IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

IJARSCT

ISSN: 2581-9429

Volume 5, Issue 2, May 2025

Impact Factor: 7.67

Enhancing Model Interpretability: A Study on Explainable Machine Learning Techniques

Anil Kumar Chikatimarla[1], Katyayini Gona[2], Teja Sri Oleti[3]

Head, Department of Computer Science[1]

Lecturer, Department of Computer Science[2,3]

A.G. & S.G. Siddhartha Degree College of Arts & Science, Vuyyuru, Andhra Pradesh, India

Abstract: *The importance of machine learning models that are simple for humans to comprehend and apply has grown in fields such as healthcare, autonomous systems, and finance, where people's lives are at stake. Despite these models' effectiveness, user trust and regulatory compliance will be significantly diminished due to their opaque nature. Four state-of-the-art XAI methods—LIME, SHAP, PDP, and CFLE—are compared in this article. We evaluate the algorithms based on how well they express their computations in terms of clarity, efficiency, and accuracy. A group of data scientists and domain experts conducted a user-centered assessment to test these approaches in an actual setting.Evidence from the findings shows that SHAP is accurate, but its computing cost is too high for real-time jobs. The complexity of the ideas is best shown by LIME, even if PDPs and Counterfactual Explanations seem simple at first glance. There is no silver bullet since accuracy and clarity are not mutually exclusive. While considering how to make XAI approaches more relevant, it is crucial to consider potential future research objectives for the field. Here you may discover hybrid explainability approaches, area-specific evaluations, and explanations that happen in real time. This study aims to analyze the current and future of explainability approaches to make machine learning more interpretable.*

Keywords: Explainable Artificial Intelligence (XAI), Machine Learning Interpretability, LIME, SHAP, Model Transparency, Trust in AI

DOI: 10.48175/IJARSCT-26225