

A comparative study of deep learning-based network intrusion detection system with explainable artificial intelligence

Tan Juan Kai, Lee-Yeng Ong, Meng-Chew Leow

Faculty of Information Science and Technology (FIST), Multimedia University, Melaka, Malaysia

Article Info

Article history:

Received Oct 24, 2024

Revised Mar 16, 2025

Accepted May 23, 2025

Keywords:

AWID3 dataset

Deep learning

Explainable artificial intelligence

Local interpretable model-agnostic explanation

Network intrusion detection system

Shapley additive explanation

TabNet

ABSTRACT

In the rapidly evolving landscape of cybersecurity, robust network intrusion detection systems (NIDS) are crucial to countering increasingly sophisticated cyber threats, including zero-day attacks. Deep learning approaches in NIDS offer promising improvements in intrusion detection rates and reduction of false positives. However, the inherent opacity of deep learning models presents significant challenges, hindering the understanding and trust in their decision-making processes. This study explores the efficacy of explainable artificial intelligence (XAI) techniques, specifically Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME), in enhancing the transparency and trustworthiness of NIDS systems. With the implementation of TabNet architecture on the AWID3 dataset, it is able to achieve a remarkable accuracy of 99.99%. Despite this high performance, concerns regarding the interpretability of the TabNet model's decisions persist. By employing SHAP and LIME, this study aims to elucidate the intricacies of model interpretability, focusing on both global and local aspects of the TabNet model's decision-making processes. Ultimately, this study underscores the pivotal role of XAI in improving understanding and fostering trust in deep learning -based NIDS systems. The robustness of the model is also being tested by adding the signal-to-noise ratio (SNR) to the datasets.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Lee-Yeng Ong

Faculty of Information Science and Technology, Multimedia University

Jalan Ayer Keroh Lama 75450 Melaka, Malaysia

Email: lyong@mmu.edu.my

1. INTRODUCTION

As the use of the internet continues to grow, maintaining robust security measures becomes increasingly important. Moreover, the prevalence of zero-day attacks adds a layer of urgency on developing and implementing such measures [1]. As such, anomaly-based network intrusion detection system (NIDS) is introduced to effectively detect zero-day attacks throughout comparison of network traffic profiles, utilizing power of machine learning or deep learning approaches [2]. Deep learning approaches in NIDS model development had proven to be more effective as it often has a better performance in terms of producing a high detection rate while keeping the low positive rate [3]. Despite the effectiveness of deep learning-based approaches in NIDS, their decision-making processes often lack transparency and clarity [4]. The explainability of prediction and classification models is typically inversely proportional to their learning performance, especially for deep learning approaches which are often referred to as "black boxes" due to their complex structures and opaque decision-making processes [5], [6].

This lack of interpretability poses significant challenges for network administrators who rely on these systems to identify and respond to abnormal network behaviors. Moreover, it is crucial to understand

the model's behavior, as this understanding allows them to trust the system's alerts and take informed actions based on the model's output. Without this knowledge, administrators may struggle to distinguish between true threats and false positives, potentially leading to either unnecessary disruptions or missed attacks. As such, the emergence of explainable AI (XAI) has become vital in the realm of network intrusion detection, enhancing the transparency and interpretability of AI models. This study employs Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME), which are well-regarded XAI techniques, to elucidate the decision-making processes of these models.

The contributions of this paper are threefold: first, it develops a NIDS model using a deep learning approach that achieves a high intrusion detection rate with a low false positive rate. Secondly, it evaluates the robustness of the NIDS model in noisy environments. Lastly, it interprets the model's decision-making steps through the application of SHAP and LIME. The remaining part of the paper is structured as follows. Section 2 discusses related work on deep learning-based NIDS and XAI approaches. Section 3 concentrates on the methodology of the experiment pipeline. Section 4 shows the experimental results and discussion, and finally section 5 concludes the contributions and future work of the study.

2. RELATED WORK

2.1. Deep learning approaches in NIDS development

Deep learning approaches in NIDS consist of deep neural network (DNN), convolution neural network (CNN) and long-short term memory (LSTM). In terms of DNN, Tang *et al.* [7] had developed a software-defined network-based NIDS using DNN and manage to hit an accuracy of 75.75% on NSL-KDD datasets. Similarly, Wang *et al.* [8] found out that DNN emerges in terms of intrusion detection for the CES-CIC-IDS 2018 datasets after comparing the results with other five deep learning models and manage to hit the accuracy of 98.79% accuracy using five hidden layers with 256 nodes.

In terms of CNN, Ahmad *et al.* [9] had proposed a CNN model using AWID3 datasets after encoding and converting the tabular data into images using Gramian angular field approach. The proposed model of the architecture 2D-CNN-1 layer achieved the best performance and managed to hit an accuracy of 99.77%, with a precision of 99.59%, recall of 99.73% and F1-score of 99.66%. Moreover, an LSTM-based model for intrusion detection in in-vehicle CAN bus communications was employed by Hossain *et al.* [10], achieving an impressive accuracy of 99.995% using self-collected datasets.

Hybrid-based approaches of CNN and LSTM have also commonly used in the development of NIDS. For instance, Deore and Bhosale [11] developed CNN-LSTM model by using the CNN architecture for feature extraction and using LSTM as its classifier, through integration with chimp chicken swarm optimization approach. The CNN-LSTM model manages to hit an accuracy of 93.97% for non-attack profile and 98.88% for the intrusions attempt in NSL-KDD dataset, while hitting an accuracy of 98.88% for non-attack profile and 90.58% accuracy of attack profile in the BoT-IoT dataset. The same approach was customized in the work of [12], which managed to hit an accuracy of 99.84% for binary classification and 99.80 accuracy for multiclass classification in X-IIoTID dataset. In addition, the customized architecture of CNN-LSTM also achieved an accuracy of 93.21% for binary classification and 92.9% for multiclass classification in UNSW-NB15 dataset.

2.2. Explainable AI approaches in NIDS

Explainable AI (XAI) approach can basically be divided into two main categories, which are global interpretability and local interpretability [13]. Global interpretability refers to understanding the overall behavior and decision-making process of the entire model, providing insights into how the model makes predictions across all inputs. Local interpretability, on the other hand, focuses on explaining individual predictions, offering a detailed understanding of why the model made a specific decision for a particular input instance.

For global interpretability, SHAP is normally used to access the overall behavior of NIDS model which are reported in various research works [14]–[18] using different approaches. For instance, [14], [16], [17] used the summary plot of SHAP to view the overall feature importance of data and show the features contribution to the corresponding labels in both binary classification and multiclass classification tasks. Meanwhile, study [18] utilized bee swarm plot to interpret the decision-making steps for binary class through different classifiers. Other methods that could be used to access the global interpretability of deep learning models such as, permutation importance (PI), contextual importance and utility (CIU) [14] and rule fit [15].

Moving onto the context of local interpretability, LIME is generally used as a tool for analyzing the interpretation of individual prediction. Common utilization of LIME is similar to the approach described in [17], where local probability predictions are displayed alongside with the features that contributed to those predictions. Meanwhile, study [18] uses LIME to plot the frequent features to analyze the most important features in the particular prediction. On the other hand, study [15] highlighted the features that often lead to

correct or wrong predictions by analyzing the local interpretability of positive and negative scenarios. SHAP can also be utilized for local interpretation purposes. For instance, studies [17], [19] used SHAP waterfall plot to illustrate the effect of features on a particular classification made based on the selected features relative to the index scores.

3. METHOD

In the proposed pipeline for developing deep learning-based NIDS with XAI, AWID3 dataset is utilized, whereby it consists of 13 types of intrusions in WPA2 networks with a total sample of 30,387,099 normal traffic and 6,526,404 malicious traffic [20]. Initially, data preprocessing is performed to clean and prepare the data for the latter stage. This is followed by the application of a feature selection algorithm to identify the most relevant features for the model. Subsequently, model development is conducted using the selected features to create a predictive model. The performance of this model is then evaluated to assess its effectiveness. Additionally, results are interpreted using XAI technique to show the transparency of the model's decision-making process.

3.1. Data preprocessing

Among the 13 types of intrusions available in the AWID3 datasets, 7 specific intrusions relevant to the network access layer of the TCP/IP model have been selected. These intrusions include deauthentication attacks, disassociation attacks, (re)association attacks, Rogue access point (AP) attacks, Evil Twin attacks, KRACK attacks, and Kr00k attacks. The attack labels are categorized into three groups: denial-of-service (DoS) attacks, man-in-the-middle (MiTM) attacks, and traffic decryption attacks. The outcome of label mapping is illustrated in Table 1. Features with more than 80% missing values are excluded, and data imputation techniques are used to address the remaining missing values. Categorical data are pre-processed using ordinal encoding, while numerical data are processed using min-max scaling.

Table 1. Label mapping of AWID3 dataset

Original intrusion	Normal traffic	Malicious traffic	Label mapping
Deauthentication	1,587,527	38,942	Denial-of-service (DoS)
Disassociation	1,938,585	75,131	
(Re)association	1,838,430	5,502	Man-in-the-Middle (MiTM)
Rogue AP	1,971,875	1310	
Evil Twin	3,673,854	104,827	
KRACK	1,388,498	49,990	Traffic decryption
Kr00k	2,708,637	186,173	

3.2. Feature selection

In order to obtain the optimal feature sets, feature selection algorithm named phi-K is being utilized as it is able to compute the correlation between categorical data and numerical data [21]. The phi-K matrix scores and their corresponding significance values are computed. The top 15 features with the highest phi-K scores and significant values are selected to reduce the dimensionality of data. These selected features with the associated values and description are presented in Table 2.

3.3. Model development

TabNet is employed in the development of the NIDS model due to its robust capabilities in handling tabular data [22]. TabNet is a deep learning architecture designed specifically for tabular data, utilizing gradient descent-based optimization to enable flexible end-to-end learning, which consists of feature and attentive transformers and fully connected layers. Before fitting the data into the models, it is split into three sets: 75% for training, 15% for validation, and 15% for testing. The parameters and model architecture of TabNet are listed in Table 3. Note that the parameter *weight* in TabNet is set to 1 in order to automatically distribute the weights among the classes to solve the class imbalanced issue.

3.4. Performance and robustness evaluation

The performance of the model is evaluated using a confusion matrix, accuracy, recall, precision, and F1-score. Subsequently, the model's performance is compared with four state-of-the-art (SOTA) models to benchmark its effectiveness. To assess the robustness of the model, the same performance evaluation metrics are applied to the AWID3 dataset with the addition of signal-to-noise ratio (SNR) from the range of 15 to 30. The inclusion of SNR is intended to simulate the level of desired signal relative to background noise, providing a realistic scenario to test the model's ability to handle noisy data in a real time environment. The SNR values are computed based on (1) as referenced in source [23].

$$SNR_{dB} = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right) \quad (1)$$

Table 2. Label mapping of AWID3 dataset

Features	Phik matrix	Phik significance	Description
radiotap.length	0.130962	716.129378	The length of the radiotap header, which provides metadata about the wireless packet.
frame.len	0.140262	597.823233	The total length of the frame, including headers and payload.
radiotap.channel.flags.cck	0.161951	211.391586	Indicates whether the complementary code keying (CCK) modulation is used in the channel.
wlan_radio.phy	0.180651	785.503059	Specifies the physical layer type used for the wireless transmission, such as 802.11a, b, g, n, or ac.
radiotap.present.tsft	0.241758	629.869990	A flag indicating the presence of the time synchronization function timer (TSFT) field in the radiotap header.
wlan.fc.ds	0.241886	693.658164	The distribution system (DS) status field in the 802.11 frame control field, indicating the direction of the frame relative to the distribution system.
wlan.fc.protected	0.253686	670.374653	Indicates whether the frame is protected by encryption.
radiotap.timestamp.ts	0.269606	950.682039	The timestamp of when the frame was captured, provided by the radiotap header.
frame.time_relative	0.298154	1,041.196104	The relative time from the beginning of the capture to when the frame was captured, typically measured in seconds.
wlan_radio.channel	0.323927	680.270172	The radio channel on which the frame was transmitted.
wlan.fc.type	0.437673	1,375.656433	The type of frame, such as management, control, or data frame.
wlan_radio.data_rate	0.472091	1,374.488746	The data rate at which the frame was transmitted.
wlan_radio.signal_dbm	0.580638	1,070.573341	The signal strength of the frame in decibels-milliwatts (dBm).
radiotap.dbm_antsignal	0.737401	1,463.181892	The signal strength received by the antenna in decibels-milliwatts (dBm).
wlan.fc.subtype	0.756285	1,529.576742	The subtype of the frame, providing more specific information about the frame's purpose, such as association request, data and acknowledgment.

Table 3. Model architecture and parameters of TabNet

Parameter	Value
n_steps	3
optimizer_fn	Adam
optimizer_params	dict(lr=0.005)
n_d	14
n_a	14
scheduler_params	{"step_size": 3, "gamma": 0.7}
scheduler_fn	torch.optim.lr_scheduler.StepLR
weight	1

3.5. Interpreting model prediction with explainable AI

To achieve a comprehensive understanding of the model's predictions, both global and local interpretation methods are employed. For global interpretation, the SHAP Kernel explainer is utilized. SHAP values provide a consistent measure of feature importance by quantifying the contribution of each feature to the model's predictions, thus offering transparency into the model's overall behavior. For local interpretation, LIME is used to construct interpretable local models around each prediction. This approach enables the explanation of individual prediction by approximating the model locally with a simpler and more interpretable model.

4. EXPERIMENT RESULTS AND DISCUSSION

To comprehensively evaluate the TabNet model on the AWID3 datasets, three experiments were conducted. Firstly, the model's performance metrics, including accuracy, precision, recall, F1-score, undetected intrusions, and false alarm rates, were assessed and compared to SOTA models to benchmark its effectiveness in detecting network intrusions. Secondly, the model's robustness was tested under varying SNR conditions to ensure high detection accuracy and low false positives in noisy environments. Lastly, interpretability was examined using SHAP and LIME techniques. SHAP provided insights into the global feature importance for different attack types, while LIME offered local interpretations of individual predictions, highlighting feature contributions to correct and incorrect classifications.

4.1. Performance evaluation of TabNet model

The confusion metrics of the TabNet model are being shown in Figure 1. Notably, the model exhibits commendable effectiveness in detecting network intrusion, with only 4 intrusion attempts going undetected across the entire testing datasets. Furthermore, the TabNet model demonstrates proficiency in addressing the prevalent issue of high false alarm rates, as shown by the generation of only 213 false alarms out of 2,336,237 testing samples.

The performance metrics listed in Table 4 are used to evaluate the comparison between TabNet with other SOTA models. Compared with other SOTA results, it is evident that TabNet outperforms other methods while utilizing more comprehensive datasets. In terms of intrusion detection, TabNet achieved a remarkable indicator, demonstrated by its 99.99% precision. Notably, the low false alarm rate, indicated by its recall, matches the high precision, showcasing the model's reliability and accuracy. This performance suggests that TabNet is exceptionally effective in identifying and mitigating various types of attacks, including deauthentication, disassociation, reassociation, Rogue AP, Krack, Kr00k, and Evil Twin. The balanced high scores across all metrics highlight TabNet's superiority in maintaining security and accurately detecting intrusions, making it a robust choice for intrusion detection systems.

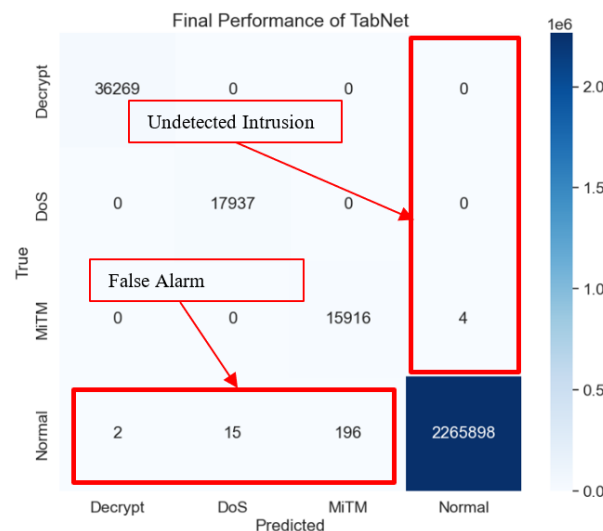


Figure 1. Confusion matrix of TabNet on AWID3 datasets

Table 4. Performance evaluation with other SOTA on AWID3 datasets

Model used	Utilization of datasets	Performance metrics (%)	
Bagging classifier [24]	Botnet, Malware, SSH, SQL injection, SSDP amplification, and Website spoofing	Accuracy: 96.70	Precision: 96.84
K-nearest neighbors K-NN [25]	Around 1 million subsets taken from the dataset (15% of datasets)	Recall : 95.03	F1-score : 88.07
		Accuracy: 99.00	Precision: N/A
2D-CNN-1 layer [9]	20% of Deauthentication, Disassociation, Reassociation, Rogue AP, Krack, Kr00k, and Evil Twin data	Recall : N/A	F1-score : N/A
		Accuracy: 99.77	Precision: 99.59
Extra tree [26]	Deauthentication, Disassociation, Reassociation, Rogue AP, Krack, Kr00k, and Evil Twin	Recall : 99.73	F1-score : 99.66
		Accuracy: 99.96	Precision: 99.75
TabNet	Deauthentication, Disassociation, Reassociation, Rogue AP, Krack, Kr00k, and Evil Twin	Recall : 99.28	F1-score : 99.52
		Accuracy: 99.99	Precision: 99.99
		Recall : 99.99	F1-score : 99.99

4.2. Robustness evaluation

SNR serves as a metric to assess the efficacy of the NIDS in managing noise within real-time wireless networks. By quantifying the ratio of signal power to background noise power, SNR facilitates an understanding of the system's capability to detect intrusions amidst varying levels of interference. The lower the SNR values being utilized, the more the noise overwhelms the signal. SNR values ranging from 15 to 30 are employed throughout the second experiment to encompass weak to strong signal conditions. Figure 2 presents a visual depiction of TabNet performance across different SNR levels, shedding light on its behavior under varying noise intensities.

Across the range of SNR, the recall metric, which is represented by the gray bars, consistently maintains a stable pattern. This consistency proved the capability of TabNet to accurately identify genuine intrusions remains unaffected by the fluctuations of signal quality. This resilience in recall underscores the model's effectiveness in detecting intrusions, using true positives as indicators, irrespective of noise levels.

Conversely, precision, indicated by the orange bars, exhibits noticeable variability across different SNR values. Particularly evident at lower SNR levels, such as SNR 10, precision tends to be lower relative to higher SNR values. This indicates that there is a greater likelihood of affected network packet content being misclassified as anomalies when the environment consists of a higher level of noise. As a result, the model tends to produce more false positives in the situation of poorer signal quality, leading to a decrease in precision.

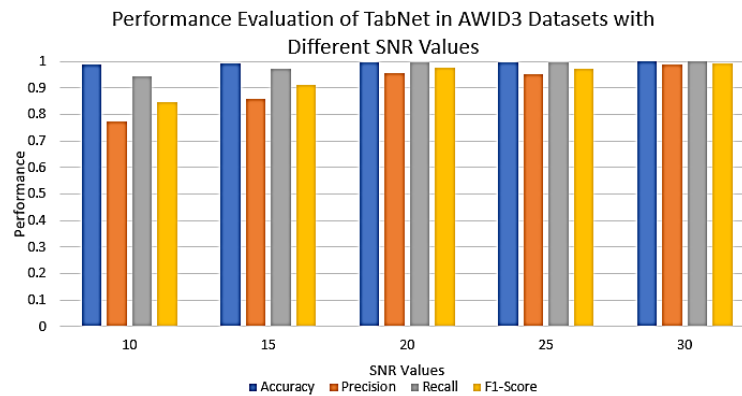


Figure 2. Performance visualization in AWID3 with different SNR values

4.3. Global interpretation of TabNet model using SHAP

SHAP summary plots in Figure 3 to 5 provide a comprehensive visualization of the impact of each feature on the model's output. The features are listed along the y-axis, ordered by their overall influence on the prediction. Each point represents a SHAP value for a feature in a particular instance, with the color indicating the feature value, where blue signifies low and red signifies high. This ordering helps to quickly identify which features are the most influential in determining the model's predictions.

SHAP feature value distribution in Figure 3 sheds light on MiTM attacks. Notably, features like *wlan.fc.type*, *wlan.fc.subtype*, and *frame.len* exhibit higher values, which are consistently shown in red plots. *Wlan.fc.type* signifies the general category of transmitted frames, while *wlan.fc.subtype* provides more specific details within that category. Rogue APs and Evil Twins, aiming to impersonate legitimate AP, often use beacon frames to lure users. These beacon frames generate subtype 8 packets, categorized as data frames, with *wlan.fc.type* numbers corresponding to 2. Additionally, for differentiation, the author filters Rogue AP attacks based on *frame.len* being less than 264 and Evil Twin attacks with *frame.len* less than 242. Moreover, an extra filter is applied to Evil Twin attacks, involving deauthentication frames to disconnect devices from the original AP, facilitating their connection to the malicious one.

Moving onto the SHAP subplot in DoS attacks illustrated in Figure 4, shows that the feature *wlan.fc.subtype* has several red points. This positioning suggests that higher values of this feature are closely linked to an increased likelihood of a DoS attack occurrence. Consequently, it implies a strong association between specific frame types and heightened risks of DoS attacks. For instance, the NIDS model scrutinizes network packets to detect potential flooding of certain frame types. Going deep into the features, deauthentication attacks correspond to subtype 10, disassociation attacks to subtype 12, and reassociation attacks to subtypes 0, 2, and 8 as per filter applied by the authors [20]. As a result, this implies that the TabNet model classifies DoS attacks in a manner that closely resembles how network administrators evaluate such attacks in real-world environments.

Moreover, Figure 5 shows the SHAP summary subplots on traffic decryption attacks. Based on the SHAP distribution, it can be observed that the feature *wlan.fc.subtype* has the highest impact value, followed by *wlan.radio.channel*. This scenario may happen due to the methodology of the author in collecting the AWID3 datasets on the KRACK and Kr00k attacks. Specifically, the significant impact of the *wlan.fc.subtype* feature aligns with the dataset authors' method of filtering and labeling network packets. They labeled packets where the feature *wlan.fc.type_subtype* is equivalent to 10 as Kr00k attacks. Next, the

feature of *wlan_radio.channel* which indicates the network channel of the network packet is located. KRACK attacks primarily conducted on channels 2 and 13 according to the authors [20], which is contradicted to the normal packet profile and other intrusions that are initially collected on channel 36. As a result, the model tends to classify the network packet allocated to channel outside of channel 36, specifically channel 2 and channel 13 as KRACK attack attempts. Moreover, *wlan.fc.protected* is the only feature that has a high feature value (mixed with red color) as compared to Figures 3 and 4 which have only low feature value (entirely blue color). It is due to the nature of traffic decryption attacks which causes the encryption key of the network packet content to be reset to an all-zero value which mean no encryption protection is available. This assertion is demonstrated by the authors using a Wireshark filter, specifically by setting *wlan.fc.protected* to zero, to identify and label the Kr00k attacks.

In conclusion, the global interpretation of SHAP values provides valuable insights into the alignment between the NIDS model's comprehension of overall results and the author's data filtering methodology alongside the intrinsic characteristics of the attacks. The discernible correspondence between the SHAP values and the applied data filtering approach highlights not only the efficacy of the feature selection process but also contributes to a deeper understanding of the decision-making framework employed by the model and hence increases the trustworthiness of the intrusion detections made among the network administrators. This concordance between the model's interpretation and the observed attack patterns serves.

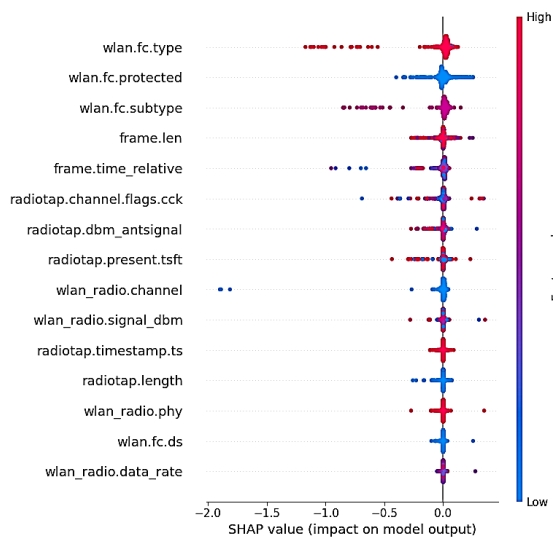


Figure 3. SHAP subplots for MiTM attacks

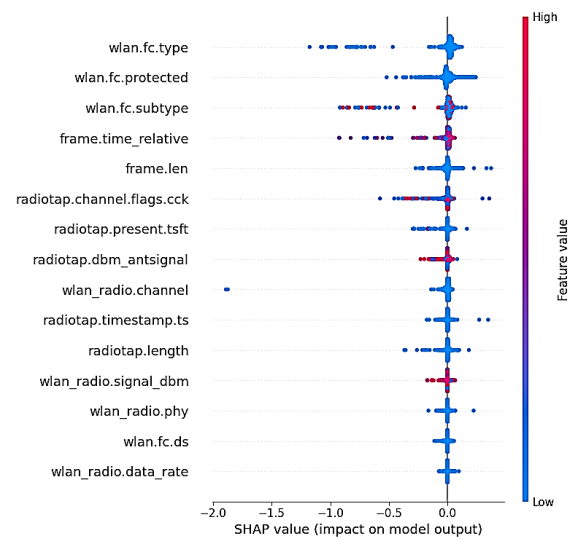


Figure 4. SHAP subplots for DoS attacks

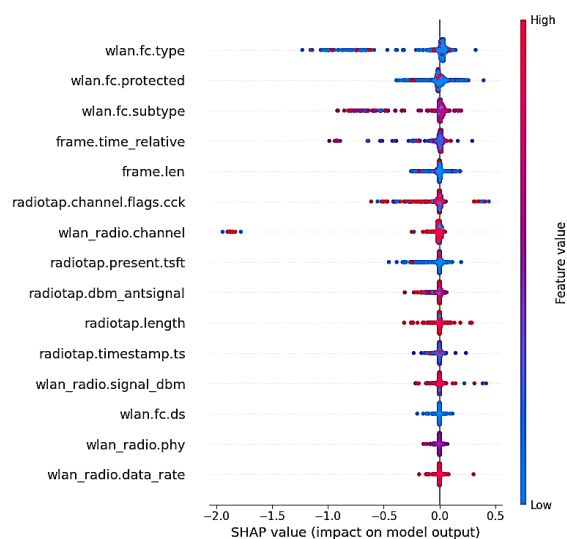


Figure 5. SHAP subplots for traffic decryption attacks

4.4. Local interpretation of TabNet model using LIME

The LIME interpretation figures provided in this study consist of three parts. The left table shows the prediction probabilities of the model interpretation which also indicates the confidence score of the model making the predictions. The right two-sided bar chart shows the detailed breakdown of the contribution of various features to prediction results, whereby the red bar at the left side shows the negative indicator to the predictions while green bar at the right side shows the positive indicator to the predictions. These magnitude levels shown in the right two-sided bar are the indicator of contributions on how the TabNet model make the classification results, whereby a positive magnitude level contributes to the classification made, and a negative magnitude level is opposing the classification results made.

Figure 6 illustrates a local interpretation using LIME for a model prediction marked as a DoS attack. The interpretation highlights accurately classified intrusion attempts of DoS attacks in the left table. The prediction probabilities indicate 100% confidence that the network packet is a DoS attack, with zero probabilities for other classes, including traffic decryption attacks, MiTM, and normal.

The right two-sided bar chart of a DoS attack uses the length of the bars to represent the magnitude of each feature's contribution to the prediction, with longer bars indicating a stronger influence. Green bars represent features that support the DoS classification, including *frame.time_relative* with a magnitude level more than 0.008, *wlan.fc.subtype* with a magnitude level around 0.006, *wlan.fc.protected* and *radiotap.timestamp.ts* with magnitude level of slightly less than 0.006, *radiotap.dbm_antsignal*, *radiotap.channel.flags.ckk* with a value of 0.0035 and lastly *wlan_radio.data_rate* with magnitude level of 0.002. Conversely, red bars indicate features that oppose to the prediction, such as *radiotap.present.tsft* with magnitude level of -0.006 and *wlan_radio.phy* with magnitude level around -0.0035. Based on the contributions of the magnitude level as per indicated in the right bar chart, it could be observed that most of the magnitude votings is towards the positive side in DoS classification and hence, TabNet model is able correctly classify the particular network packet as a DoS attempt.

Looking into the specific contributions of each feature, the local interpretation aligns with established principles in network security, as well as the global interpretation derived from SHAP values for DoS attack classification. In network security, certain features such as *wlan.fc.protected* being 0, indicating unprotected frames, and *wlan_radio.phy* being 1, indicating the utilization of physical radio settings, are crucial features indicators of potential DoS activity. Additionally, as previously discussed in the global interpretation of DoS attacks, the accurate classification of DoS attacks involves recognizing *wlan.fc.subtype* as a pivotal indicator. Furthermore, the positive direction on the bar of *radiotap.dbm_antsignal* shown in Figure 6 reinforces this classification, as this feature shows signal strength condition in real-time environment, which means that the model is capable to detect the abnormal signal strength occurred.

In contrast, Figure 7 for a false alarm scenario where a normal network packet is incorrectly classified as a DoS attack. The prediction probabilities show a 98% likelihood for the DoS class, with very low probabilities for other classes, despite the true label being 'Normal'. This misclassification highlights the model's error. The right-side two-sided bar chart shows that certain features negatively impact the classification of the packet as a DoS attack, suggesting it should be correctly classified as a normal packet. Specifically, the features *wlan.fc.subtype* with a magnitude slightly lower than -0.006, *wlan_radio.phy* with magnitude of approximately -0.006, *radiotap.present.tsft* with magnitude around -0.004, and *radiotap.dbm_antsignal* with magnitude around -0.002 contribute negatively to the DoS classification. However, the majority voting of the remaining features and magnitude are more towards to the positive direction, causing a false alarm scenario, whereby the normal packet is being misclassified as a DoS attempt.

Moving on to the perspective of the network security field, the NIDS model has identified key factors for correctly classifying the network packet as a normal packet. This scenario could be found in the feature *wlan.fc.subtype* and *radiotap.dbm.ant_signal* whereby the negative side in the right bar chart indicates that the model realizes that these features oppose the classification of the particular network packet as a DoS attack. To be more specific, the feature *wlan.fc.subtype* shows that a normal type of network packet is being transmitted, while the feature *radiotap.dbm.ant_signal* indicates that the signal strength of the network is actually normal. However, the NIDS model gets confused when certain features create ambiguity, such as *wlan.fc.protected* and *radiotap.channel.flags.ckk*.

In the feature *wlan.fc.protected*, this confusion arises when unencrypted frames are transmitted, which does not necessarily indicate an intrusion attempt. For instance, during the transmission of probe requests, the frames are unencrypted as part of normal network operations. Moreover, the feature *radiotap.channel.flags.ckk* indicates the presence of certain modulation schemes, which are common in both normal and attack scenarios. This overlap can lead the model to misinterpret normal modulation as a potential threat. Consequently, the model's challenge lies in disentangling these ambiguous signals to make a correct classification.

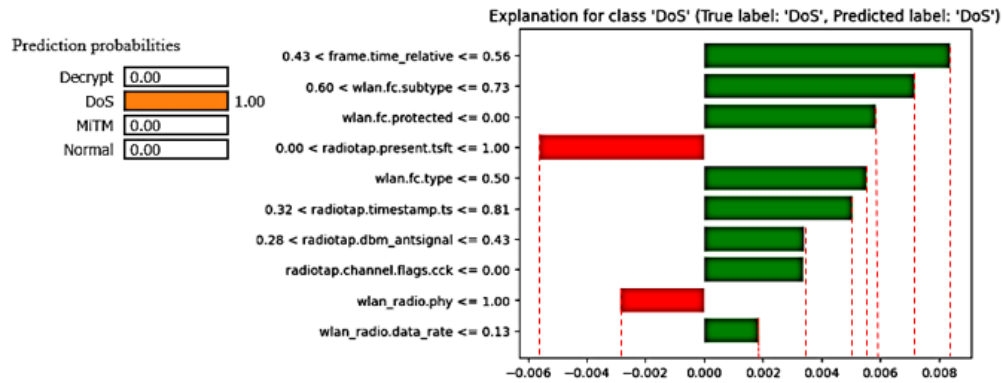


Figure 6. LIME interpretation for correct prediction in DoS attacks

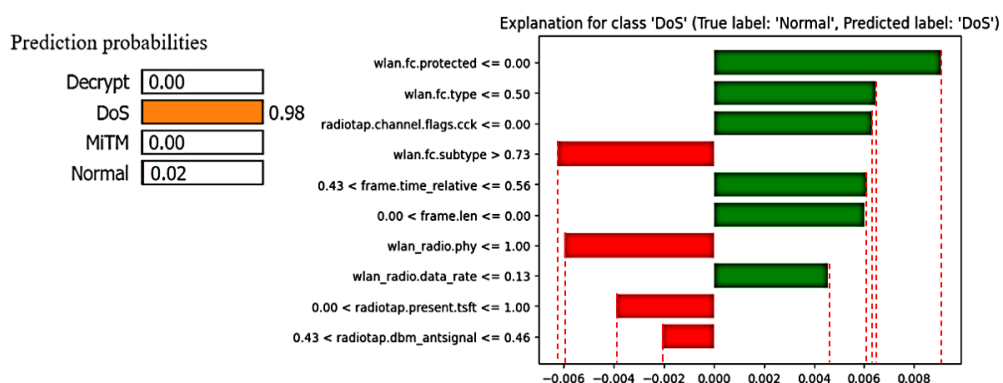


Figure 7. LIME interpretation for false alarm in DoS attacks

5. CONCLUSION

This paper presents a comprehensive study analyzing the performance of TabNet in NIDS model development. Using feature sets obtained through the phi-K method, the model achieved impressive accuracy, recall, and precision of 99.99%, surpassing the performance of four existing works. These results demonstrate that the TabNet model is highly effective in intrusion detection. Additionally, it successfully mitigates the risk of alarm fatigue, which is often caused by a large volume of false alarms.

The robustness of the model is evaluated by introducing SNR values ranging from 15 to 30. The model's consistent performance in terms of precision has demonstrated that the TabNet model can detect intrusions regardless of the noise level in the wireless channel. However, it is notable that the noise level can cause normal profiles to be misclassified as intrusion attempts. This is evident from the decrease in recall when the SNR value is lower, indicating that high noise level is present in the network profile.

Finally, the decision-making steps of the NIDS model are interpreted through the XAI approaches, particularly using SHAP and LIME. By examining both the global and local interpretations of the model, key features that significantly influence the model's predictions are identified. These insights help in understanding how the model differentiates between normal and malicious activities, thereby enhancing the transparency and trustworthiness of the intrusion detection process. From the perspective of NIDS interpretation, its method of determining abnormal traffic closely aligns with the way security experts identify abnormal traffic in real life. This demonstrates that it is reliable and effective in accurately detecting security threats.

As such, this study has investigated the reliability of NIDS through the comprehensive model and robustness evaluation, effective feature selection, and the integration of interpretability approaches. These contributions proven the capability of the TabNet model in enhancing network security by accurately identifying and mitigating various threats. Future work is suggested to focus on minimizing false alarms in noisy environments by using hybrid approaches. Specifically, rule-based methods should be integrated into the model to reduce false alarms across different noise levels.

FUNDING INFORMATION

This work was supported by the Telekom Malaysia Research and Development under Grant RDTG/241125 (MMUE/240066).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Juan-Kai Tan	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Lee-Yeng Ong					✓	✓	✓	✓		✓	✓	✓	✓	
Meng-Chew Leow		✓			✓		✓			✓				✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in AWID3 at 10.1109/ACCESS.2021.3061609, reference number [26].




REFERENCES

- [1] R. Ahmad, I. Alsmadi, W. Alhamdani, and L. Tawalbeh, "Zero-day attack detection: a systematic literature review," *Artificial Intelligence Review*, vol. 56, no. 10, pp. 10733–10811, Oct. 2023, doi: 10.1007/s10462-023-10437-z.
- [2] B. M. Serinelli, A. Collen, and N. A. Nijdam, "Training guidance with KDD cup 1999 and NSL-KDD data sets of ANIDINR: anomaly-based network intrusion detection system," *Procedia Computer Science*, vol. 175, pp. 560–565, 2020, doi: 10.1016/j.procs.2020.07.080.
- [3] A. Nisioti, A. Mylonas, P. D. Yoo, and V. Katos, "From intrusion detection to attacker attribution: a comprehensive survey of unsupervised methods," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3369–3388, 2018, doi: 10.1109/COMST.2018.2854724.
- [4] S. Roy, J. Li, V. Pandey, and Y. Bai, "An explainable deep neural framework for trustworthy network intrusion detection," in *2022 10th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud)*, Aug. 2022, pp. 25–30, doi: 10.1109/MobileCloud55333.2022.00011.
- [5] D. Gaspar, P. Silva, and C. Silva, "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," *IEEE Access*, vol. 12, pp. 30164–30175, 2024, doi: 10.1109/ACCESS.2024.3368377.
- [6] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, Jan. 2024, doi: 10.1007/s12559-023-10179-8.
- [7] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep learning approach for network intrusion detection in software defined networking," in *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Oct. 2016, pp. 258–263, doi: 10.1109/WINCOM.2016.7777224.
- [8] Y.-C. Wang, Y.-C. Hwang, H.-X. Chen, and S.-M. Tseng, "Network anomaly intrusion detection based on deep learning approach," *Sensors*, vol. 23, no. 4, p. 2171, Feb. 2023, doi: 10.3390/s23042171.
- [9] R. S. Ahmad, A. H. Ali, S. M. Kazim, and Q. Niyaz, "A GAF and CNN based Wi-Fi network intrusion detection system," in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHOPS)*, May 2023, pp. 1–6, doi: 10.1109/INFOCOMWKSHOPS57453.2023.10226036.
- [10] M. D. Hossain, H. Inoue, H. Ochiai, D. Fall, and Y. Kadobayashi, "LSTM-based intrusion detection system for in-vehicle can bus communications," *IEEE Access*, vol. 8, pp. 185489–185502, 2020, doi: 10.1109/ACCESS.2020.3029307.
- [11] B. Deore and S. Bhosale, "Hybrid optimization enabled robust CNN-LSTM technique for network intrusion detection," *IEEE Access*, vol. 10, pp. 65611–65622, 2022, doi: 10.1109/ACCESS.2022.3183213.
- [12] H. C. Altunay and Z. Albayrak, "A hybrid CNN+LSTM-based intrusion detection system for industrial IoT networks," *Engineering Science and Technology, an International Journal*, vol. 38, p. 101322, Feb. 2023, doi: 10.1016/j.jestech.2022.101322.
- [13] M. Graziani *et al.*, "A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3473–3504, Apr. 2023, doi: 10.1007/s10462-022-10256-8.
- [14] S. Hariharan, R. R. Rejmol Robinson, R. R. Prasad, C. Thomas, and N. Balakrishnan, "XAI for intrusion detection system: comparing explanations based on global and local scope," *Journal of Computer Virology and Hacking Techniques*, vol. 19, no. 2, pp. 217–239, Jul. 2022, doi: 10.1007/s11416-022-00441-2.
- [15] Z. A. El Houda, B. Brik, and L. Khokhi, "Why should i trust your IDS?: An explainable deep learning framework for intrusion




- detection systems in internet of things networks,” *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1164–1176, 2022, doi: 10.1109/OJCOMS.2022.3188750.
- [16] R. Younis, A. Ahmad, and Q. Abu Al-Haija, “Explaining intrusion detection-based convolutional neural networks using Shapley additive explanations (SHAP),” *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 126, Oct. 2022, doi: 10.3390/bdcc6040126.
- [17] P. Ramyavarshini, G. K. Sriram, U. Rajasekaran, and A. Malini, “Explainable AI for intrusion detection systems,” in *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, Dec. 2022, pp. 1563–1567, doi: 10.1109/IC3I56241.2022.10073356.
- [18] T. Senevirathna, B. Siniarski, M. Liyanage, and S. Wang, “Deceiving post-hoc explainable AI (XAI) methods in network intrusion detection,” in *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, Jan. 2024, pp. 107–112, doi: 10.1109/CCNC51664.2024.10454633.
- [19] P. Barnard, N. Marchetti, and L. A. DaSilva, “Robust network intrusion detection through explainable artificial intelligence (XAI),” *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, Sep. 2022, doi: 10.1109/LNET.2022.3186589.
- [20] M. Baak, R. Koopman, H. Snoek, and S. Klous, “A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics,” *Computational Statistics & Data Analysis*, vol. 152, p. 107043, Dec. 2020, doi: 10.1016/j.csda.2020.107043.
- [21] S. Arık and T. Pfister, “TabNet: attentive interpretable tabular learning,” *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 8A, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.
- [22] J. Price and T. Goble, “Signals and noise,” in *Telecommunications Engineer’s Reference Book*, Elsevier, 1993, pp. 10-1–10-15.
- [23] E. Chatzoglou, G. Kambourakis, C. Smiliotopoulos, and C. Kolias, “Best of both worlds: detecting application layer attacks through 802.11 and non-802.11 features,” *Sensors*, vol. 22, no. 15, p. 5633, Jul. 2022, doi: 10.3390/s22155633.
- [24] A. Mughaid *et al.*, “Improved dropping attacks detecting system in 5G networks using machine learning and deep learning approaches,” *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 13973–13995, Apr. 2023, doi: 10.1007/s11042-022-13914-9.
- [25] E. Chatzoglou, G. Kambourakis, C. Kolias, and C. Smiliotopoulos, “Pick quality over quantity: expert feature selection and data preprocessing for 802.11 intrusion detection systems,” *IEEE Access*, vol. 10, pp. 64761–64784, 2022, doi: 10.1109/ACCESS.2022.3183597.
- [26] E. Chatzoglou, G. Kambourakis, and C. Kolias, “Empirical evaluation of attacks against IEEE 802.11 enterprise networks: the AWID3 dataset,” *IEEE Access*, vol. 9, pp. 34188–34205, 2021, doi: 10.1109/ACCESS.2021.3061609.

BIOGRAPHIES OF AUTHORS






Tan Juan Kai    is currently studying Bachelor of Computer Science (Hons.) Artificial Intelligence in Multimedia University, Melaka, Malaysia. His area of interest includes machine learning, deep learning, sustainable AI and cybersecurity. He can be contacted at email: jktan1588@gmail.com.



Lee-Yeng Ong    received Ph.D. degree in computer vision from Multimedia University, Malaysia. She is currently working as an assistant professor with Faculty of Information Science and Technology, Multimedia University, Malaysia. Her research interests include object tracking, computer vision, data science and big data analytics. She can be contacted at email: lyong@mmu.edu.my.



Meng-Chew Leow    received his Doctor of Philosophy from Multimedia University. His research interest is in game-based learning, specifically role-playing game-based learning. He is also interested in system science, practical spirituality and philosophy. He can be contacted at email: mcleow@mmu.edu.my.