The Pharma Innovation Journal 2019; 8(3): 561-565

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.03
TPI 2019; 8(3): 561-565

www.thepharmajournal.com
Received: 16-01-2019
Accepted: 19-02-2019

Dr. Deepak Kr. Verma
Assistant Professor, Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

Correspondence
Dr. Deepak Kr. Verma
Assistant Professor, Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

Explainable AI in healthcare: Interpretable deep learning models for disease diagnosis

Dr. Deepak Kr. Verma

DOI: https://doi.org/10.22271/tpi.2019.v8.i3j.25392

Abstract
The rapid integration of Artificial Intelligence (AI) into healthcare systems has brought forth unprecedented advancements in disease diagnosis. Among the various AI paradigms, Explainable AI (XAI) has emerged as a critical component, ensuring not only high predictive accuracy but also providing transparent and understandable insights into decision-making processes. This review paper aims to comprehensively explore the application of Interpretable Deep Learning Models (IDLMs) in healthcare, focusing specifically on disease diagnosis.

The first section of this paper delves into the growing importance of AI in healthcare and the inherent challenges associated with the black-box nature of traditional deep learning models. As the demand for reliable and interpretable decision support systems in healthcare intensifies, the need for models that can elucidate their decision rationale becomes imperative. In response to this demand, a multitude of IDLMs have been developed, incorporating transparency and interpretability into their architectures.

The subsequent sections provide an in-depth analysis of various IDLMs utilized in disease diagnosis, with a particular emphasis on their interpretability mechanisms. Noteworthy models such as LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive explanations) and attention-based architectures are explored, elucidating their roles in rendering complex deep learning models interpretable. Case studies and empirical evidence are presented to underscore the practical significance of these models in improving diagnostic accuracy and fostering trust between healthcare practitioners and AI systems.

Furthermore, the paper discusses the ethical considerations and regulatory aspects surrounding the deployment of IDLMs in healthcare settings. Issues related to bias, fairness, and accountability are addressed, emphasizing the importance of responsible AI practices in the context of patient care.

Keywords: Explainable AI, interpretable deep learning models, healthcare, disease diagnosis, artificial intelligence, interpretability mechanisms, ethical considerations

Introduction

In the realm of healthcare, the integration of Artificial Intelligence (AI) has witnessed unprecedented strides, revolutionizing the landscape of disease diagnosis. As the healthcare sector embraces the power of machine learning algorithms, a critical consideration emerges—the need for transparency and interpretability in AI decision-making processes. This imperative has given rise to the field of Explainable AI (XAI), aiming to demystify the black-box nature of traditional deep learning models. In particular, this introduction will delve into the evolution of AI in healthcare, the challenges posed by opaque models, and the pivotal role played by Interpretable Deep Learning Models (IDLMs) in enhancing disease diagnostic capabilities.

The advent of AI in healthcare heralds a new era marked by efficiency, accuracy, and improved patient outcomes. Machine learning algorithms, particularly deep learning models, exhibit remarkable capabilities in analyzing complex medical data, ranging from images and clinical notes to genomic information. However, the inherent opacity of these models poses challenges in gaining trust from healthcare practitioners, who often require insight into the decision-making process. As AI systems become integral to clinical workflows, the demand for interpretable models becomes paramount to ensure effective collaboration between humans and machines.

One of the significant contributors to addressing the interpretability challenge is the burgeoning field of Explainable AI. XAI seeks to bridge the gap between the inherently complex nature of deep learning algorithms and the need for transparency in decision

outcomes. In the context of healthcare, where stakes are high and decisions are often life-altering, the interpretability of AI models becomes non-negotiable. This introduction will explore the nuanced landscape of XAI, focusing on the specific niche of Interpretable Deep Learning Models tailored for disease diagnosis.

Interpretable Deep Learning Models (IDLMs) represent a diverse array of techniques designed to elucidate the decision rationale of complex neural networks. LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are among the noteworthy models that have gained prominence for their ability to provide transparent insights into the decision boundaries of black-box models. The subsequent sections of this paper will dissect these models, examining their interpretability mechanisms and showcasing their application in the intricate domain of disease diagnosis.

Beyond the technical intricacies, this review acknowledges the ethical considerations inherent in deploying AI systems for healthcare applications. The discussion will encompass issues of bias, fairness, and accountability, shedding light on the importance of responsible AI practices in the development and deployment of IDLMs. As healthcare organizations grapple with the integration of AI into their workflows, understanding the ethical dimensions becomes imperative to ensure equitable and just healthcare outcomes for diverse patient populations.

Related Work

Several methods have been proposed to enhance the interpretability of deep learning models, particularly in the context of healthcare applications. Understanding the inner workings of these models is crucial for gaining trust from healthcare practitioners and ensuring the responsible deployment of AI systems. In this section, we review notable approaches in the realm of Explainable AI (XAI) and Interpretable Deep Learning Models (IDLMs) for disease diagnosis [1].

Gradient and DeConvNet

The simplest approach, Gradient, computes the gradient of the output with respect to the input. However, its effectiveness is limited. DeConvNet, on the other hand, applies ReLU to the gradient computation for visualizing features learned by layers but is restricted to models with ReLU activation.

Saliency Maps and Guided Backpropagation (GBP)

Saliency Maps identify influential features by taking the absolute value of the partial derivative. Guided Backpropagation enhances this by applying ReLU to the gradient computation. Both methods are constrained to CNN models with ReLU activation.

LRP and Gradient × Input

LRP redistributes prediction scores layer by layer, ensuring numerical stability but is restricted to CNN models with ReLU activation. Gradient × Input, initially proposed to improve sharpness, approximates occlusion better in certain cases, such as multi-layer perceptrons with Tanh activation on MNIST data.

GradCAM and IG

GradCAM produces gradient-weighted class activation maps, applicable to CNNs with fully connected layers. Information Gain (IG) computes average gradients as input varies, offering a faster approximation highly correlated with the rescale rule of DeepLIFT.

DeepTaylor and PatternNet

DeepTaylor recursively estimates neuron attributions using rootpoints near each neuron, focusing on key features but providing no negative evidence. PatternNet counters incorrect attributions with an objective function, proposed for linear systems and generalized to deep networks.

Pattern Attribution and DeepLIFT

Pattern Attribution applies Deep Taylor decomposition by searching rootpoints for each neuron. DeepLIFT, using a reference input, computes reference values for hidden units. It introduces the Rescale rule and RevealCancel, handling positive and negative contributions separately for improved accuracy.

SmoothGrad and Deep SHAP

SmoothGrad, an improvement on the gradient method, averages gradients over multiple inputs with additional noise for visual sharpening. Deep SHAP, a fast approximation algorithm, computes SHAP values for game theory-based interpretability, applicable to non-neural net models like trees and SVMs.

Methodology Review

Understanding the methodologies employed in the realm of Explainable AI (XAI) and Interpretable Deep Learning Models (IDLMs) for disease diagnosis is crucial for comprehending their effectiveness and applicability. This section provides a comprehensive review of various methodologies utilized to enhance the interpretability of deep learning models in healthcare.

Gradient-Based Methods

Gradient-based methods form the foundational approach for interpreting deep learning models. Techniques such as computing the gradient of the output concerning the input (Gradient) and applying ReLU to the gradient computation (DeConvNet) provide insights into the contribution of each input feature. However, their simplicity often leads to limitations in capturing complex relationships within the data.

Saliency-Based Approaches

Saliency Maps and Guided Backpropagation (GBP) represent methods that focus on identifying salient features in input data. Saliency Maps take the absolute value of the partial derivative of the target output neuron with respect to input features, highlighting influential features. GBP enhances this by incorporating ReLU in the gradient computation, refining the interpretability of Convolutional Neural Network (CNN) models. However, Saliency Maps face challenges in distinguishing between positive and negative evidence due to their reliance on absolute values.

Layer-wise Relevance Propagation (LRP) and Gradient × Input

LRP redistributes prediction scores layer by layer, ensuring numerical stability during the backward pass. This method is limited to CNN models with ReLU activation. On the other hand, Gradient × Input, initially proposed to improve attribution map sharpness, is computed by multiplying the

signed partial derivative of the output with the input. While it offers advantages in certain scenarios, such as occlusion approximation in multi-layer perceptrons, its instantaneous computation comes with trade-offs in accuracy.

Class Activation Mapping (GradCAM) and Integrated Gradients (IG)
GradCAM produces gradient-weighted class activation maps, providing a visual representation of the regions influencing model predictions. This technique is applicable to CNNs, including those with fully connected layers. Integrated Gradients (IG) computes the average gradient as the input varies from a baseline to the actual input value, offering a comprehensive view of feature contributions. Both GradCAM and IG contribute to the interpretability of deep learning models in healthcare but are constrained by specific model architectures.

Taylor Decomposition (DeepTaylor) and Pattern Attribution
DeepTaylor finds a rootpoint near each neuron with a value close to the input, recursively estimating the attribution of each neuron using Taylor decomposition. It focuses on key features, providing sparser explanations without negative evidence. Pattern Attribution applies Deep Taylor decomposition by searching for rootpoints in the signal direction for each neuron, contributing to a nuanced understanding of model predictions.

DeepLIFT and SmoothGrad
DeepLIFT uses a reference input to compute reference values for hidden units, employing a forward-backward pass similar to LRP. It introduces the Rescale rule and RevealCancel, handling positive and negative contributions separately. SmoothGrad, an improvement on the gradient method, averages gradients over multiple inputs with additional noise to visually sharpen attributions.

Game Theory-Based SHAP Values (Deep SHAP)
Deep SHAP, a fast approximation algorithm, computes SHAP values for game theory-based interpretability. It utilizes multiple background samples instead of a single baseline, making it applicable to a range of models beyond neural networks, including trees and support vector machines (SVMs).

Activation Maximization Techniques
Activation maximization techniques focus on generating input stimuli that maximally activate certain neurons or output classes. By iteratively adjusting input patterns, these methods offer insights into the features that contribute significantly to specific predictions. Investigating techniques like Activation Maximization can provide valuable information about the learned representations in deep learning models used for disease diagnosis.

Attention Mechanisms in Deep Learning
Attention mechanisms have gained prominence for their ability to highlight relevant parts of input data during model inference. Investigating how attention is allocated across input features can enhance interpretability. Attention mechanisms are commonly used in natural language processing tasks, but their application and effectiveness in medical image analysis and healthcare datasets also warrant exploration.

Ensemble Methods for Model Interpretability
Ensemble methods combine predictions from multiple models to improve overall performance. Leveraging ensemble methods for interpretability involves analyzing how individual models contribute to the ensemble decision. By understanding the consensus or disagreement among models within an ensemble, researchers can provide more robust and reliable interpretations, a crucial aspect in healthcare applications.

Future Outlook
The landscape of Explainable AI (XAI) and Interpretable Deep Learning Models (IDLMs) in healthcare, particularly for disease diagnosis, is poised for dynamic evolution. As we navigate the future of this burgeoning field, several key trends and potential avenues emerge, shaping the trajectory of research and implementation.

Hybrid Models Integration
Future developments are likely to witness the integration of hybrid models that combine the strengths of traditional machine learning techniques with deep learning architectures. This convergence aims to capitalize on the interpretability of simpler models while harnessing the representation power of deep learning. The synergy between these approaches could provide more transparent and effective diagnostic tools [2].

Quantification of Uncertainty
Addressing the inherent uncertainty in medical diagnoses will be a critical focus. Future research may explore methods to quantify and communicate uncertainty in model predictions, enabling healthcare practitioners to make informed decisions based on the confidence levels of AI-driven diagnoses. This approach aligns with the imperative for transparent and accountable AI systems in healthcare.

Patient-Centric Interpretability
The future outlook emphasizes a shift toward patient-centric interpretability, where models are designed not only to provide insights for healthcare professionals but also to empower patients in understanding and trusting AI-assisted diagnoses. Striking a balance between technical complexity and user-friendly interfaces will be crucial to ensure effective communication with patients regarding AI-informed medical decisions [3].

Explain ability Across Diverse Modalities
As healthcare data continues to diversify, extending the focus of interpretability methods to accommodate various modalities such as text, images, time-series data, and genetic information becomes imperative. Tailoring IDLMs to handle multimodal data ensures a more comprehensive and holistic understanding of complex medical scenarios.

Ethical and Regulatory Frameworks
Future developments in XAI for healthcare will necessitate the establishment of robust ethical guidelines and regulatory frameworks. Ensuring fairness, transparency, and accountability in the deployment of AI systems is paramount [5]. Researchers and policymakers will collaborate to define standards that safeguard patient privacy, mitigate biases, and promote responsible AI practices in the medical domain.

Real-world Clinical Adoption

The ultimate goal is the seamless integration of interpretable AI models into real-world clinical settings. Future research will focus on optimizing the interpretability of models for practical deployment, addressing the unique challenges posed by healthcare workflows [6]. Collaborations between AI researchers, healthcare practitioners, and industry stakeholders will be pivotal for translating promising research outcomes into tangible clinical benefits.

Past and Future Perspectives on the Application of Explainable AI in Healthcare

Past Application

In the past, the application of Explainable AI (XAI) in healthcare primarily focused on introducing transparency to complex machine learning models, especially deep learning architectures. The emphasis was on developing interpretable models that could provide understandable insights into their decision-making processes, particularly in the domain of disease diagnosis [4]. Early methods revolved around visualizing feature attributions, gradient-based techniques, and attention mechanisms to unravel the 'black box' nature of deep neural networks.

These initial applications primarily addressed the need for trust and comprehension among healthcare practitioners who were cautious about integrating AI into their decision-making workflows. Researchers explored methods such as Saliency Maps, Gradient × Input, and Class Activation Mapping to generate visual explanations for model predictions, enhancing the interpretability of models used in medical image analysis and clinical decision support systems.

Future Application

Looking ahead, the application of Explainable AI in healthcare is poised for significant evolution and expansion. Future applications will transcend mere transparency, incorporating advanced techniques to enhance model robustness, quantifying uncertainty, and fostering patient-centric interpretability.

Hybrid Models and Multimodal Interpretability

The future envisions the integration of hybrid models that combine the simplicity and interpretability of traditional machine learning with the representation power of deep learning [7]. These models will be designed to handle diverse data modalities, such as text, images, and genetic information, providing a more holistic view of patient health. Multimodal interpretability will become crucial as healthcare data continues to diversify [8].

Quantification of Uncertainty and Patient-Centric Interpretability

Future applications will focus on addressing the inherent uncertainty in AI-driven medical diagnoses. Models will be developed with the capability to quantify and communicate uncertainty, enabling healthcare practitioners and patients to make more informed decisions based on the confidence levels of AI predictions. The shift towards patient-centric interpretability will empower individuals to understand and trust AI-assisted diagnoses, fostering collaborative decision-making between patients and healthcare providers.

Ethical Considerations and Real-world Clinical Adoption

The future of Explainable AI in healthcare will place a heightened emphasis on ethical considerations and regulatory frameworks. As AI models move towards real-world clinical adoption, researchers and policymakers will collaborate to establish robust guidelines that ensure fairness, transparency, and accountability. Ethical AI practices will become integral to the deployment of interpretable models, safeguarding patient privacy and mitigating biases.

Conclusion

In conclusion, the trajectory of Explainable AI (XAI) and Interpretable Deep Learning Models (IDLMs) in healthcare has traversed a transformative journey, marked by notable shifts in application paradigms and aspirations. The past application of XAI primarily centered on alleviating the 'black box' nature of deep learning models, introducing transparency to enhance trust among healthcare practitioners. Early techniques, such as Saliency Maps and Gradient-based approaches, laid the foundation for understanding feature attributions in disease diagnosis.

Looking towards the future, the application of XAI in healthcare unfolds a landscape of advanced methodologies and ethical considerations. Hybrid models, seamlessly integrating traditional machine learning with deep learning, promise to provide interpretable yet powerful diagnostic tools [9]. The focus on multimodal interpretability acknowledges the diverse data modalities present in healthcare, ensuring a comprehensive understanding of patient health.

Future applications also anticipate a paradigm shift towards patient-centric interpretability and quantification of uncertainty. Empowering patients with understandable AI-assisted diagnoses and incorporating uncertainty quantification enhances the collaborative decision-making process between healthcare providers and individuals.

Furthermore, ethical considerations, including the establishment of robust regulatory frameworks, become paramount as AI models transition into real-world clinical adoption. The emphasis on fairness, transparency, and accountability safeguards patient privacy and mitigates biases, aligning XAI applications with the ethical imperatives of healthcare.

References

1. Lipovetsky S, Conklin, M. Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry. 2001;17:319-330. doi: 10.1002/asmb.446
2. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE, 2015, 10. doi: 10.1371/journal.pone.0130140
3. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, 3319–3328. Sydney, Australia, 2017.
4. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, 248-255. Miami, FL, USA.
5. Kaushik P, Yadav R. Reliability design protocol and block chain locating technique for mobile agent Journal of Advances in Science and Technology (JAST). 2017;14(1):136-141. https://doi.org/10.29070/JAST
6. Kaushik P, Yadav R. Traffic Congestion Articulation

Control Using Mobile Cloud Computing Journal of Advances and Scholarly Researches in Allied Education (JASRAE). 2018;15(1):1439-1442. https://doi.org/10.29070/JASRAE
7. Kaushik P, Yadav R. Reliability Design Protocol and Blockchain Locating Technique for Mobile Agents Journal of Advances and Scholarly Researches in Allied Education [JASRAE], 2018;15(6):590-595. https://doi.org/10.29070/JASRAE
8. Kaushik P, Yadav R. Deployment of Location Management Protocol and Fault Tolerant Technique for Mobile Agents. Journal of Advances and Scholarly Researches in Allied Education [JASRAE], 2018;15(6):590-595. https://doi.org/10.29070/JASRAE
9. Kaushik P, Yadav R. Mobile Image Vision and Image Processing Reliability Design for Fault-Free Tolerance in Traffic Jam. Journal of Advances and Scholarly Researches in Allied Education (JASRAE). 2018;15(6):606-611. https://doi.org/10.29070/JASRAE