Group 02: Project Update 05

Rakibul Hasan Ridoy

1731339042

This week, I focused on reviewing several research papers and came across one titled P. Peisker et al., 'Evaluating Explainable AI for Deep Learning-Based Network Intrusion Detection,” arXiv, 2025. Our project aims to develop an interpretable ML pipeline for zero-day attack detection in IoT environments, while Peisker et al. (2025) evaluate XAI methods such as SHAP and LIME in deep learning–based intrusion detection systems. Their study provides a strong foundation for the explainability and benchmarking aspects of our project. Although their work primarily focuses on known-attack detection, their insights into XAI’s potential for generalising to unseen attacks directly support our objectives. We can leverage their findings as benchmarks for SHAP and LIME on datasets like NSL-KDD and CIC-IDS2017, and adopt their XAI evaluation metrics — fidelity (accuracy of explanations), stability (consistency across similar samples), and human interpretability (clarity and usefulness of explanations to users).