IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, March 2025

Interpretable AI: Enhancing Transparency and Fairness in Decision-Making

Rajeev Nair[1], Rosemol Thomas[1], Sandra MV[1], Ms. Siji K B[2]
Students, MCA, Vidya Academy of Science and Technology Thalakkottukara, Thrissur, India[1]
Assistant Professor, Department of Computer Applications[2]
Vidya Academy of Science and Technology, Thalakkottukara, Thrissur, India

Abstract: *Interpretable Artificial Intelligence aims to make machine learning models more transparent, interpretable, and accountable, addressing the "black box" nature of traditional AI systems. As AI plays a critical role in high-stakes domains like healthcare, finance, and autonomous systems, ensuring trust and fairness in decision-making has become essential and this paper also explores key techniques in AI.*
This study adopts a mixed-methods approach to analyse, evaluate, and compare XAI techniques across key domains. This research examines a three-phase approach in XAI, focusing on exploring different methods, evaluating their impact in real- world applications, and analysing the trade-offs between interpretability and model performance. XAI enhances transparency, bias detection, and user trust, but still it faces challenges, such as the trade-off between interpretability and accuracy, as well as computational complexity. Future studies focus on improving model interpretability, enhancing human-AI interaction, and promoting fairness in AI-driven decisions.ncy.

Keywords: Explainable Artificial Intelligence , Transparency & Trust , LIME & SHAP , Model-Agnostic Methods, Gradient-Based Methods , Propagation-Based Methods , Meta- Explanations , Accuracy vs. Simplicity , High-Stakes Domains