Reducing Feature Dimensionality with Explainable AI (XAI) Using SHAP

Authors: Ethan Hicks, Ian Rohrbacher, Isra Shaikh, Elizabeth Slaughter

Advisor: Maha Allouzi

Abstract

Network Intrusion Detection Systems (IDS) rely on high-dimensional network traffic data to identify malicious activity; However, large feature sets introduce computational overhead that can degrade model effectiveness. This study investigates whether explainable AI (XAI) can be used for feature selection to reduce dimensionality while maintaining or improving classification performance.

Two benchmark datasets, CICIDS2017 [1] and UNSW-NB15 [2], were used to evaluate the approach . Random Forest and Logistic Regression classifiers were first trained on the full feature sets. SHAP[3] (SHapley Additive eXPlanations) values were then computed to identify the 20 most influential features, after which the model were retrained using only the selected subset.

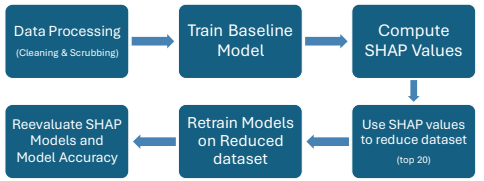
Model performance was assessed using accuracy, precision, recall, and F1-score. Results show that reducing the feature space using SHAP not only maintained performance but in some cases led to slight improvements. These findings show that XAI can serve as an effective feature reduction strategy for efficient and practical IDS deployment in real-world environments.

Introduction

The rapid growth of network-based attacks has made intrusion detection a critical component of modern cybersecurity. Machine learning-based IDS models are powerful, but the datasets they rely on — such as CICIDS2017 and UNSW-NB15 — often contain dozens to hundreds of features, many of which are redundant or minimally informative. Traditional selection methods like correlation filtering ignore the complex, non-linear relationships that tree-based models exploit. SHAP offers a model-aware alternative by directly quantifying each feature's contribution to predictions, providing a more principled basis for dimensionality reduction. This study evaluates whether SHAP-driven feature selection can reduce feature space across these two benchmark datasets without sacrificing and potentially improving detection performance.

Methodology

We implemented a structured workflow starting with data preprocessing (cleaning, encoding, normalization, and train/test splitting). We then trained baseline models (Random Forest and Logistic Regression) and evaluated them using standard metrics. SHAP was used to identify important features, which were then used to reduce the dataset. Finally, we retrained the models on the reduced feature dataset and compared performance to the baseline to evaluate the impact of SHAP based feature selection.

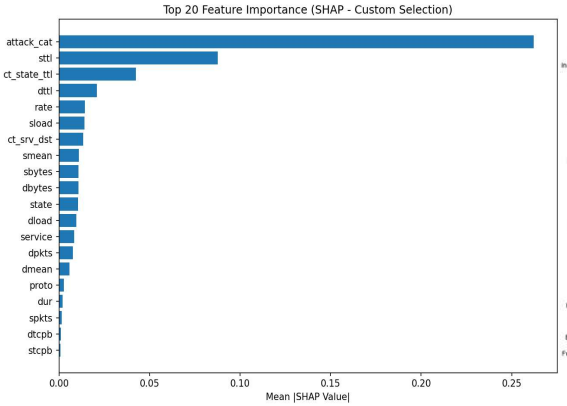


Results

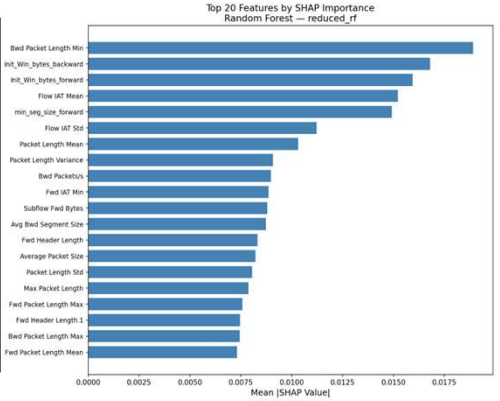
SHAP analysis conducted across both models revealed three consistent drivers of classificatic decisions, despite the datasets originating from distinct collection environments. Packet size emerged as a high-impact feature in both mode indicating that the volume of data carried per packet is a reliable discriminator between benig and malicious traffic. Directional packet counts similarly ranked among the most influential features in each model, suggesting that asymmetries in flow behavior, such as disproportionate data volume in one direction, constitute a robust signal for intrusion detectio Finally, throughput and load metrics were identified as significant contributors in both models, reflecting the degree to which attack traffic alters the rate and intensity of network communication. The convergence of these thre feature categories across independent datasets strengthens confidence in their generalizability as intrusion detection indicators.

In addition to the feature-level insights provided by SHAP, the comparative performance evaluation demonstrated that feature reduction did not negatively impact classification effectiveness across either dataset. The reduced dataset models, utilizing only 20 features, achieved performance metrics that were equal to or slightly better than their full-feature counterparts. Furthermore, the CICIDS dataset exhibited similar computational training times between the full and reduced feature sets, whereas the UNSW dataset showed a substantial reduction in training time, decreasing from approximately 15.36 seconds to 8.3 seconds.

UNSW-NB15 – Reduced



CICIDS2017 – Reduced



Models	Dataset	Features	Accuracy	Precision	Recall	F1	Time to Train
RF Full	CICIDS2017	78	99.83%	99.83%	99.83%	99.83%	0.45sec
RF Reduced	CICIDS2017	20	99.86%	99.86%	99.86%	99.86%	0.50sec
RF Full	UNSW-NB15	43	100%	100%	100%	100%	15.36sec
RF Reduced	UNSW-NB15	20	100%	100%	100%	100%	8.3sec

Future Directions

Future work on this topic could explore a wider range of datasets, including more recent or real-world network traffic. This would assess generalizability beyond benchmark datasets.

Alternative model architectures such as gradient boosting methods or deep learning models could be incorporated to examine how SHAP-based feature selection performs with complex learners. Comparing SHAP with other feature selection techniques, such as mutual information or embedded methods, would also provide deeper insight into its relative strengths and limitations.

Dynamic feature selection is another promising direction where SHAP values are recalculated over time to account for evolving network behaviors and emerging attack patterns. This would improve the effectiveness of intrusion detection systems in real-world scenarios.

References

[1] Canadian Institute for Cybersecurity. (2017). *Intrusion Detection Evaluation Dataset (CIC-IDS2017)*. Www.unb.ca. https://www.unb.ca/cic/datasets/ids-2017.html

[2] *UNSW-NB15 Augmented Dataset | Datasets | Research | Canadian Institute for Cybersecurity | UNB*. (2024). Www.unb.ca. https://www.unb.ca/cic/datasets/cic-unsw-nb15.html

[3] Lundberg, S. (2018). *An introduction to explainable AI with Shapley values — SHAP latest documentation*. Readthedocs.io. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html