Interpretable Machine Learning Bridging Performance and Transparency in AI Systems

Gusti Muhamad Sardana
http://orcid.org/0009-0001-4281-2432
Universitas Esa Unggul, Indonesia

ABSTRACT

Advances in artificial intelligence (AI) technology, particularly through machine learning and deep learning, have resulted in systems that are highly accurate but often difficult to understand. This article explains the importance of interpretability in AI, especially in the context of critical applications such as healthcare, transportation, and public services. Highlighting the differences between inherently interpretable and "black-box" models, the article reviews methods such as SHAP, LIME, and counterfactual explanations to enhance transparency. It also discusses the challenges of implementing interpretability in smart city environments based on IoT and edge computing, as well as its relation to ethics, cybersecurity, and regulations such as GDPR. The article emphasizes that interpretability is not just a technical feature, but an ethical and social foundation for building fair, transparent, and trustworthy AI systems.

1. INTRODUCTION TO INTERPRETABLE MACHINE LEARNING: BRIDGING PERFORMANCE AND TRANSPARENCY IN AI SYSTEMS

The advancement of artificial intelligence technology, especially through machine learning and deep learning techniques, has created a system that is able to solve various complex problems with a high level of accuracy. This system has been widely used in various fields, such as disease detection in the medical world, object identification in digital images, sentiment analysis in social media, and in recommendation systems on e-commerce platforms. Despite its extraordinary performance, one major challenge that arises is the lack of transparency in decision-making by AI models, especially in complex models such as deep neural networks. Complex machine learning models such as Random Forest and deep learning are often considered "black-box" because it is difficult to understand how they make decisions. "A major limitation associated with most machine learning models is the fact that they are so-called 'black-boxes' one cannot understand how the decision was arrived at," so Explainable AI (XAI) methods are needed

DOI: 10.4018/407566

to increase transparency and user trust in critical systems such as traffic accident prediction. (Alotaibi, 2025; Javed et al., 2023).

Concerns about the closed nature of the decision-making process in AI systems have given rise to a field of study called interpretable machine learning or explainable AI. This field focuses on explaining the internal mechanisms of AI models to users, researchers, and policymakers so that the results provided by the system are not only accurate but also accountable. Interpretability becomes especially important when AI models are used in critical sectors such as health, finance, law, and government, where the resulting decisions can have a major impact on human life. In this context, transparency is not just an added value, but a moral and legal imperative. According to (Mathew et al., 2025), Explainable Artificial Intelligence (XAI) lies in its potential to bridge the gap between the complexity of advanced AI models and the need for human understanding and trust," making interpretability a crucial element in building socially acceptable and ethical AI systems.

Modern smart cities increasingly depend on AI-driven systems to manage transportation, public safety, energy distribution, citizen services, and urban governance. As these systems automate decisions that directly affect public welfare, the need for interpretable and accountable AI becomes critical. Interpretability provides a structured way to understand how models generate predictions, detect failures, and ensure that automated decisions remain transparent and aligned with regulatory, ethical, and societal expectations. In the context of smart city infrastructures where AI models operate within complex, real-time environments interpretability is not merely a technical requirement; it is fundamental to sustaining public trust, supporting cross-agency coordination, and enabling responsible deployment of data-driven urban intelligence. This chapter adopts smart cities as its central application domain, using it as the guiding thread through which interpretability concepts, methods, and challenges are contextualized.

Interpretable Machine Learning: Bridging Performance and Transparency in AI Systems takes a deep dive into the concepts, approaches, and practices of building humanly explainable AI systems. Early in the chapter, it defines interpretability as the ability to explain or provide meaningful understanding of a model's behavior. Interpretability can be viewed in terms of several dimensions, such as predictive accuracy, descriptive accuracy, and user relevance. Good interpretability not only explains the results, but also reliably shows how the model made decisions, and whether those explanations are relevant and understandable to users from a variety of backgrounds.

One of the main goals of interpretability in AI is to build trust between humans and intelligent systems. With good explanations, users can assess whether the results of a model are trustworthy and free from bias or systemic errors. In addition, interpretability aids the debugging process, allowing developers to understand and correct unexpected model behavior. In the context of law and regulation, interpretability is especially important, especially with policies such as the GDPR in Europe that give individuals the right to know how algorithmic decisions that affect them are made.

This book describes various approaches to achieving interpretability, both in explainable models and black-box models. For interpretable models, such as decision trees, rule-based learning, and generalized additive models, these techniques are considered sufficient. However, in black-box models, special explanation methods are needed that can provide local and global understanding of the predictions produced. Local approaches such as LIME, SHAP, and counterfactual explanations are very useful for explaining decisions on a single data instance. Meanwhile, global approaches such as partial dependence plots (PDP), accumulated local effects (ALE), and surrogate models aim to provide an overview of how inputs affect the overall model output.