MANTECH Publications
Recent Trends in Computer Science and Software Technology
Volume 2, Issue 1, January-April 2017

Explainable AI Models for Transparent Decision-Making

Dr. V. Ramesh
Associate Professor
Department of Computer Science and Engineering
Meenakshi Institute of Technology, Tirunelveli, Tamil Nadu
Email: *drv.ramesh@yahoo.co.in*

Ms. Divya M.
B.Tech Student
Department of Artificial Intelligence
Sri Krishna College of Engineering and Technology, Dindigul, Tamil Nadu
Email: *divyam.tech@rocketmail.com*

Abstract

Explainable Artificial Intelligence (XAI) addresses the growing need for transparency in AI decision-making systems. As AI systems are increasingly deployed in critical domains such as healthcare, finance, and legal sectors, understanding the rationale behind their decisions becomes imperative. This paper explores recent developments in interpretable models, post-hoc explanation techniques, and evaluation methods. It emphasizes the balance between model accuracy and interpretability and provides real-world examples showcasing the necessity of XAI in high-stakes applications.

INTRODUCTION

Artificial Intelligence (AI) systems have rapidly become integral to many industries, offering automated solutions that rival or exceed human performance. However, the black-box nature of many AI models, particularly deep learning systems, raises concerns about the accountability and trustworthiness of their outputs. Explainable AI (XAI) has emerged as a field focused on creating interpretable systems or explaining the decisions of complex models. It addresses the critical need to understand

'why' and 'how' an AI system makes a decision, especially in applications involving ethical, legal, or life-critical outcomes.

NEED FOR EXPLAINABILITY

In many real-world applications such as medical diagnosis, loan approval, and criminal justice, stakeholders require more than just a prediction. They demand comprehensible justifications to trust the model's outcomes. Furthermore, regulatory frameworks such as GDPR emphasize the 'right to explanation,' mandating organizations to explain automated decisions. XAI thus becomes essential for model validation, bias detection, fairness, and regulatory compliance.

TYPES OF EXPLAINABLE MODELS

Intrinsic Interpretable Models

These are models that are inherently understandable by humans. Examples include decision trees, linear regression, and rule-based systems. Although they are interpretable, they often lack the predictive power of complex models like neural networks. They are suitable for scenarios where interpretability is more critical than accuracy.

Post-Hoc Explanation Techniques

Post-hoc explanations are methods used to interpret black-box models after training. Popular techniques include Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and Partial Dependence Plots (PDP). These methods approximate how input features affect model predictions and help users understand model behavior.

MODEL EVALUATION IN XAI

Evaluating the quality of explanations is a major challenge. The effectiveness of XAI is typically assessed using criteria such as fidelity, consistency, stability, and human understandability. Metrics like comprehensibility scores and explanation satisfaction ratings are also used in user studies to validate the impact of XAI in practical applications.

APPLICATIONS AND CASE STUDIES

1. **Healthcare**: In medical imaging, XAI models help radiologists understand which image regions contributed to a diagnosis.
2. **Finance**: Credit scoring models utilize SHAP values to highlight which customer features influence loan decisions.
3. **Autonomous Vehicles**: XAI is applied to explain why a vehicle chose a particular action, enhancing trust in its decision-making during critical scenarios.
4. **HR and Recruitment**: XAI is used to audit algorithmic bias in hiring tools and ensure fairness across different demographics.

Table 1: Overview and comparison of widely used XAI methods. Table height fixed at approximately 0.9 cm for uniformity.

Technique	Type	Strengths
LIME	Post-Hoc	Local explanations, model-agnostic
SHAP	Post-Hoc	Consistent, interpretable contributions
Decision Trees	Intrinsic	Human-readable, globally interpretable
PDP	Post-Hoc	Visualizes feature influence

CHALLENGES IN EXPLAINABLE AI

Despite significant progress, XAI faces challenges such as scalability, explanation faithfulness, and usability. Balancing accuracy and interpretability is complex, as more explainable models tend to be less performant. Moreover, explanations must be adapted to diverse user needs, ranging from technical developers to non-expert end-users.

CONCLUSION

Explainable AI is not just a technical feature but a necessity in the ethical deployment of AI systems. From regulatory compliance to public trust, transparency enhances the broader acceptance of AI. Future advancements will likely focus on hybrid models that offer both high accuracy and interpretability. As AI continues to evolve, XAI will play a pivotal role in ensuring responsible and accountable AI adoption.

REFERENCES

1. [1] Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD.
2. Lundberg, S.M., & Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems.
3. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.
4. Holzinger, A. et al. (2019). What do we need to build explainable AI systems for the medical domain? Review and perspectives. Methods of Information in Medicine.
5. Caruana, R. et al. (2015). Intelligible Models for Healthcare. Proceedings of the 21st ACM SIGKDD.
6. Gilpin, L. et al. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. IEEE Big Data.
7. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/
8. Goodfellow, I. et al. (2016). Deep Learning. MIT Press.