A transformer-based hierarchical learning model for the detection of cybersecurity threats

J. M. Martínez-Ramírez[1], M. Lucena[1], O. Cordón[2] and C. J. Carmona [1,3,4]

[1] Department of Computer Science, University of Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[2] Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, Avenida del Conocimiento, 37, 18016 Granada, Spain
[3] Andalusian Research Institute in Data Science and Computational Intelligence, University of Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
4 Leicester School of Pharmacy, DeMontfort University, Gateway House, LE1 7RH Leicester, United Kingdom

Abstract. Cybersecurity is often stagnant, fighting a silent war against new attacks while developing much more slowly than other technologies. Due to the wide variety of attacks that we can find in technology, several branches of cybersecurity have also appeared. Deep learning has recently emerged as the machine learning technology best suited to predicting these attacks. This contribution presents a new hierarchical model based on deep learning able to handle accurately two different cybersecurity threat detection tasks. First, it will determine whether a given connection is an attack or not, thus dealing with a binary classification problem. Then, it will classify the malicious connections within a family of attacks. The proposed model offers accurate results when compared with those of state-of-the-art proposals, especially in the second task tackled. This study has been tested on three different datasets of real attack data, obtaining predictions with an accuracy of 99.92% on dataset CIC-IDS2017, 98.39% on CIC-CSE-IDS2018 and 93.74% on CIC-DDoS2019.

Keywords: Cybersecurity, Intrusion Detection, Deep Learning, Transformers, Hierarchical Models.

1 Introduction

Cybersecurity is fighting a silent war, as defense mechanisms often develop much more slowly than the attacks they are meant to thwart. As technologies advance, attackers design specific exploits to target them, rendering classic defense mechanisms insufficient. Within the broad field of cybersecurity, which includes user access authentication, network situation awareness, and abnormal traffic identification, this research focuses specifically on dangerous behavior monitoring, the primary task of Intrusion Detection Systems (IDS).

Algorithms supported by Deep Learning (DL) and Machine Learning (ML) are able to analyze large datasets and extract patterns within data that may be used to properly

discern between malicious and benign traffic and distinguish between different families of attack. These techniques could reduce the amount of false positives, which are the greatest issue in IDS as they have to be manually reviewed by humans.

2 The HiLMIDS model

The core contribution of this work is the Hierarchical Learning Model for Intrusion Detection Systems (HiLMIDS), a hierarchical system that combines classical machine learning with Transformers. The model operates in two distinct stages:

- Stage 1: Binary classification (Random Forest): The system first receives unlabeled network traffic and determines if the connection is benign or malicious. A Random Forest (RF) algorithm is used due to its high efficiency and robustness, which proves highly effective in the binary task. If a connection is deemed benign, the process stops, which saves computational energy and time.
- Stage 2: Multiclass Classification (Transformers): Connections identified as malicious in the first stage are forwarded to a second stage in which a model based on Transformers (specifically a modified BERT model) classifies the attack into a specific family, such as DDoS, Portscan, or Infiltration.

This hierarchical approach allows HiLMIDS to leverage the strengths of both models: the speed and accuracy of Random Forest offers a whitelist-like behavior in the first stage, while the superior pattern recognition capabilities of Transformers are used for more complex multiclass problems.

3 Technical implementation and model design

The architecture of the model can be seen in figure 1.

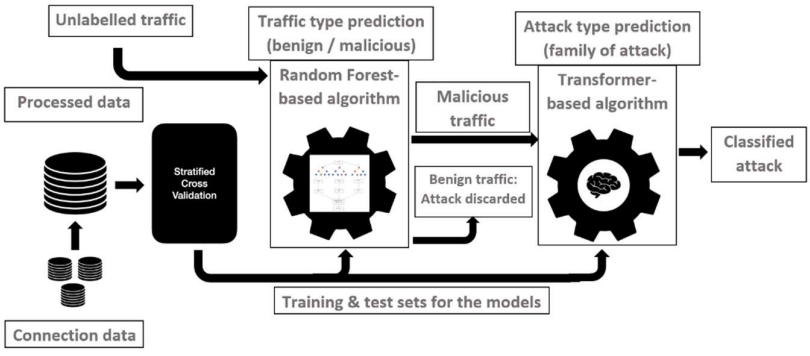


Fig. 1. Architecture of HiLMIDS.

3.1 Data pre-processing

The datasets used (CIC-IDS2017, CIC-CSE-IDS2018, and CIC-DDoS2019) presented significant challenges due to class imbalance (some classes had over 1 million instances while others had as few as 11). To address this, undersampling techniques were applied to majority classes, reducing them by up to 50% based on the amount of examples they had. Oversampling techniques were avoided to avoid the introduction of noise. Some attributes, such as timestamps and specific IP directions, were additionally removed to prevent the model from becoming biased toward specific time intervals or flow origins and destinations.

3.2 Stage 1: Random Forest

The RF model consists of an ensemble of 100 C4.5 trees using the Gini index as the construction criterion. It acts as a primary filter that aims to detect even 0-day attacks (new, unknown threats), as it marks any traffic that does not closely resemble known benign connections as malicious.

3.3 Stage 2: BERT Transformer

While Transformers were originally designed for Natural Language Processing (NLP), they can be used in several other tasks. In this case, the model is trained by receiving strings of connection data, each of which represents a network data flow. A BERT-base-cased model featuring 12 hidden layers and 768 cells per layer was used. Although Transformer training is computationally expensive and slow, its prediction speed once trained is high, making it viable for real-time IDS applications.

4 Experimental framework and methodology

The models were evaluated using stratified 10 fold cross-validation, ensuring that each training and test set maintained a distribution of classes similar to the original data. The study utilized high-performance hardware, including an NVIDIA RTX 2070 GPU and 32 GB of RAM, to ensure a fair comparison across all tested algorithms.

The primary metrics for success were Accuracy (the percentage of correct predictions) and Recall. Recall is particularly vital in cybersecurity because it is far more dangerous to misclassify a real attack as benign than to flag a benign connection as an attack. Execution times were also monitored, and though HiLMIDS takes roughly twice as long as simpler models to make a prediction, it remains within acceptable bounds for an effective IDS.

5 Results and experimental analysis

HiLMIDS demonstrated superior performance across all three real-world datasets:

- CIC-IDS2017: 99.9% accuracy and 99.8% recall.

- CIC-CSE-IDS2018: 98.4% accuracy and 98.8% recall.
- CIC-DDoS2019: 93.7% accuracy and 96.7% recall.

The models shows superiority over individual models, as outperforms using either RF or Transformers alone by up to 2%, a significant margin when processing millions of connections. Moreover, the Random Forest algorithm may offer some insight on which features are more important when discerning malicious traffic from benign connections, and the study shows that flag count, forwarded packets per second and flow duration are the most critical features when identifying malicious traffic.

The model was additionally compared with several other state-of-the-art models in an ablation study. Other models that were tested include recurrent neural networks (RNN), convolutional neural networks (CNN), and deep neural networks (DNN), most of which were outperformed by HiLMDS.

Lastly, pairwise Wilcoxon tests, Friedman tests and post-hoc Holm tests confirmed that the accuracy improvements offered by HiLMIDS were statistically significant compared to all aforementioned models.

6 Discussion and limitations

While HiLMIDS is highly effective, it faces two primary challenges: explainability and dataset imbalance. Both RF and Transformers are black box models, meaning it is difficult for a human to understand exactly why a specific connection was flagged. Furthermore, the extreme scarcity of data for certain attack types makes them harder to predict accurately.

HiLMIDS represents a robust step forward in intelligent network defense. Future work will focus on applying post-hoc explainability techniques to extract knowledge from the model's predictions, allowing security experts to understand the reasoning behind the alerts and further refine the system's capabilities. By combining the efficiency of classical machine learning with the pattern recognition of Transformers, HiLMIDS provides a high-accuracy tool for both broad threat detection and specific attack classification.

Acknowledgments. The research carried out in this study is part of the project "Advances in the development of trustworthy AI models to contribute to the adoption and use of responsible AI in healthcare (TAIH)" financed by the Ministry of Science, Innovation and Universities with code PID2023-149511OB-I00.

Disclosure of Interests. Authors declare no conflict of interest.

References

1. Martínez-Ramírez, J. M., Lucena, M., Cordón, O., & Carmona, C. J. A transformer-based hierarchical learning model for the detection of cybersecurity threats. Progress in Artificial Intelligence. https://doi.org/10.1007/s13748-025-00398-5. (2025)