. Paper presented at the 2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS).

Pordeli Shahreki, A., Hosseini-Baharanchi, F. S., & Roudbari, M. (2024). Diagnosis of Tuberculosis Using Medical Images by Convolutional Neural Networks. Journal of Kerman University of Medical Sciences, 31(4), 180-186.

Ranjan, N., & Savakis, A. (2024). Lrp-qvit: Mixed-precision vision transformer quantization via layer-wise relevance propagation. arXiv preprint arXiv:2401.11243.

Rashid, A., Rashid, A., Ali, A., Rubab, R., Ali, R., & Ullah, H. (2024). RISK FACTORS, PREVALENCE, AND SURGICAL OUTCOMES OF PALATAL FISTULAS: A GLOBAL AND PAKISTANI PERSPECTIVE. The Research of Medical Science Review, 2(3), 438-451. http://www.thermsr.com/index.php/Journal/article/view/111

Rutger F. R. van Mierlo, B. S., Joost G. E. (2024). Optimizing cardiovascular risk management in primary care with a personalized eCoach solution enriched by an AI-driven clinical prediction model: a study protocol of the CARRIER consortium (Preprint)

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Santhi K, S. K. (2025). A Comprehensive Review on Brain Tumor Detection Using Advanced Learning Algorithms. DOI: https://doi.org/10.47392/IRJAEH.2025.0008

Shafiq, M., Sami, M. A., Bano, N., Bano, R., & Rashid, M. (2025). Artificial Intelligence in Physics Education: Transforming Learning from Primary to University Level. Indus Journal of Social Sciences, 3(1), 717-733. DOI: https://doi.org/10.59075/ijss.v3i1.807

Sheng-Yi Hsu, Mau-Hsiang Shih., Wu-Hsiung Wu., Hao-Ren Yao., Feng-Sheng Tsai. (30 June 2024). Gene reduction for cancer detection using layer-wise relevance propagation.

Spitzer, H., Ripart, M., Whitaker, K., Napolitano, A., De Palma, L., De Benedictis, A., Wagstyl, K. . (2023). Automated lesion prediction and characterisation of focal cortical abnormalities: a MELD study

Taiyeb Khosroshahi, M., Morsali, S., Gharakhanlou, S., Motamedi, A., Hassanbaghlou, S., Vahedi, H., . . . Jafarizadeh, A. (2025). Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease. Diagnostics, 15(5), 612.

Tan, S., Zhang, Z., Cai, Y., Ergu, D., Wu, L., Hu, B., . . . Zhao, Y. (2024). Segstitch: Multidimensional transformer for robust and efficient medical imaging segmentation. arXiv preprint arXiv:2408.00496.

Tekkesinoglu, S., & Pudas, S. (2024). Explaining graph convolutional network predictions for clinicians—An explainable AI approach to Alzheimer's disease classification. Frontiers in Artificial Intelligence, 6, 1334613.

Trigos, A. S., Pasam, A., Inderjeeth, A. J., Cain, L. D., Weng, S., Gupta, V., ... & Sandhu, S. (2023). Heterogeneity of canonical prostate cancer markers across lesions in metastatic castration-resistant prostate cancer.

Tulsani, V., Sahatiya, P., Parmar, J., & Parmar, J. (2023). XAI Applications in Medical Imaging: A Survey of Methods and Challenges.

van Daalen, K. R., Zhang, D., Kaptoge, S., Paige, E., Di Angelantonio, E., & Pennells, L. (2024). Risk estimation for the primary prevention of cardiovascular disease: considerations for appropriate risk prediction model selection. The Lancet Global Health, 12(8), e1343-e1358.

Vibishan, B., Harshavardhan, B. V., & Dey, S. (2024). A resource-based mechanistic framework for castration-resistant prostate cancer (CRPC). Journal of Theoretical Biology, 587, 111806.

Yan, X., Sun, S., Han, K., Le, T.-T., Ma, H., You, C., & Xie, X. (2024). After-sam: Adapting sam with axial fusion transformer for medical

imaging segmentation. Paper presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

Yang, J., Chereda, H., Dönitz, J., Bleckmann, A., & Beißbarth, T. (2024). Deciphering BRCAness Phenotype in Cancer: A Graph Convolutional Neural Network Approach with Layer-wise Relevance Propagation Analysis. An automated data integration platform for interpreting genomic data and reporting treatment options in molecular tumor boards, 42.

Zafari-Ghadim, Y., Soliman, A., Yousif, Y., Ibrahim, A., Rashed, E. A., & Mabrok, M. (2024). Deep models for stroke segmentation: do complex architectures always perform better? arXiv preprint arXiv:2403.17177.

Zakaria, M., & Mamun, M. A. I. (2024). Deep Neural Networks in Medical Imaging: Advances, Challenges, and Future Directions for Precision Healthcare.

Zeynali, A., Tinati, M. A., & Tazehkand, B. M. (2024). Hybrid CNN-Transformer Architecture with Xception-Based Feature Enhancement for Accurate Breast Cancer Classification. IEEE Access.

Zhou, D., Xu, Q., Zhang, J., Wu, L., Xu, H., Kettunen, L., . . . Cong, F. (2024). Interpretable sleep stage classification based on layer-wise relevance propagation. IEEE Transactions on Instrumentation and Measurement.