A Data-Driven Solution for Improving Transferability of Traffic Flow Feature Selection

Pegah Golchin [*], Nima Rafiee [†], Ralf Kundel [*]
[*]*Multimedia Communications Lab (KOM), Technical University of Darmstadt, Germany*
[†]*Zalando, Berlin, Germany*
Contact: pegah.golchin@kom.tu-darmstadt.de

***Abstract*—The expansion of Internet connectivity has increased cyber threats in computer networks. Machine Learning (ML)-based Intrusion Detection Systems (IDS) have emerged as a promising candidate, leveraging ML models to analyze network traffic features and differentiate between malicious and benign flows. However, before using ML models, a crucial preprocessing step called feature selection is performed in ML-based IDS to identify the most relevant features that can enhance detection accuracy, streamline ML models, and reduce computational complexity. The selected features need to be transferable across diverse network traffic datasets, which is challenging due to variations in attack types, network architectures, and complex relationships among their flow features. In this work, we present a Data-Driven Ensemble Feature Selection (*DD-EFS*) to improve the transferability of the selected features across various network traffic datasets. Our results demonstrate an average increase in detection performance of up to 6.8%, 5.1%, and 4.3% across two distinct, previously unseen network traffic datasets for the Random Forest, Logistic Regression, and Multi-Layer Perceptron models, respectively.**

***Index Terms*—Intrusion Detection, Machine Learning**

I. Introduction

Nowadays, the demand for Network Intrusion Detection Systems (IDS) has surged, driven by the rise in cyber attacks [1]. Conventional IDSs, referred to as signature-based, relied on storing known attack signatures [2]. However, as new attacks emerge and zero-day attacks proliferate, these systems struggle to detect any attack not already cataloged in their databases. Conversely, Machine Learning (ML) models possess the ability to learn representations of network traffic flows. This allows them to differentiate between malicious attacks and benign flows, enabling detection not only of known attacks but also of those exhibiting different patterns from the learned representations [2].

To employ ML models effectively in IDS, training them on network traffic flow features is essential. Supervised ML models establish connections between statistical features of the available data and their corresponding labels (Benign/Attack). Employing a proper preprocessing pipeline can enhance their learning capacity and detection accuracy [3]. Feature selection techniques are particularly beneficial in this regard, as they trim down the noise in the training dataset by focusing on the most relevant features. This not only reduces the ML model's complexity but also facilitates faster model retraining when required.

ISBN 978-3-903176-63-8 ©2024 IFIP

However, network traffic datasets inherently exhibit heterogeneity due to their origin from distinct network architectures governed by unique management rules and routing protocols [4]. This inherent diversity leads to differences between datasets. Additionally, variations in attack types across different datasets contribute to discrepancies in the relationships between flow features. Consequently, these discrepancies pose challenges for feature selection approaches, particularly when focusing solely on one network traffic dataset. Features selected from one dataset may effectively capture patterns in network traffic flows with similar characteristics. However, given the potential for concept drift in network traffic and the emergence of new attack types [5], there is a likelihood of encountering flows with different patterns. Thus, it becomes imperative to select flow features that are transferable across various network traffic datasets.

To address this challenge, in this work, we show a Data-Driven Ensemble Feature Selection (DD-EFS) approach, aiming to improve the transferability of selected relevant flow features when faced with new flows from diverse datasets.

II. Related Work

In [6], authors proposed an ensemble feature selection method that iteratively chose a selector with a replacement strategy based on their contribution to ensemble performance. Models were added to the ensemble if their averaged performance maximizes overall performance. In [3], the authors devised a multi-aspect feature selection method and created new sub-datasets, including each attack and benign flows separately, to extract the most relevant features and enhance the generalization of ML-IDS. However, their findings indicated successful results on similar flow patterns. Authors in [4] examined the issue of generalization in supervised learning methods' detection performance. They trained ML models on a limited number of network traffic datasets and evaluated the results on various datasets. However, their work did not address the preprocessing step of feature selection for supervised ML models and the issue of transferability of the selected features.

III. Case Study

Our proposed method, *DD-EFS*, employs a data-driven approach to an ensemble feature selection technique to enhance the transferability of selected flow features across diverse network traffic datasets.

A. Implementation of DD-EFS

To extract the most relevant flow features from each network traffic dataset utilizing *DD-EFS*, the following modules are required.

1) Utilized Network Traffic Dataset: To tackle the challenges arising from differences between network traffic datasets and propose a data-driven solution, we incorporated five different network traffic files: CICIDS17 [7], UNSW-NB [8], Botnet [9], SlowDoS [10], and CTU13 [11]. These selections were made based on the availability of their traffic files in pcap format and their inclusion of various types of attacks.

2) Flow Features Extraction: In the initial stage, we convert the network traffic files to a feature set using the NF-Stream tool [12]. This tool can extract various types of flow features, including post-mortem, time-related, architecture-related, and statistical features, resulting in a total of 88 flow features. In this study, flows are defined as a collection of received packets sharing the same 5-tuple information, comprising source and destination IP addresses, source and destination MAC addresses, and protocol [13].

3) Pruning Certain Flow Features: In the subsequent stage, we filter out certain features that may lead to information leakage or bias during the training of the ML model. These features encompass the 5-tuple features, as they could potentially reveal the ground-truth labels (Benign/Attack). Additionally, we exclude time-related features, as they tend to vary across different network architectures and traffic management scenarios. Lastly, features with zero variance are removed, as they do not provide meaningful information for training the ML model. This pruning process results in reducing feature dimension from 88 to 45 features.

4) Data-Driven Ensemble Feature Selection (DD-EFS): Our proposed ensemble feature selection approach incorporates three ML-based methods: Random Forest (RF) Gini impurity, Lasso-regularization Logistic Regression (LR), and Lasso-regularization Support Vector Machine (SVM). By employing multiple methods, the approach gains robustness, as each method may perform better with certain traffic datasets.

Random Forest's Gini Importance [14] provides interpretability, efficiency, and robustness. Quantitating feature impact on predictive performance enhances model interpretability and identifies key contributors to classification accuracy. The averaging of results across multiple decision trees enhances efficiency and reduces the risk of overfitting, ensuring resilience to noisy data and outliers. In Lasso-regularization LR, the Lasso regularization penalizes the absolute values of the coefficients associated with each feature, thereby promoting sparsity in the coefficient values. Consequently, some coefficients are driven to zero, resulting in automatic feature selection [15]. Lasso-regularized SVMs introduce an additional layer of complexity compared to traditional SVMs. While traditional SVMs focus on maximizing the margin between different classes and penalizing classification errors, Lasso-regularized SVMs penalize the absolute value of the weights [16]. This regularization technique is particularly effective for datasets where feature selection is crucial, such as high-dimensional spaces with numerous features but relatively few samples. The output of each feature selection approach is normalized to ensure consistency in range. Then, the coefficients from each feature selection method are summed for each flow feature, as depicted in Figure 1. The results reveal variations in the importance of flow features across different network traffic datasets, underscoring the uniqueness of the selected feature set for each dataset and the differences between them.

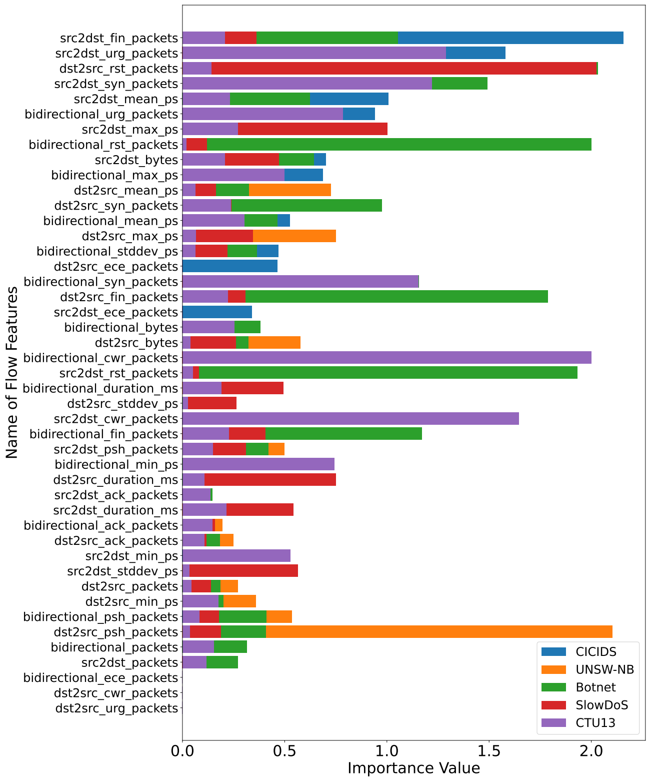

Fig. 1: Flow Feature importance for each individual network traffic dataset.

In this stage, we extract the top 25 features from 70% of three datasets: CICIDS17, UNSWNB, and CTU13. We then find the intersection of these features, resulting in the selection of five features: ”bidirectional maximum packet size”, ”bidirectional mean packet size”, ”source to destination maximum packet size”, ”destination to source SYN packets”, and ”bidirectional duration in milliseconds”. These features are considered as the output of the proposed *DD-EFS* approach.

B. Preliminary Evaluation Results

For extracting features of the proposed *DD-EFS*, we utilized 70% of the CICIDS17, UNSW-NB, and CTU-13 datasets. Meanwhile, the SlowDoS and Botnet datasets were excluded from this process and reserved to train on the output of *DD-EFS* (the 5 features explained in Section III-A). Three supervised ML approaches were utilized, including an RF model with 200 decision trees, an LR model with maximum iteration of 200, and a Multi-Layer Perceptron (MLP) model with four layers (20, 32, 20, and 1 neurons). They were

TABLE I: Investigating transferability of the selected features from *DD-EFS*. Here, DS1, DS2, and DS3 refer to CICIDS17, UNSW-NB, and CTU13 datasets, respectively.

Dataset	**F1-Score (%))**			
	5 top features of DS1	5 top features of DS2	5 top features of DS3	*DD-EFS* features
SlowDoS				
RF	96.3	98.1	76.8	**99.0**
LR	71.6	64.3	71.2	**72.7**
MLP	91.1	95.2	76.4	**95.5**
Botnet				
RF	85.9	82.7	70.4	**91.8**
LR	60.3	59.3	60.8	**64.8**
MLP	79.2	70.6	70.4	**83.2**

trained on the selected relevant features to evaluate detection performance. The detection performance is measured using macro-average F1-Score, which ensures equal consideration of each class, regardless of its frequency or potential imbalance in the dataset [3]. To perform the proposed *DD-EFS* and training models, we utilize an Ubuntu server equipped with 250GB RAM and 4 CPUs(Intel Core i9-10900K).

Transferability of extracted features from *DD-EFS*: To assess the transferability of selected features from *DD-EFS*, three ML models (RF, LR, and MLP) are trained using these features on two excluded network traffic datasets (SlowDoS and Botnet) as explained in Section III-B. These datasets possess potential differences compared to other utilized network traffic datasets. The obtained F1-Score is compared with the scenario in which these ML models are trained on the top five features extracted from each of the CICIDS17, UNSW-NB, and CTU13 datasets individually. This comparison highlights the potential limitations of other feature selection approaches that rely solely on a single dataset to extract flow features. Table I illustrates this comparison, demonstrating that features extracted from *DD-EFS* achieve higher detection performance when encountering new traffic flows that may differ from those used in training. This highlights the transferability of the extracted features from the *DD-EFS* approach (explained in Section III-A). The highest F1-Scores are indicated in bold.

IV. Conclusion

In this study, we highlight the challenges encountered by selected relevant features from a single network traffic dataset in supervised Machine Learning-based Intrusion Detection Systems (ML-IDSs). As network traffic datasets are heterogeneous and exhibit diverse distribution patterns, and contain different attack types, the relationships between their features may vary. Moreover, in real-world scenarios, ML-IDSs may encounter different traffic patterns due to concept drift in network traffic. Therefore, it is crucial not to rely solely on a single network traffic dataset for extracting the most relevant features. To address this challenge, we propose *DD-EFS*, a Data-Driven Ensemble Feature Selection approach, which demonstrates that selected features can achieve higher detection performance across various network traffic datasets using different supervised ML models. To expand on this study, we aim to assess how the inclusion of additional network traffic datasets influences the transferability of the selected features. Additionally, we will determine the optimal number of network traffic datasets needed to ensure the transferability of the extracted relevant features.

Acknowledgment

This work is funded by the Federal Ministry of Education and Research of Germany (BMBF) through the CELTIC-NEXT Flagship Project AI-NET-PROTECT and Software Campus Grant 01IS17050 (ML-based NIDS).

References

[1] "Cisco Annual Internet Report (2018–2023) ," (Accessed on 13.11.2023). [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] M. Hajizadeh, S. Barua, and P. Golchin, "Fsa-ids: A flow-based self-active intrusion detection system," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2023, pp. 1–9.

[3] P. Golchin, R. Kundel, T. Steuer, R. Hark, and R. Steinmetz, "Improving ddos attack detection leveraging a multi-aspect ensemble feature selection," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2022, pp. 1–5.

[4] L. D'hooge, T. Wauters, B. Volckaert, and F. De Turck, "Inter-dataset generalization strength of supervised machine learning methods for intrusion detection," *Journal of Information Security and Applications*, vol. 54, p. 102564, 2020.

[5] P. Golchin, J. Weil, R. Kundel, and R. Steinmetz, "Dynamic network intrusion detection system in software-defined networking," in *2nd Workshop on Machine Learning & Networking (MaLeNe), co-located with the 5th International Conference on Networked Systems (NetSys 2023)*, 2023.

[6] Y. Akhiat, K. Touchanti, A. Zinedine, and M. Chahhou, "Ids-efs: Ensemble feature selection-based method for intrusion detection system," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 12 917–12 937, 2024.

[7] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." *ICISSp*, vol. 1, pp. 108–116, 2018.

[8] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

[9] E. B. Beigi, H. H. Jazi, N. Stakhanova, and A. A. Ghorbani, "Towards effective feature selection in machine learning-based botnet detection approaches," in *2014 IEEE Conference on Communications and Network Security*. IEEE, 2014, pp. 247–255.

[10] H. H. Jazi, H. Gonzalez, N. Stakhanova, and A. A. Ghorbani, "Detecting http-based application layer dos attacks on web servers in the presence of sampling," *Computer Networks*, vol. 121, pp. 25–36, 2017.

[11] S. Garcia, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *computers & security*, vol. 45, pp. 100–123, 2014.

[12] Z. Aouini and A. Pekar, "Nfstream: A flexible network data analysis framework," *Computer Networks*, vol. 204, p. 108719, 2022.

[13] P. Golchin, C. Zhou, P. Agnihotri, M. Hajizadeh, R. Kundel, and R. Steinmetz, "Cml-ids: Enhancing intrusion detection in sdn through collaborative machine learning," in *2023 19th International Conference on Network and Service Management (CNSM)*. IEEE, 2023, pp. 1–9.

[14] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[15] R. A. Shah, Y. Qian, D. Kumar, M. Ali, and M. B. Alvi, "Network intrusion detection through discriminative feature selection by using sparse logistic regression," *Future Internet*, vol. 9, no. 4, p. 81, 2017.

[16] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1184–1199, 2011.