The research focuses on the increasing need for artificial intelligence (AI) techniques in intrusion detection systems (IDS) due to the rising number of network intrusions.

The utilization of explainable AI (XAI) methods is crucial for enhancing the transparency and interpretability of AI models in real-world IDS.

The study introduces an end-to-end framework for evaluating black-box XAI methods for network intrusion detection, aiming to provide insights into the strengths and limitations of these techniques.

**Motivation**

- **The lack of interpretability** in current AI-based IDS poses challenges for security analysts, requiring them to sift through large volumes of data to identify abnormal behaviors.
- **Achieving high accuracy** in results and providing comprehensible explanations for AI algorithms are **essential** in network intrusion detection to build trust and understanding among security analysts.
- **XAI frameworks** play a vital role in assisting analysts by elucidating AI decisions, facilitating efficient investigations, and ultimately saving time for human security analysts in network security tasks.

**Our Contribution**

- **The research proposes an end-to-end framework for evaluating XAI techniques for network intrusion detection tasks**, assessing both global and local explanations.
- **Six different evaluation metrics** are analyzed for two popular black-box XAI techniques, SHAP and LIME, under three network intrusion datasets and seven AI models.
- The **source codes** of the evaluation framework are released to the community, serving as a baseline for XAI evaluation in network intrusion detection and encouraging further development with new datasets and models.

**Framework**

- High-level: Blackbox AI, XAI Methods, XAI Evaluation Metrics
- Low-level: NSL-KDD, DNN, LIME, Shap, MLP, Feature

**Key Takeaways**

- **Creation of a novel end-to-end framework** for evaluating XAI techniques for network intrusion detection tasks, assessing both global and local explanations.

**Results**

- **Descriptive Accuracy**
- **Sparsity**
- **Stability**
- **Completeness**
- **Efficiency**

**Acknowledgement**

This work was supported by AnalytixN, Enhanced Mentoring Program with Opportunities for Ways to Excel in Research (EMPOWER); by the 1st Year Research Immersion Program (1RIP) Grants from the Office of the Vice Chancellor for Research at Indiana University–Purdue University Indianapolis; and in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.