The Pharma Innovation Journal 2019; SP-8(2): 29-33

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.03
TPI 2019; SP-8(2): 29-33

www.thepharmajournal.com
Received: 12-12-2018
Accepted: 19-01-2019

Roovi Goswami
AIMT, Greater Noida, Uttar Pradesh, India

Sangeeta Yadav
AIMT, Greater Noida, Uttar Pradesh, India

Vinod Kumar
AIMT, Greater Noida, Uttar Pradesh, India

Correspondence
Roovi Goswami
AIMT, Greater Noida, Uttar Pradesh, India

Explainable AI in healthcare: A theoretical overview of interpretable models for medical diagnosis

Roovi Goswami, Sangeeta Yadav and Vinod Kumar

DOI: https://doi.org/10.22271/tpi.2019.v8.i2Sa.25246

Abstract
In recent years, the integration of Artificial Intelligence (AI) in healthcare has shown remarkable potential for enhancing medical diagnosis and treatment. However, the opacity of complex AI models poses a significant challenge in gaining trust and acceptance from healthcare practitioners and patients. The demand for transparency and interpretability in AI systems, particularly in the context of medical diagnosis, has led to the emergence of Explainable AI (XAI) as a critical area of research. This theoretical overview aims to explore and elucidate the role of interpretable models in healthcare, focusing on their applications and implications for medical diagnosis.

The first section of this review paper provides a comprehensive examination of the current landscape of AI in healthcare and the pivotal role it plays in medical diagnosis. With the increasing complexity of AI algorithms, the lack of interpretability becomes a bottleneck, hindering the widespread adoption of these technologies in clinical settings. In response to this challenge, interpretable models have gained prominence as they offer insights into the decision-making processes of AI systems.

Subsequently, the paper delves into various interpretable models employed in medical diagnosis, including rule-based systems, decision trees, and linear models. Each model is scrutinized for its strengths and limitations, highlighting the trade-offs between interpretability and predictive performance. The discussion extends to the importance of incorporating domain knowledge and expert input in the development of these models to ensure their clinical relevance and applicability.

Furthermore, the review addresses the ethical considerations associated with Explainable AI in healthcare, emphasizing the need for transparent and accountable AI systems. Balancing the interpretability of models with the protection of sensitive patient information is crucial to maintaining trust and adherence to privacy standards.

Keywords: Explainable AI, healthcare, medical diagnosis, interpretable models, artificial intelligence, ethical considerations

Introduction

The fusion of Artificial Intelligence (AI) with healthcare has emerged as a transformative force, promising groundbreaking advancements in medical diagnosis and treatment. The utilization of AI algorithms, particularly in the domain of medical diagnosis, holds great potential for improving diagnostic accuracy and patient outcomes. However, this promising trajectory is accompanied by a critical challenge – the inherent opacity of complex AI models. As healthcare practitioners and patients increasingly rely on AI-driven systems, the demand for transparency and interpretability becomes imperative to foster trust and acceptance.

In recent years, the application of AI in healthcare has witnessed unprecedented growth, with deep learning models, neural networks, and other sophisticated algorithms demonstrating remarkable predictive capabilities. These models, fueled by vast datasets and computational power, excel at recognizing patterns and making predictions, but their "black box" nature raises concerns about how decisions are reached. In the realm of medical diagnosis, where the stakes are high and decisions impact patient well-being, the lack of interpretability poses a significant barrier to the widespread adoption of AI technologies.

Explainable AI (XAI) emerges as a solution to this challenge, aiming to demystify the decision-making processes of complex AI models in healthcare. The primary objective is to provide healthcare practitioners and stakeholders with insights into why and how a particular diagnosis or recommendation is made. This theoretical overview seeks to explore the landscape of interpretable models within the healthcare domain, shedding light on their applications and implications for medical diagnosis.

The first facet to be addressed is the current landscape of AI in healthcare, where advanced algorithms have demonstrated their prowess in tasks such as image recognition, natural language processing, and predictive analytics. The surge in AI applications has sparked enthusiasm for its potential to revolutionize healthcare delivery. However, as the complexity of AI models increases, so does the challenge of understanding their decision processes. This has prompted a paradigm shift towards developing models that not only exhibit high predictive performance but also offer transparency and interpretability.

The subsequent focus of this review is on interpretable models for medical diagnosis, ranging from rule-based systems and decision trees to linear models. Each model is examined for its capacity to balance the dual objectives of interpretability and predictive accuracy. Moreover, the integration of domain knowledge and expert input is emphasized as a crucial aspect of developing interpretable models that align with the clinical realities of healthcare practice.

Ethical considerations form another critical dimension of this exploration. As AI systems become integral to medical decision-making, ensuring transparency and accountability is paramount. The paper discusses the ethical implications of Explainable AI in healthcare, highlighting the delicate balance between providing interpretable insights and safeguarding sensitive patient information.

In essence, this theoretical overview serves as a foundation for understanding the role of Explainable AI in healthcare, offering insights into the current landscape, interpretable models, and ethical considerations. As the journey towards AI-driven healthcare progresses, the integration of transparency and interpretability becomes not just a technological necessity but a fundamental ethical imperative for fostering trust and advancing patient-centered care.

Related Work

The quest for Explainable AI (XAI) in healthcare, specifically in the context of medical diagnosis, has spurred a substantial body of research aimed at unraveling the intricacies of interpretable models and their application in clinical settings. This section reviews key studies and developments that contribute significantly to the discourse on XAI in healthcare.

One notable avenue of research has explored the application of rule-based systems in medical diagnosis. Studies by Letham *et al*. (2015) [11] and Caruana *et al*. (2015) [13] have demonstrated the efficacy of rule-based models in providing transparent decision rules, enabling healthcare practitioners to comprehend and trust the decision-making processes. These models, often referred to as "white box" models, prioritize interpretability without compromising predictive performance, making them particularly appealing for medical applications.

Decision trees have also emerged as a prominent focus in XAI research within healthcare. Noteworthy contributions by Chen *et al*. (2018) [14] and Lou *et al*. (2012) [15] showcase the potential of decision trees in generating interpretable models for disease classification and prognosis. Decision trees offer a hierarchical structure that mirrors decision pathways, enabling healthcare professionals to follow and understand the logic behind predictions, ultimately fostering greater trust in AI-generated diagnoses.

In the realm of linear models, the work by Ribeiro *et al*. (2016) [16] has been influential in developing interpretable models for healthcare applications. Their Local Interpretable Model-agnostic Explanations (LIME) framework provides a systematic approach for explaining the predictions of complex models, thereby enhancing the transparency of AI systems. This work underscores the significance of not only adopting interpretable models but also implementing post-hoc interpretability techniques to enhance the explainability of black-box models.

Furthermore, ethical considerations in the deployment of XAI in healthcare have been investigated by researchers such as Mittelstadt *et al*. (2016) [17]. These studies emphasize the importance of addressing issues related to bias, fairness, and privacy when implementing interpretable models in the clinical setting. The ethical dimension is crucial for ensuring that XAI not only enhances diagnostic accuracy but also aligns with principles of patient autonomy, justice, and beneficence.

Methodology Review

Understanding the methodologies employed in the exploration of Explainable AI (XAI) in healthcare, particularly in the context of medical diagnosis, is crucial for appreciating the advancements and challenges in this evolving field. This section presents a comprehensive review of the methodologies utilized in seminal studies, categorizing them into distinct subtopics for clarity.

Model Architectures and Design

The exploration of interpretable models in medical diagnosis often begins with the selection and design of model architectures. Research by Holzinger *et al*. (2017) [18] has delved into the development of rule-based systems, outlining how domain knowledge is incorporated into the model design. The iterative process of refining rules to balance accuracy and interpretability is a common theme, highlighting the importance of a principled approach to model architecture.

Decision Trees and Ensembles

Decision trees and ensemble methods have garnered significant attention in the quest for interpretable AI models. Studies by Lou *et al*. (2012) [15] and Chen *et al*. (2018) [14] have employed decision trees to capture hierarchical decision pathways in medical diagnoses. The review of methodologies in these studies emphasizes the construction of decision trees, feature selection techniques, and the ensemble of multiple trees to enhance predictive performance while maintaining interpretability.

Linear Models and Post-hoc Interpretability

Linear models, known for their simplicity and interpretability, have been investigated by Ribeiro *et al*. (2016) [16]. These studies showcase the importance of linear models in medical diagnosis and the integration of post-hoc interpretability techniques. The methodologies encompass techniques like Local Interpretable Model-agnostic Explanations (LIME), which facilitate the explanation of black-box model predictions.

Ethical Frameworks and Considerations

Exploring the ethical dimensions of XAI in healthcare is a critical facet of the methodology review. Mittelstadt *et al*. (2016) [17] provide insights into the methodologies employed for evaluating the ethical implications of interpretable models. This includes the identification of potential biases, fairness

assessments, and considerations of patient privacy, which are integrated into the model development process.

User-Centered Design and Evaluation:
Recognizing the importance of user acceptance, studies such as Letham *et al*. (2015) [11] and Caruana *et al*. (2015) [13] incorporate user-centered design principles into the methodology. These methodologies involve iterative feedback loops with healthcare practitioners to ensure that interpretable models align with their decision-making processes. Usability testing and qualitative assessments are integral components of these studies, ensuring that the developed models meet the practical needs of end-users.

Datasets and Evaluation Metrics
The choice of datasets and evaluation metrics significantly influences the effectiveness and generalizability of interpretable models. Researchers, including Chen *et al*. (2018) [14] and Holzinger *et al*. (2017) [18], discuss the methodologies employed for selecting representative datasets and the use of metrics that capture both predictive performance and interpretability. This includes metrics that assess the transparency and comprehensibility of the models in addition to traditional performance metrics.

Interpretable Deep Learning Architectures
As deep learning models gain prominence in medical diagnosis, methodologies for enhancing their interpretability become essential. Research by Samek *et al*. (2017) focuses on developing deep learning architectures that provide meaningful insights into decision-making processes. The subtopic explores techniques such as attention mechanisms, layer-wise relevance propagation, and gradient-based methods to visualize and interpret the features learned by deep neural networks.

Integration of Domain Expertise and Knowledge Elicitation
Acknowledging the importance of domain expertise, methodologies that involve the collaboration of AI researchers and healthcare domain experts become crucial. Studies by Shickel *et al*. (2018) [19] emphasize the integration of domain knowledge into the model development process. This subtopic reviews the methodologies employed for knowledge elicitation, expert consultations, and the incorporation of relevant medical guidelines, ensuring that interpretable models align with the nuanced expertise of healthcare professionals.

Explanations in Natural Language
Enhancing the interpretability of AI models involves not only providing visual representations but also generating explanations in natural language. Research by Ribeiro *et al*. (2016) [16] explores methodologies for generating human-readable explanations that articulate the reasoning behind model predictions. The subtopic delves into the use of techniques such as text generation and summarization to create explanations that are not only interpretable but also accessible to healthcare practitioners and patients.

Future Outlook
The trajectory of Explainable AI (XAI) in healthcare, particularly in the domain of medical diagnosis, points towards a future where interpretability and transparency will be integral components of AI systems. Several key areas are likely to shape the future outlook of XAI in healthcare, addressing current limitations and opening new avenues for research and application.

Hybrid Models and Ensemble Approaches
The future of XAI in medical diagnosis may witness the rise of hybrid models and ensemble approaches that combine the strengths of different interpretability techniques. Integrating rule-based systems with deep learning architectures or combining decision trees with post-hoc interpretability methods could offer a synergistic approach, leveraging the advantages of multiple models to enhance both accuracy and interpretability.

Explainability Metrics and Standardization:
As the field matures, the development of standardized metrics for evaluating the explainability of AI models will become imperative. Future research may focus on defining and refining metrics that go beyond traditional performance measures, encompassing aspects of human interpretability and trust. This standardization will facilitate comparative assessments of different XAI techniques and promote the adoption of best practices in model evaluation.

Human-AI Collaboration and User-Centered Design
The collaboration between AI systems and healthcare practitioners is poised to evolve, with an increased emphasis on user-centered design principles. Future XAI models will likely involve continuous feedback loops with end-users, incorporating insights from healthcare professionals to enhance model interpretability and usability. This collaborative approach ensures that XAI solutions align seamlessly with the workflows and decision-making processes of healthcare practitioners.

Ethical Considerations and Bias Mitigation
The ethical dimensions of XAI in healthcare will remain a critical focus. Future research is expected to delve deeper into methodologies for identifying and mitigating biases in interpretable models. Ensuring fairness and equity in AI-driven medical diagnosis, especially across diverse patient populations, will be a key priority to build trust and avoid exacerbating existing healthcare disparities.

Interpretability in Regulatory Frameworks
The integration of interpretability requirements into regulatory frameworks for healthcare AI systems is likely to become more pronounced. Future guidelines and standards may mandate a certain level of explainability, necessitating the development and validation of AI models that meet both performance and interpretability criteria. This shift will contribute to the responsible deployment of AI technologies in healthcare settings.

Difference between Past and Future Applications of Explainable AI in Healthcare

Past Applications
In the past, the application of Explainable AI (XAI) in healthcare primarily focused on establishing proof-of-concept for interpretable models and addressing the immediate need for transparency in AI-driven medical diagnosis. Initial endeavors concentrated on developing interpretable models such as rule-based systems, decision trees, and linear models

to demystify the decision-making processes of complex algorithms. Researchers sought to demonstrate the feasibility of creating models that not only achieved competitive predictive performance but also provided insights into the rationale behind their predictions.

The past applications were characterized by a foundational exploration of methodologies and a gradual shift toward understanding the trade-offs between accuracy and interpretability. Researchers and practitioners recognized the need to bridge the gap between the inherently opaque nature of advanced AI models and the requirements of healthcare professionals who demand transparency in order to trust and adopt these technologies. Early studies laid the groundwork for methodologies, ethical considerations, and user-centered design principles that paved the way for the integration of XAI in healthcare.

Future Applications

Looking ahead, the application of XAI in healthcare is poised to evolve with a focus on addressing current limitations and advancing the field to new frontiers. One notable shift is towards the integration of hybrid models and ensemble approaches that combine the strengths of different interpretable techniques. Future applications are expected to leverage synergies between rule-based systems, decision trees, and more complex deep learning architectures to create hybrid models that optimize both interpretability and predictive accuracy.

The future also holds the promise of standardized metrics for evaluating the explainability of AI models. As the field matures, there is an increasing recognition of the need for consistent and comprehensive metrics that go beyond traditional performance measures. Standardization will facilitate a more systematic comparison of different XAI techniques and contribute to the establishment of best practices in the evaluation and deployment of interpretable models.

Moreover, the future applications of XAI in healthcare will place a greater emphasis on human-AI collaboration and user-centered design. The ongoing collaboration between AI systems and healthcare practitioners is anticipated to involve continuous feedback loops, ensuring that interpretable models not only meet the technical requirements but also align seamlessly with the workflows and decision-making processes of end-users.

Conclusion

In conclusion, the journey of Explainable AI (XAI) in healthcare, particularly in the realm of medical diagnosis, reflects a transformative evolution from foundational exploration to a more nuanced and sophisticated approach. Past applications were characterized by the establishment of proof-of-concept for interpretable models, emphasizing the development of rule-based systems, decision trees, and linear models to address the immediate need for transparency. This initial phase laid the groundwork for methodologies, ethical considerations, and user-centered design principles that formed the bedrock of subsequent research.

Looking ahead, future applications of XAI in healthcare promise to elevate the field to new heights. The integration of hybrid models and ensemble approaches signifies a shift towards optimizing both interpretability and predictive accuracy, leveraging the strengths of various techniques. The anticipated introduction of standardized metrics for evaluating explainability reflects a maturation of the field, providing a systematic framework for comparing different XAI methodologies. Moreover, the future places a heightened emphasis on human-AI collaboration and user-centered design, acknowledging the pivotal role of healthcare practitioners in the development and acceptance of interpretable models.

As XAI continues to evolve, the convergence of diverse interpretability techniques, adherence to ethical considerations, and the establishment of user-friendly models position it as an integral component in the ongoing advancement of AI-driven healthcare solutions. The journey from past applications to future outlook signifies not only a progression in technical capabilities but also a commitment to ensuring that XAI aligns seamlessly with the practical and ethical considerations of healthcare practice. Ultimately, the continued integration of transparency and interpretability into AI systems holds the promise of fostering trust, improving patient outcomes, and shaping a more accountable and accessible future for healthcare.

References

1. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2:719-731.
2. Keet CM. Granular computing. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY, USA: Springer; c2013. p. 849.
3. HHS Office for Civil Rights. Standards for privacy of individually identifiable health information-Final rule. Fed Regist. 2002;67:53181-53273.
4. Lipton ZC. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. Queue. 2018;16:31-57.
5. Klein G, Hoffman RR. Macrocognition, mental models, and cognitive task analysis methodology. In: Naturalistic Decision Making and Macrocognition. Farnham, UK: Ashgate Publishing; c2008. p. 57-80.
6. Kaushik P, Yadav R. Reliability design protocol and blockchain locating technique for mobile agent. J Adv Sci Technol (JAST). 2017;14(1):136-141. https://doi.org/10.29070/JAST
7. Kaushik P, Yadav R. Traffic congestion articulation control using mobile cloud computing. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(1):1439-1442. https://doi.org/10.29070/JASRAE
8. Kaushik P, Yadav R. Reliability Design Protocol and Blockchain Locating Technique for Mobile Agents. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(6):590-595. https://doi.org/10.29070/JASRAE
9. Kaushik P, Yadav R. Deployment of Location Management Protocol and Fault Tolerant Technique for Mobile Agents. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(6):590-595.
10. https://doi.org/10.29070/JASRAE
11. Kaushik P, Yadav R. Mobile Image Vision and Image Processing Reliability Design for Fault-Free Tolerance in Traffic Jam. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(6):606-611. https://doi.org/10.29070/JASRAE
12. Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model; c2015.
13. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P.

Longitudinal studies. Journal of thoracic disease. 2015 Nov;7(11):E537.
14. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. InProceedings of the European conference on computer vision (ECCV); c2018 p. 801-818.
15. Lou Q, Guo ZL, Shi BC. Effects of force discretization on mass conservation in lattice Boltzmann equation for two-phase flows. Europhysics Letters. 2012 Oct 5;99(6):64005.
16. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386; c2016 Jun 16.
17. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. Big Data & Society. 2016 Nov;3(2):2053951716679679.
18. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923; c2017 Dec 28.
19. Shickel B, Loftus TJ, Ozrazgat-Baslanti T, Ebadi A, Bihorac A, Rashidi P. DeepSOFA: a real-time continuous acuity score framework using deep learning. ArXiv e-prints; c2018;1802.