Transformer-Based Intrusion Detection with Feature-Selected Inputs for Cross-Day Generalization and Unseen Attack Detection

1st Syed Aon Ali Naqvi, 2ndAmit K. Shukla, 3rd Petri Välisuo
School of Technology and Innovations, University of Vaasa, Vaasa, Finland
{syed.aon.naqvi, amit.shukla, petri.valisuo}@uwasa.fi

4th Muhammad Faheem
VTT Technical Research Centre of Finland, Espoo, Finland
muhammad.faheem@vtt.fi

***Abstract*—Most Intrusion Detection Systems (IDS) are evaluated using random data splits, leading to inflated performance that does not reflect the real-world deployment conditions. In practical environments, IDS must generalize temporally separated traffic and previously unseen attack patterns. To address this limitation, we propose a Transformer-based intrusion detection framework with a two-stage feature selection strategy. The proposed method combines Random Forest (RF) and permutation importance to identify informative features, reduce input noise and improve representation learning. The Transformer encoder further learns contextual feature representations through multi-head self-attention, which improves the modeling of complex dependencies. The proposed framework is evaluated on the CIC-IDS 2018 dataset under a cross-day experimental setting, where training and testing data contain different attack categories. Extensive experimental results demonstrate that the proposed framework significantly outperforms existing Machine Learning (ML) and Deep Learning (DL) baselines. The proposed framework improves the cross-day F1-score by up to 14% and achieves 97% F1 score, demonstrating its effectiveness in detecting unseen attacks in dynamic environments.**

***Index Terms*—Intrusion Detection, Transformer, Generalization, Cybersecurity, Features Selection, Unseen Attacks Detection**

I. Introduction

Networked systems are critical part of the modern society today, supporting various sectors including banking, healthcare, industrial control, smart homes and cloud computing. While the rapid growth of Internet has increased the convenience and connectivity in daily life, in parallel it also made us more vulnerable to cyber threats as we were never before. In 2018, more than 16 companies became the victim of cyber intrusions, in result lost approximaltely $13.0 billion [1]. In 2019 an attack was launched on Venezuela's power house, which resulted in the world longest blackout in the 20 out of 23 state for 6 days. As a consequence, essential services i.e., schools, hospitals and industries were severely disrupted [2]. According to the Nokia 2025 threat intelligence report [3], the Distributive Denial of Service (DDoS) attacks are occuring five times more frequent than ever, with the peaks traffic volume ranging from 5-10 Tbps.

A Network Intrusion Detection System (NIDS) is one of the essential security mechanisms to detect the malicious traffic and protect network from the harmful activities. NIDS distinguishes malicious traffic from normal traffic. Methodologically, NIDS can be divided into signature-based and anomaly-based Intrusion Detection System (IDS) [4]. Signature-based IDS relies on predefined rules and attack signatures and can only recognize the previously learned attacks, makes it ineffective against unknown threats. In contrast, anomaly-based IDS is capable of detecting abnormal behavior and zero-day attacks.

In past few years, Machine Learning (ML) techniques have shown promising performance in the field of NIDS. Tree-based classification methods such as Random Forest (RF) and XGboost (XGB) have frequently achieved high accuracy, sometimes close to 100%, when evaluated on the publicly available datasets. However, as attacks strategy continuously evolve, their increasing complexity demands more robust and adaptive solutions, which can deal with the emerging unseen problems. Deep Learning (DL), as a sub-domain of ML utilizes artificial neural networks for complex pattern recognition and unlike traditional ML methods, automatically learns features from raw data. Consequently, researchers enhanced NIDS solutions using DL models such as Multi Layer Perceptron (MLP), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and in recent years, Transformer-based architectures [5] as well.

Most existing DL methods being utilized in the NIDS are sequential and iterative in nature, which makes them computationally expensive and time-consuming during training. Furthermore, many of these methods focus on the spatial features of data by neglecting the contextual dynamics [6], which are crucial for detecting long-duaration and coordinated attacks such as DDoS. As a result, misclassification of the correlated attack pattern may occur, allowing threats to evolve over time [7]. Therefore, there is a strong requirement for models which are able to detect unseen and evolving attacks

This research is funded by The Finnish Doctoral Program Network in Artificial Intelligence, AI-DOC (decision number VN/ 313/2024-OKM-6).

efficiently. One the other hand, Transformer networks provide more powerful contextual modeling through a self-attention mechanism and enable parallel data processing, which makes it more efficient than conventional DL approaches. Moreover, Transformers are capable of capturing both local and global features interactions, which has led to their increasing adoption in the NIDS research. Most Existing IDS studies evaluate their models using random splits from the same dataset, which often leads to inflated performance that does not generalizable to real-world environments. D'hooge et al. [8] demonstrated that models achieving nearly perfect accuracy on intra-dataset evaluation can fail dramatically when tested on related but unseen datasets, highlighting the challenge of inter-dataset generalization.

To address these challenges, we propose a Transformer-based intrusion detection framework combined with an effective feature selection strategy. The proposed model captures complex relationships among network traffic features and learns robust representations for accurate attack detection. In real network environments, intrusion detection models are required to detect previously unseen attacks, makes generalization a critical requirement. This cross-attack scenario resembles a zero-day setting, where attack patterns differ significantly from historical data. In this work we focus on cross-day evaluation, in which the trainin set contains different attack categories than the testing set, enabling a rigorous assessment of model generalization under unseen and different evolving attack conditions.

The main contributions of this work are as follows:

- We propose a transformer-based intrusion detection framework for tabular network traffic data, capable of modeling complex feature dependencies using self-attention mechanisms.
- We design an effective feature selection strategy based on RF and permutation importance to enhance the detection performance and reduce computational complexity.
- We conduct extensive cross-day and cross-attack experiments to evaluate the generalization ability of the proposed model on unseen attack types, to reflect realistic deployment scenarios.
- We analyze the limitations of conventional random train-test evaluation and demonstrate the performance of cross-distribution validation of IDS.
- We compared the proposed approach with classical ML and DL baselines, showing consistent improvements in detection accuracy and generalization capability.

II. Related Work

Early IDS primarily relied on the classical ML approaches such as XGB, RF, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and related methods. These approaches demonstrated strong performances on public benchmark datasets. For instance, Chen et al. [9] utilized RF to develop an IDS for wireless networks, where their model was able to detect malicious signals and traffic effectively. Rohit et al. [10] proposed an ensemble approach based on three different algorithms: Naïve Bayes (NB), PART and Adaptive Boost combining their outputs and reducing variance error through bagging. Mohiuddin et al. [11] utilized modified the wrapper-based Whale optimization algorithm with the Sine Cosine algorithm for feature selection and employed weighted XGB for final classification.

With the rapid growth of network traffic and attack complexity, the cybersecurity community has increasingly shifting toward DL techniques due to their ability to process large-scale dataset efficiently. CNN is one of the power approach in the DL which has the ability to effectively capture spatial feature hierarchies. Several researchers network traffic into images to exploit CNN-Based feature extraction. Li et al. [12] transformed features intro grayscale images and utilized CNN-based model, achieving better results than classical ML approaches. RNN are capable of modeling sequential information, Kasangoo [13] proposed RNN based IDS that outperformed ML based ensemble methods. LSTM an extension of RNN addressing addresses vanishing gradient issues through its memory cells and gating mechanisms. Islam et al. [14] designed a stacked LSTM framework and compared with existing DL-based solutions. Although DL-based methods improve detection accuracy still many are suffer from high computation cost and limited capability to model long-range dependencies.

The emergence of Transformer network brought the significant advances in sequence modeling due to their self-attention mechanism, parallel processing capability and scalability. Wu et al. [1] presented a Transformer-based IDS and beaten the existing NIDS approaches, however its encoder-decoder architecture incurred higher training overhead than existing approaches. Subsequently, many researcher proposed modified encoder-based and hybrid Transformer approaches. For example Wang and Li [15] utilized Transformer for features extraction and CNN for classification. Han et al. [16] introduced a time-aware transformer which incorporate time encoding instead of positional encoding, enabling effective modeling of sequential attacks.

Most existing IDS solutions rely on a random train-test split from the same dataset, which often leads to overly optimistic performance estimation as proved with experiment by D'hooge et al. in [8]. To address the unknown intrusions, Jemili et al. [17] proposed an ensemble-based framework that integrate multiple datasources and employs future fusion and Principal Component Analysis (PCA) for dimensionality reduction. Their method demonstrate strong performance on cross-dataset evaluation, however, it mainly relies on classical ensemble learning and hand crafted feature fusion strategies, which may limit the adaptability to evolving attack patterns. Yao et al. [18] employed XGB for feature screening and utilized CNN-Transformer hybrid for the classification, evaluating generalization through 10-fold cross validation. However their study primarily reported only the accuracy which can be misleading in the highly imbalance dataset. Mao et al. [19] integrated Fully Convolutional Network (FCN) with Transformer encoder to extract multi-level features and applied few-shot learning

through an inductive module to recognize novel attack types. While these methods improve generalization to unseen attacks, their evaluation setting often involve partial overlap between training and testing distribution.

In contrast to existing studies, our work focuses on cross-day generalization under realistic deployment scenario. Moreover, our framework incorporate a feature reduction strategy to identify discriminative and relevant features to reduce the computational complexity. The proposed model is trained on network traffic collected from different days and evaluated on the unseen days containing novel attack types. Unlike prior work, our evaluation does not involve any fine-tuning on target-day data, which provides more realistic environment for the IDS capability to handle the novel and unseen attacks in real conditions.

III. Proposed Methodology

This section presents our proposed Transformer framework, referred to as Proposed-TF. The framework consists of four main components: data preprocessing, feature selection, representation learning and classification. The overall architecture of the proposed framework can be visualized in Figure 1. First, the raw CSV files are loaded and preprocessed to remove noise, normalize feature values and handle missing values. Then, a feature selection strategy is performed to retain only the most relevant attributes and reduce computation overhead. The selected features are subsequently embedded and fed into the Transformer encoder to learn discriminative representations. Finally, a classification head perform attack or benign prediction based on the learned feature representations.

A. Data Preprocessing

The raw network traffic data is stored in the CSV format file, where each day of traffic is stored in separate file. Each record initially contain 80 traffic-related features. Some CSV files contain repeated header rows appearing within the data, which are first identified and removed. Subsequently, missing (NaN) and infinite values are detected and replaced with zero to ensure numerical stability during training. Columns containing non-numeric values or constant zero values are eliminated, as they do not contribute to model learning. The timestamp information is also included in the features, it has also been removed to prevent potential information leakage during model validation. Finally, categorical labels are encoded into a numerical format for model training. After preprocessing, the cleaned dataset based on 68 attributes is used for feature selection.

B. Feature Selection

We applied a robust two-stage feature selection procedure by combining model-based and model-agnostic importance measures. This approach aims to identify stable and discriminative features while reducing redundancy and noise. Initially, a RF classifier was trained on the binary intrusion detection dataset and features were ranked according to the Gini-based importance derived from impurity reduction. It reflects the contribution of each features to the decision tree splitting. Let f_i denote the i-th feature. Its RF importance score is defined as:

$$I_{RF}(f_i) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n\in N_i^{(t)}} \Delta G(n), \tag{1}$$

where T denotes the number of trees, $\mathcal{N}_i^{(t)}$ represents the set of nodes in tree t where feature f_i is used for splitting, and $\Delta G(n)$ is the reduction in Gini impurity at node n.

In the second stage, permutation importance is computed using the trained RF model on the test set to enhance robustness and reduce model-specific biasness. It evaluates the importance of features by measuring the decrease in model performance when the feature's values are randomly permuted. For each feature f_i, its permutation importance is calculated as:

$$I_{PI}(f_i) = \mathcal{L}(D, M) - \mathcal{L}(D^{(i)}, M), \tag{2}$$

where $\mathcal{L}(\cdot)$ denotes the classification loss, $\mathcal{D}$ represents the original dataset, $\mathcal{D}^{(i)}$ denotes the dataset obtained after randomly permuting feature f_i, and M is the trained RF model.

To ensure stability, only features that appear within the top–40 rankings according to both methods are retained. Let $\mathcal{S}_{\mathrm{RF}}$ and $\mathcal{S}_{\mathrm{PI}}$ denote the top–40 feature sets obtained from RF and permutation importance, respectively. The stable feature subset is defined as:

$$S = S_{RF} \cap S_{PI}. \tag{3}$$

The top features among both methods are intersected to retain only features consistently identified as important by both methods. Finally, S are then re-ranked based on the RF importance score and the highest scoring subset was selected based on 20 features were selected. This dual-importance and intersection strategy reduces noise, mitigates model-specific biased and preserves only consistent information features.

C. Transformer-based Model

Transformer architecture is highly effective in capturing the long range dependencies, for modelling complex network traffic patterns. Conventionally, the Transformer consists of encoder-decoder structure, where the encoders process the input in parallel manner to extract the latent representations and decoder generate sequential output based on these representations. Since intrusion detection is formulated as a classification task without sequential output generation, only the encoder is employed in this framework.

The encoder is responsible for learning contextual relationships among input features and transform them into discriminative representations suitable for classification. Each encoder layer consists of two main components: multi-head self-attention and point-wise Feed-Forward Network (FFN). The encoder is based on multiple identical stacked layers to enhance the representation learning.

In this study, the input feature vector is treated as a sequence of tokens, where each feature corresponds to a token in the sequence. Unlike sequential traffic modeling, positional encoding in our framework represents feature index ordering, enabling consistent interaction learning across heterogeneous flow attributes. Self-attention enables each feature to interact

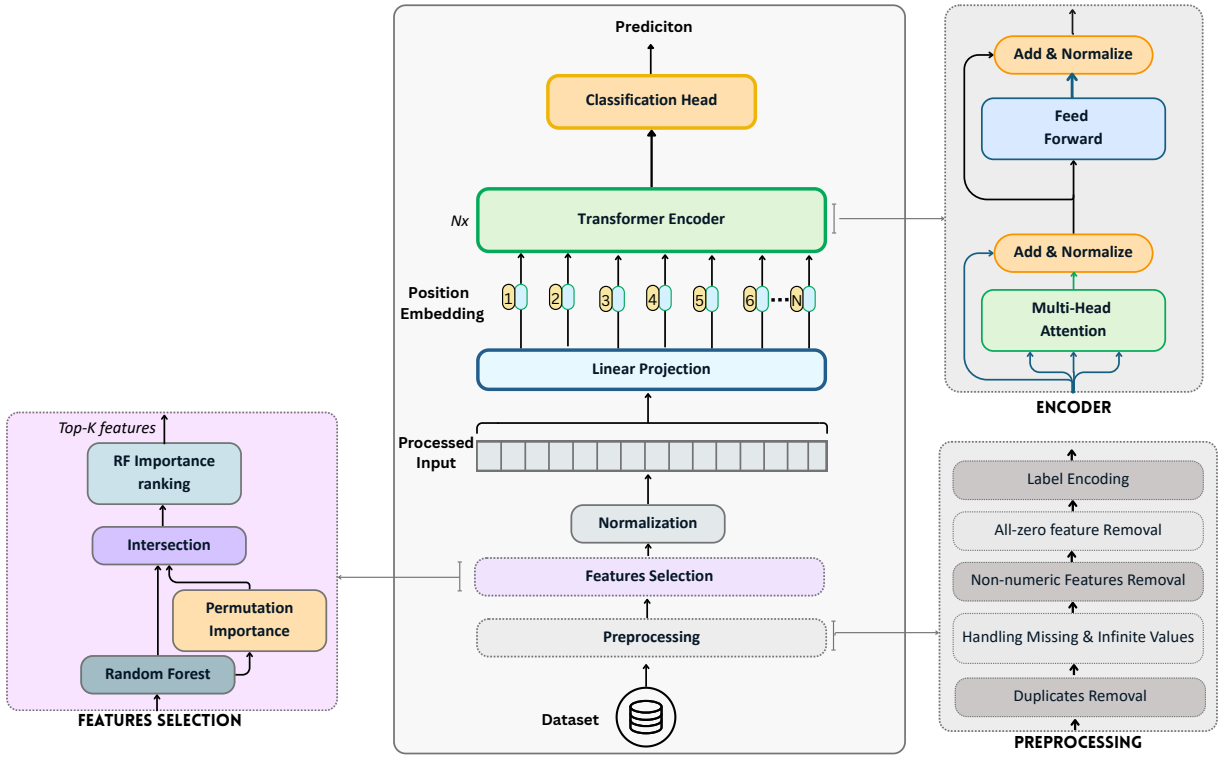

Fig. 1: Overview of the proposed framework, including data preprocessing, feature selection, and the Transformer encoder

with all the other features, allowing the model to to learn the relative importance and contextual dependencies.

The self-attention mechanism operates by projecting the input sequence in the form of three distinct matrices: Query (Q), Key (K) and Value (V). which are further used to compute the attention weights and generate context-aware feature representation: The scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{4}$$

Further multi-head attention is computed in parallel as:

$$\text{MHA}(\mathbf{E}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}_O, \tag{5}$$

where each head is defined as :

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i), \tag{6}$$

The embedded features are then processed through stacked encoder layers, each consisting of a multi-head self-attention followed by a position-wise FFN. Residual connections and layer normalization are employed to stabilize the training and facilitate gradient propagation.

Each encoder layer generates the output:

$$Z = \text{LayerNorm}\big(E + \text{MHA}(E)\big). \tag{7}$$

The final encoder output is aggregated using global average pooling and passed to a fully connected classification layer, followed by a SoftMax function to produce the predicted class probabilities:

$$\hat{y} = \text{softmax}(\mathbf{W}_c\mathbf{h} + \mathbf{b}_c), \tag{8}$$

where h denotes the pooled encoder representation, and $W_c$and b_c are trainable parameters.

IV. Experimental Evaluation

A. Dataset

The CIC-IDS 2018 dataset is considered for experimental evaluation. It was developed by Canadian Institute of Cybersecurity (CIC) in collaboration with Communication Security Establishment (CSE), and represents an updated end extended version of CIC-IDS2017 dataset. During its construction, network traffic was captured for the period of 10 days under realistic network conditions, where different attack scenarios were executed on different days. In this study, we considered a subset of a dataset consisting of three consecutive days. The first two days are used for training, while the third day is reserved for the evaluation in order to assess the generalization capability of the proposed model.

The training data contain Benign Traffic with the attack types FTP-BruteForce, SSH-Bruteforce, DoS attacks-SlowHTTPTest and DoS attacks-Hulk. On the other side in evaluation data include previously unseen attack variants, specifically DoS attacks-GoldenEye and DoS attacks-Slowloris, occurring under different attack patterns. Although these attack belong to the same general DoS Family as those in training data, their underlying characteristics and execution patterns differ significantly, and helping in forming a realistic cross-day and cross-attack evaluation scenario.

B. Experimental Setup

The proposed model was implemented using TensorFlow/Keras with the following hyperparameters: embedding dimension $d_{model} = 64$, number of attention heads = 4, feed-forward dimension $d_{ff} = 128$, number of encoder layers

= 2, dropout rate = 0.2, batch size = 1024, learning rate = 10^{-3}, and training epochs = 30. Early stopping was applied by monitoring the validation loss with a patience of 3 epochs. The model was trained using the Adam optimizer with binary cross-entropy loss and 10% of the training data was used for validation. Experiments were conducted on the Puhti supercomputing infrastructure provided by CSC – IT Center for Science, Finland.

C. Baseline Models

To evaluate the effectiveness of the proposed Transformer framework, we compare it with several widely used ML and DL models for intrusion detection. These models include RF, XGB, Logistic Regression (LR) and MLP which have demonstrated strong performance in the previous studies. In addition, a Transformer-based classifier without features selection is considered to assess the impact of the proposed feature selection strategy. All baseline models are trained under the same experimental setting for fair comparison.

D. Results

To establish baseline performance, we first evaluate all the models using a conventional 80/20 random train-test split on the selected two-day dataset. As expected, most model achieve high F1-score and Matthews Correlation Coefficient (MCC) values around 95%, which is consistent with prior studies, as shown in Table I. We are monitoring F1 and MCC, because accuracy can be misleading when the data is highly imbalanced. The results demonstrate that the dataset is relatively easy under random splits. However, such results do not reflect a realistic deployment scenario, where traffic patterns evolve overtime and attack characteristics continuously change.

TABLE I: Same distribution performance

Model	MCC	Accuracy	F1	Precision	Recall
RF	0.92	0.97	0.94	0.96	0.91
XGB	**0.94**	0.97	**0.95**	0.99	0.91
LR	0.84	0.94	0.87	0.92	0.83
MLP	**0.94**	0.97	**0.95**	0.99	0.90
Proposed-TF	**0.94**	0.97	**0.95**	0.99	0.91

To evaluate robustness and unseen-attack detection capability, we adopt a cross-day experimental setup, where models are trained on first two days data and tested on third day data which contain unseen DoS variants. As shown in Table II, performance drops dramatically across all the models. Although the Proposed-TF achieves relatively higher performance because of its generalization ability, but the most models collapse severely, exposing their lack of robustness under realistic temporal distribution shifts.

TABLE II: Cross-Day Evaluation Performance

Model	MCC	Accuracy	F1	Precision	Recall
RF	0.14	0.95	0.04	0.99	0.02
XGB	0.00	0.94	0.00	0.98	0.00
LR	0.81	0.94	0.80	0.98	0.67
MLP	0.79	0.98	0.77	0.99	0.63
Proposed-TF	**0.83**	0.98	**0.83**	0.99	0.71

E. Ablation Study

To reduce noise and retain only the most informative attributes, we investigate the impact of feature selection through an ablation study. Initially, all available features are used for training. Subsequently, we apply a RF feature selector and chosen top 20 ranked features. We experimentally evaluated multiple feature subset sizes (10, 15, 20, 30, and 40 features) and observed that retaining 20 features provided the best trade-off between detection performance and computational efficiency. This step leads to consistent performance improvement across all the classifiers, specially Proposed-TF model exhibit a substantial performance gain, as shown in the Table III. Notably, XGB which previously yielded nearly zero F1-score, improves to 0.77, demonstrating the effectiveness of eliminating irrelevant and redundant features.

TABLE III: Cross-Day Evaluation with RF-feature selection

Model	MCC	Accuracy	F1	Precision	Recall
RF	0.38	0.95	0.26	0.99	0.15
XGB	0.77	0.98	0.76	0.99	0.61
LR	0.87	0.98	0.87	0.78	0.99
MLP	0.84	0.98	0.84	0.99	0.72
Proposed-TF	**0.91**	0.99	**0.91**	0.98	0.86

Furthermore, we propose a two-stage feature selection strategy by intersecting the top-ranked features obtained by RF importance and permutation importance. This results in substantial improvement, particularly for Proposed-TF which achieve F1-score of 0.97, as reported in table IV. Figure 2 summarizes the evolution of cross-day performance under different feature selection configurations. The proposed method consistently outperformed all alternative settings, confirming its effectiveness for robust features selection in intrusion-detection scenarios. These results demonstrates that selecting stable and discriminative features is crucial for enhancing generalization of the model. The final feature subset is determined based on a comprehensive ablation analysis, as summarized in Table V.

TABLE IV: Cross-Day Evaluation with RF Importance ∩ Permutation Importance Feature Selection

Model	MCC	Accuracy	F1	Precision	Recall
RF	0.40	0.95	0.28	0.99	0.16
XGB	0.76	0.98	0.75	0.99	0.60
LR	0.83	0.95	0.84	0.78	0.90
MLP	0.81	0.98	0.81	0.99	0.68
Proposed-TF	**0.97**	0.99	**0.97**	0.99	0.94

TABLE V: Impact of Different Feature Selection Methods on F1-Score

Method	**Effect on F1**	**Remarks**
Mutual Information	↓	Performance degradation
Logistic Regression	↔	Failed to reduce features
PCA	↓	Loss of discriminative information
Information Gain	↓	Unstable feature ranking
ANOVA	↑	Slight improvement
RF	↑	Significant improvement
Extra Trees	↑	Lower improvement than RF
RF + Permutation	↑↑	Proposed method (best)

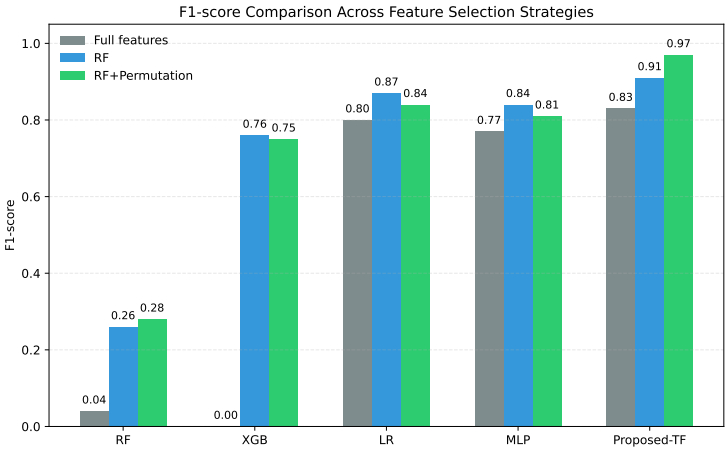

Fig. 2: F1-score comparison across models using Full-features, RF selection, and RF+Permutation feature-selection strategies.

F. Discussion

The results of the experiment clearly demonstrate that conventional evaluation protocols may significantly overestimate IDS performance. Although most classifiers perform well on random splits, their effectiveness fails under cross-attack conditions. By incorporating a robust feature selection strategy and Transformer-based representation learning, our framework improves generalization to unseen attack scenarios. This finding is particularly important for real-world deployment, where labeled data from future attack types is unavailable. Although the proposed approach achieves strong performance, further investigation is required to evaluate its scalability across additional datasets.

V. Conclusion

In this study, we investigated the limitations of conventional IDS under realistic cross-day and unseen-attack scenarios. Experimental results demonstrate that the model achieving high results on the random-train test split often fail when it comes to unseen and new attack variants. To address this challenge, we proposed a Transformer-based intrusion detection framework integrated with a robust two-stage feature selection strategy. By combining RF importance and permutation-based analysis, the framework effectively identifies informative features and enhances representation learning on the unseen traffic. It enables Proposed-TF to effectively capture contextual dependencies by improving generalization across diffrent attack distributions. Although, the proposed approach exhibit strong performance in the designed scenario, several directions remain for future exploration. These include extending the framework to cross-dataset evaluation, integration of self-supervised and contrastive learning techniques for label-efficient training and validating the model on large-scale real-world datasets.

Declaration of AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT only to improve the sentence structure and grammar, aiming to enhance the readability of the article. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "Rtids: A robust transformer-based approach for intrusion detection system," *IEEE Access*, vol. 10, pp. 64 375–64 387, 2022.

[2] D. Sun, L. Zhang, K. Jin, J. Ling, and X. Zheng, "An intrusion detection method based on hybrid machine learning and neural network in the industrial control field," *Applied Sciences*, vol. 13, no. 18, p. 10455, 2023.

[3] https://www.nokia.com/cybersecurity/threat-intelligence-report/, [Accessed 03-12-2025].

[4] M. Abdel-Basset, N. Moustafa, H. Hawash, I. Razzak, K. M. Sallam, and O. M. Elkomy, "Federated intrusion detection in blockchain-based smart transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2523–2537, 2021.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Vehicular Communications*, vol. 21, p. 100198, 2020.

[7] L. Yang, A. Moubayed, and A. Shami, "Mth-ids: A multitiered hybrid intrusion detection system for internet of vehicles," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 616–632, 2021.

[8] L. D'hooge, M. Verkerken, T. Wauters, F. De Turck, and B. Volckaert, "Investigating generalized performance of data-constrained supervised machine learning models on novel, related samples in intrusion detection," *Sensors*, vol. 23, no. 4, p. 1846, 2023.

[9] Y. Chen and F. Yuan, "Dynamic detection of malicious intrusion in wireless network based on improved random forest algorithm," in *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*. IEEE, 2022, pp. 27–32.

[10] R. K. S. Gautam and E. A. Doegar, "An ensemble approach for intrusion detection system using machine learning algorithms," in *2018 8th International conference on cloud computing, data science & engineering (confluence)*. IEEE, 2018, pp. 14–15.

[11] G. Mohiuddin, Z. Lin, J. Zheng, J. Wu, W. Li, Y. Fang, S. Wang, J. Chen, and X. Zeng, "Intrusion detection using hybridized meta-heuristic techniques with weighted xgboost classifier," *Expert Systems with Applications*, vol. 232, p. 120596, 2023.

[12] Y. Li, Y. Xu, Z. Liu, H. Hou, Y. Zheng, Y. Xin, Y. Zhao, and L. Cui, "Robust detection for network intrusion of industrial iot based on multi-cnn fusion," *Measurement*, vol. 154, p. 107450, 2020.

[13] S. M. Kasongo, "A deep learning technique for intrusion detection system using a recurrent neural networks based framework," *Computer Communications*, vol. 199, pp. 113–125, 2023.

[14] N. Islam, F. Farhin, I. Sultana, M. S. Kaiser, M. S. Rahman, M. Mahmud, A. Hosen, and G. H. Cho, "Towards machine learning based intrusion detection in iot networks," *Comput. Mater. Contin*, vol. 69, no. 2, pp. 1801–1821, 2021.

[15] H. Wang and W. Li, "Ddostc: A transformer-based network attack detection hybrid mechanism in sdn," *Sensors*, vol. 21, no. 15, p. 5047, 2021.

[16] X. Han, S. Cui, S. Liu, C. Zhang, B. Jiang, and Z. Lu, "Network intrusion detection based on n-gram frequency and time-aware transformer," *Computers & Security*, vol. 128, p. 103171, 2023.

[17] F. Jemili, K. Jouini, and O. Korbaa, "Detecting unknown intrusions from large heterogeneous data through ensemble learning," *Intelligent Systems with Applications*, vol. 25, p. 200465, 2025.

[18] R. Yao, N. Wang, P. Chen, D. Ma, and X. Sheng, "A cnn-transformer hybrid approach for an intrusion detection system in advanced metering infrastructure," *Multimedia Tools and Applications*, vol. 82, no. 13, pp. 19 463–19 486, 2023.

[19] J. Mao, X. Yang, B. Hu, Y. Lu, and G. Yin, "Intrusion detection system based on multi-level feature extraction and inductive network," *Electronics*, vol. 14, no. 1, p. 189, 2025.