ID 368

# Review on Feasibility of Building an Explainable Artificial Intelligence Model for Anti-phishing Detection

YLDH Yakandawala[1#] and MKP Madushanka[1]

[1]Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

[#]38-cs-6229@kdu.ac.lk

## Abstract

Explainable Artificial Intelligence (XAI) is a field of Artificial Intelligence (AI) that aims to make AI models interpretable and transparent, allowing humans to understand the reasoning behind the decisions made by the AI system. XAI techniques provide insights into how the AI arrives at its conclusions, enhancing trust and usability. The viability of developing an XAI model for anti-phishing detection is examined in this review. The significance of XAI, its principles, methods/types, challenges, ethical issues, and vulnerability aspects are discussed. The areas of machine learning for phishing detection, XAI models for phishing detection, developing appropriate explanation messages for warnings, feasibility issues, and a comparison with conventional approaches are all covered. The importance of XAI in enhancing the clarity and interpretability of AI models are further emphasized in the paper. It shows different XAI techniques, difficulties in striking a balance between explainability and performance, and XAI ethics. The evaluation looks at phishing scams, machine learning detection methods, and the advantages of XAI models. It suggests a thorough strategy for conveying explanatory messages and examines the viability of creating XAI models. In highlighting the promise of XAI to improve transparency and interpretability, the research also acknowledges the difficulties that must be overcome in order to create scalable and reliable XAI models for anti-phishing detection.

**Keywords**: *Explainable Artificial Intelligence, Phishing, Anti-phishing detection, Cyber Security*