Explainable AI and interpretable Machine Learning: Cementing the gap between Complexity and Transparency

By: Aamir Khan, Abhishek Acharya, Allen Ho, Ankit Paudel

Abstract

Artificial Intelligence, specifically the ones based on black-box models are rapidly being integrated into high-stake domains such as finance, healthcare, automation and other socio-political as well as economical aspects. Due to their questionable interpretability, transparency,

accountability and reliability raises numerous ethical concerns as well. In the following report, we have explored the importance of XAI and interpretable machine learning techniques such as SHAP, LIME and other methodologies that help cement the gap between human understanding and data-driven decision making. Furthermore in this report we dive into ethical aspects and implications, trade-offs between accuracy and explainability and the role of XAI and the bias created by AI as a whole as well as federated learning that emphasizes on user's privacy concerns, scalability and the challenges it poses.

Table of Contents:

Introduction

Artificial Intelligence systems based on the blackbox models are rapidly evolving and undertaking various high-stake tasks, seemingly integrating in our day to day lives. Ultimately creating a vacuum for interpretability, transparency and accountability for these system. The

increasing dependency on these AI systems based on blackbox models on machine driven decision-making has made the understanding how these system come to a solution, importance and working mechanism of these systems crucial to be studied. Especially when the results caused by these predictions and outcomes have a major impact on real world scenarios. Like how an AI powered hired tool developed by Amazon discriminated against women by favoring male candidates as it's training model was based on the resumes submitted by men over the past decade, ultimately the AI system downgraded the CVs with words like 'women's-chess clubs' or how AI bias in healthcare was less likely to recommend patients of Black or Asian origin for additional medical care than white patients with similar medical conditions.

In this report we will look at a brief overview of AI systems, BlackBox models, their working methodology and policies to implement,

Literary Overview

The book Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (Springer Nature, 2019) by W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller looks into methods to make deep learning models more transparent and interpretable as AI systems are getting complex, a proper insight in their decision-making method becomes pivotal for reliability, transparency and their accountability.

From the book Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (Springer Nature, 2019) we get a technical overview of methods such as Layer-wise Relevance Propagation(LRP), attention mechanisms and feature visualization. All of which help explain their neural network predictions. Furthermore it looks in challenges within interpretability, trade-offs between accuracy and explainability and also dives in the ethical domain of AI led decision making.

Real life Case

Explainable AI has been undertaking various high-stake jobs, like in healthcare sector, Explainable AI was pivotal for studying the gut microbiota, a complex community of microorganisms that thrive in the gastronomical tract and plays a vital role in digestion, maintaining a balanced immune system and fighting pathogens. Further looking into how dysbiosis or a dynamic imbalance in the gut microbiota is responsible for various health diseases such as diabetes, obesity or even Alzheimer's disease.

Result and Solution

By providing a transparent insights into how AI models arrive at specific solutions, clinicians and researchers can understand the underlying mechanisms and ensure the predictions or diagnostics are not only accurate but also logical. Explainable AI techniques such as feature importance visualization or decision pathway explanations help the stakeholders to validate the solutions given by the AI, looking at the potential biases and optimize the reliability of the models, likely being helpful for interpreting AI-based analysis for data-driven decision creating confidence amongst practioners and introducing and integrating the AI system into healthcare sector. Bridging the gap between advanced AI capabilities and a practical and accountable insight to ensure a protected and effective healthcare solution

Focus and vision

Highlighting both theoretical foundations and practical applications serves a valuable resource for researchers and practiconers in AI, aiding them to design, create and implement models that are not only powerful enough to perform complex calculations but are also interpretable. The primary focuses on the visualization techniques and real-world case studies further aid in building the reader's understanding. Overall playing a crucial role on the fields of Explainable AI(XAI) which is necessary for building trusts in AI systems, mainly in critical and rapidly growing areas like healthcare, finances and automation.

Methodology

In his book Interpretable Machine Learning, Molnar dives into a more comprehensive guide for a better understanding and interpretation of complex machine learning models. AI systems are rapidly evolving and undertaking high-stake domains such as healthcare, criminal justice, finance and automation. A model transparency is critical to make turst accountable for. Furthermore, we can get ideas for making machine learning models more interpretable and efficient, ensuring further exploration of numerous vast techniques to make machine learning more efficient, interpretable all while maintaining predictive performance.

Monlar furthermore outlines techniques like using Simpler, Interpretable Models that rely on Decision Trees that are inherently interpretable since they provide a clear path from input to output. while also using Linear Models that utilizes Linear regression and logistic regression are transparent and easy to interpret, especially with regularization to prevent overfitting. Using Importance Analysis techniques such as Permutation Feature Importance which measures the impact of each feature on model predictions by shuffling its values and observing changes in

performance and SHAP (Shapley Additive Explanations) which is based on game theory, that assigns each feature a contribution score for every prediction, ensuring fairness and transparency. Using Local Interpretability Methods such as LIME (Local Interpretable Model-Agnostic Explanations) approximates complex models with simpler, interpretable models around individual predictions, making AI decisions more understandable and Counterfactual Explanations that outlines the bare minimum changes to input features would alter a model's decision, helping users understand decision boundaries. Implementation of Visualization Techniques like Partial Dependence Plots (PDPs) for highlighting how a feature affects predictions on average, helping understand non-linear relationships and ICE (Individual Conditional Expectation) which are similar to PDPs but provide insights at an individualistic instance level.

Monlar also mentions Regularization for Interpretability for Sparse Models by applying L1 regularization (Lasso) that aids in feature selection, in turn making the models more interpretable by focusing primarily on the most critical variables.

These techniques ensure that machine learning models remain both effective and explainable, thus being crucial for high-stakes applications.

Interpretability methods can be classified into two classes: an intrinsic interpretability model and a post-hoc interpretability model. the Intrinsic interpretability model is inherently understandable. Such as decision trees, linear regression and rule based models. Post-hoc interpretability applies to black-box models that require an external technique for explanation such as deep neural networks and ensemble methods.

By providing a structured approach to interpretability, Interpretable Machine Learning serves as a valuable resource for researchers, data scientists, and practitioners seeking to develop more transparent and accountable AI systems.

Federated Learning

Federated learning is a decentralized paradigm and method for machine learning that primarily focuses on enabling multiple devices across multiple platforms to simultaneously collaborate on a trained model without inter-transmitting raw data. This prevents violation of user's privacy and

cuts down communication overhead. Federated Learning has a growing importance in IoT ecosystems where the same data is oftentimes distributed across multiple devices, making centralized training impractical as it violates the privacy concerns, limitations on the bandwidth and resources constraint.

Challenges

The recent advancements in Federated learning for IoT are optimization techniques, efficient communications, user privacy policies and methods, preventing classification of existing learning approaches based on the architectures they use, learning new objectives and their domains.

Heterogeneity of IoT devices, scalability, resource allocation and security threats prove to be an open threat to fully utilize the potential of Federated learning.

Conclusion

While AI systems, especially the ones based on black-box models are impacting various fields of higher-stakes, ensuring their accuracy, interpretability and transparency is pivotal in this discipline. Methods and techniques like Explainable AI and interpretable machine learning models give a critical insight into data driven-decision-making processes, catering trust, transparency and fairness in both high-stake and day to day aspects.

References

[1]

W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, “Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,” *Google Books*, 2019. (accessed Mar. 14, 2025). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning - Google Books
[2]
P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, “From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies,” *Mobile Information Systems*, vol. 2022, pp. 1–20, Jun. 2022, doi: https://doi.org/10.1155/2022/8167821.
[3]
Urvi, P. Sharma, K. Goyal, and S. Sharma, “Real-World Applications of Explainable AI in Healthcare,” *Explainable Artificial Intelligence in the Healthcare Industry*, pp. 451–466, Mar. 2025, doi: https://doi.org/10.1002/9781394249312.ch20.
[4]
K. A. Tahboub, “Human-Machine Coadaptation Based on Reinforcement Learning with Policy Gradients,” pp. 247–251, Oct. 2019, doi: https://doi.org/10.1109/icsc47195.2019.8950660.
[5]
S. Ali *et al.*, “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Information Fusion*, vol. 99, no. 101805, p. 101805, Apr. 2023, doi: https://doi.org/10.1016/j.inffus.2023.101805.
[6]
Christoph Molnar, “Interpretable Machine Learning,” *Github.io*, Aug. 27, 2019. https://christophm.github.io/interpretable-ml-book/

.