Malicious Network Connection Detection using Explainable AI

Harsh Vishwakarma

Abstract

In today's digital landscape, safeguarding against malicious network connections is critical for maintaining the integrity and security of information systems. This paper proposes a lightweight, explainable AI-based tool to detect malicious network activity using a supervised machine learning approach. By leveraging feature importance techniques and standard classification models, the system can classify new network connections as benign or malicious with considerable accuracy. The model prioritizes interpretability through feature visualization, enabling analysts to understand the rationale behind each prediction. Our results demonstrate a practical balance between predictive performance and transparency, making this tool viable for use in cybersecurity operations.

Problem Statement & Objective

The exponential growth in internet traffic has resulted in a surge of cyberattacks, many of which manifest as malicious network connections. Traditional detection systems often lack the transparency needed for human validation. This research aims to develop a simple yet effective explainable AI (XAI) tool that detects malicious connections and provides clear, visual explanations for its predictions, aiding cybersecurity analysts in decision-making.

Literature Review

1. Doshi-Velez, F., & Kim, B. (2017). "Towards A Rigorous Science of Interpretable Machine Learning"
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" - introduces LIME for local model interpretability.
3. Lundberg, S. M., & Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions" - SHAP values for feature attribution.
4. Sommer, R., & Paxson, V. (2010). "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection"
5. Tavallaee, M., et al. (2009). "A detailed analysis of the KDD CUP 99 data set"
6. Shapira, B., Rokach, L., & Freilikhman, S. (2013). "Feature selection for anomaly detection in web traffic"
7. Chandola, V., Banerjee, A., & Kumar, V. (2009). "Anomaly Detection: A Survey"
8. Zuech, R., Khoshgoftaar, T. M., & Wald, R. (2015). "Intrusion detection and Big Heterogeneous Data"

9. Xie, Y., & Yu, S. (2009). "A large-scale hidden semi-Markov model for anomaly detection on user browsing behaviors"
10. Kim, G., Lee, S., & Kim, S. (2014). "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection".

Research Methodology

- Dataset: A simplified network traffic dataset containing fields such as src_bytes, dst_bytes, duration, flag, and protocol, with a binary malicious label.
- Preprocessing: Standardization of numerical features and encoding of categorical ones.
- Model: Random Forest Classifier trained on 80% of the dataset and tested on 20%.
- Explainability: Feature importance plotted to identify which variables influenced each prediction.
- Evaluation: Accuracy, precision, recall, and F1-score were used as evaluation metrics.

Tool Implementation

The tool is implemented in Python using scikit-learn, pandas, seaborn, and matplotlib. After training on labeled connection data, the model classifies new connection records and outputs whether they are benign or malicious. Additionally, it generates a feature importance graph to visually explain the decision. The tool can be run on any system with basic Python setup.

Results & Observations

The trained model achieved an accuracy of over 95% on the test dataset. Feature importance analysis revealed that src_bytes, dst_bytes, and duration were the most influential features in classification. The tool effectively demonstrated transparency by showing why a connection was flagged, helping human analysts verify its decisions.

Ethical Impact & Market Relevance

Ethically, explainable AI reduces the risk of black-box decision-making in security, promoting responsible AI deployment. From a market perspective, cybersecurity tools with explainability features are gaining traction, especially in sectors where compliance and transparency are critical, such as finance, healthcare, and government.

Future Scope

- Extend the tool to handle real-time data streams.
- Integrate more sophisticated XAI techniques like SHAP or LIME.
- Expand the feature set and improve dataset diversity.
- Deploy as a web-based dashboard or integrate into existing SIEM platforms.

References

[See Literature Review section for full references list.]