ATTENTION MECHANISM FOR ATTACKS AND INTRUSION DETECTION

Angham Alsuhaimi [1] and Jehan janbi [2]

[1] Department of Cyber Security, College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia
[2] College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia

ABSTRACT

In the rapidly evolving field of cybersecurity, effectively detecting and preventing network-based attacks is critical to safeguarding vital infrastructures. Traditional Network Intrusion Detection Systems (NIDS) often struggle to address the complexity and sophistication of modern cyber threats. To overcome these limitations, this thesis introduces a novel deep learning-based NIDS framework that integrates TabNet with an Attention Mechanism to improve both detection accuracy and interpretability. Leveraging the CIC UNSW-NB15 Augmented Dataset, which includes nine diverse attack categories alongside benign traffic, the proposed system employs advanced preprocessing techniques, such as SMOTE, to address class imbalance and data variability. Experimental results indicate that the model achieves an overall accuracy of 74%, excelling in the detection of benign traffic and reconnaissance attacks but encountering challenges with rare attack types, including worms and shellcode. These findings underscore the promise of attention-based deep learning models for enhancing NIDS performance and emphasize the need for future research to refine detection capabilities for rare and complex attacks.

KEYWORDS

Cybersecurity, Network Intrusion Detection Systems (NIDS), Deep Learning, TabNet, Attention Mechanism, SMOTE

1. INTRODUCTION

In our current era, the importance of cybersecurity has increased significantly, as the world faces increasing and diverse threats from cyberattacks targeting individuals, companies, and governments alike. Protecting data and systems is now more crucial than ever due to the quick development of technology and our growing reliance on the Internet in every part of our everyday lives. As it seeks to create a secure and sustainable digital society, the Kingdom of Saudi Arabia's Vision 2030 is a trailblazing move in improving cybersecurity. The Kingdom is working to develop integrated strategies to protect against cyber risks.

Artificial intelligence is one of the powerful tools that enhance cybersecurity today. Thanks to its capabilities in analyzing big data and discovering patterns, artificial intelligence can play a vital role in enhancing attack detection and response systems. The use of technologies such as machine learning and deep learning makes it possible to improve the accuracy of attack detection and analyze network behaviors faster and more effectively. These systems are able to identify and respond to assaults before they cause significant harm by analyzing vast volumes of data and identifying anomalous patterns in network traffic.

In the context of these developments, Network Intrusion Detection Systems (NIDS) emerge as vital tools in protecting networks from attacks. NIDS provide continuous network monitoring across on-premises and cloud infrastructure to detect malicious activity like policy violations, lateral movement or data exfiltration [1].

As a result, machine learning and deep learning techniques have become essential for identifying anomalous behaviors in network traffic that may indicate an attack. This proposal presents an innovative approach based on deep learning by using the TabNet model. TabNet is a deep tabular data learning architecture that uses sequential attention to choose which features to reason from at each decision step [2], which will be explained in detail in the next section. This approach is applied to CIC UNSW-NB15 Augmented Dataset containing modern network traffic, including nine different categories of attacks in addition to normal traffic [3].

Compared to more conventional models like support vector machines (SVMs) or convolutional neural networks (CNNs), TabNet offers a number of benefits, such as better feature interpretation, dynamic feature selection, and effective tabular data management. With these advantages, TabNet can perform better in detecting complex cyber-attacks, which enhances the ability of systems to counter such threats. The goal of this thesis is to increase the accuracy of cyberattack detection and improve overall cyber security by investigating preprocessing techniques, feature engineering, and model construction using TabNet.

2. LITERATURE REVIEW

This section reviews prior research on the TabNet model and the Attention Mechanism, highlighting their roles in improving machines and deep learning performance.

Laghrissi et al. [4] presents an innovative approach to intrusion detection systems (IDS) by integrating Long Short-Term Memory (LSTM) networks with an Attention mechanism, alongside four dimensionality reduction algorithms: Chi-square, UMAP, Principal Component Analysis (PCA), and Mutual Information. The research primarily focuses on addressing the high false negative rates commonly associated with IDS. Based on the experimental results, the Attention-PCA model with three components achieved a 98.49% accuracy rate for multiclass classification, while the Attention-based model with all features scored an impressive 99.09% accuracy rate for binary classification. The effectiveness of the suggested procedures on the NSL-KDD dataset is demonstrated by these results, which show a notable improvement in performance when compared to earlier solutions, especially in lowering false negatives.

Zhao et al. [5] presents a novel method for detecting Distributed Denial of Service (DDoS) attacks by integrating a self-attention mechanism with Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. The approach begins with feature selection using a Random Forest algorithm combined with Pearson correlation analysis to reduce input data redundancy. Subsequently, the model employs one-dimensional CNNs to extract spatial features and BiLSTM networks for temporal features, which are then fused in parallel. The attention mechanism improves the model's capacity to classify traffic by giving features varying weights according to their significance. The effectiveness of the proposed method was validated through binary and multi-classification experiments on the CIC-IDS2017 and CIC-DDoS2019 datasets. With the maximum accuracy, precision, recall, and F1 values recorded at 95.670%, 95.824%, 95.904%, and 95.864%, respectively, the results showed that the model outperformed other models in the literature and obtained excellent performance metrics.

Dai et al. [6] introduces the CNN-BiLSTM-Attention model, a unique network intrusion detection model that combines Convolutional Neural Networks (CNN), Bidirectional Long Short-

Term Memory (BiLSTM), and an attention mechanism. The model aims to address the challenges of low detection accuracy and high false positive rates prevalent in existing intrusion detection methods. By utilizing CNN for spatial feature extraction and BiLSTM for temporal feature mining, the model effectively captures the spatiotemporal characteristics of network intrusion data. By giving the extracted characteristics varying weights and highlighting the most important ones in the classification process, the attention mechanism improves the model's performance even further. The study evaluates the model on three public datasets: NSL-KDD, UNSW-NB15, and CIC-DDoS2019. The results demonstrate that the CNN-BiLSTM-Attention model achieves impressive performance metrics, including an accuracy of 99.79% on the NSL-KDD dataset, a detection rate of 99.83%, and a low false positive rate of 0.17%. Additionally, the introduction of Equalization Loss v2 as a loss function helps improve the model's detection rate for minority class data, addressing the issue of class imbalance in the datasets. The study's overall findings demonstrate how well these cutting-edge methods work together to improve intrusion detection capabilities.

Wang and Ghaleb [7] present an attention-based convolutional neural network (CNN) model designed for intrusion detection, addressing the growing need for effective network security mechanisms. The authors implemented a novel methodology that combines CNNs with attention mechanisms to enhance the model's computational efficiency and accuracy. They utilized image generation techniques to convert network traffic data into a structured image format, which allows the CNN to effectively analyze the data. The research was conducted using a subset of the CSE-CIC-IDS2018 dataset, and the results demonstrated that the proposed model could swiftly complete the detection process while maintaining high accuracy. Specifically, the model achieved a classification accuracy exceeding 96%, indicating that it effectively retained essential information from the original dataset. Additionally, the study highlighted the model's ability to process over 500 samples per second, showcasing its efficiency in real-time applications. All things considered, the CNN framework's effectiveness in identifying intrusions was greatly enhanced by the incorporation of attention mechanisms, which made it a useful addition to the field of network security.

Pillai et al. [8] focuses on enhancing IoT security detection for smart cyber infrastructures by implementing a deep learning approach using Long Short-Term Memory (LSTM) networks with an attention mechanism. The research addresses the limitations of traditional Intrusion Detection Systems (IDS), which often struggle with low detection effectiveness and the inability to adapt to new types of intrusions. The proposed method utilizes Information Gain (IG) for feature selection and employs z-score normalization to preprocess the data. The model was evaluated using the NSL-KDD and CICIDS-2017 datasets, achieving impressive results with accuracies of 99.70% and 99.60%, respectively. These findings demonstrate the efficacy of the LSTM with attention mechanism in detecting and classifying cyberthreats in IoT environments, showing a notable improvement in detection performance when compared to current methods.

Anandh et al. [9] focuses on enhancing intrusion detection systems (IDS) within the context of the Internet of Things (IoT). The researchers employed a feed-forward neural network with a single-layer attention mechanism, utilizing a weight vector parameterized as a real number and applying Leaky ReLU as the non-linearity function. Accuracy, recall, precision, F1-score, false positive rate, and execution time are among the performance measures that the study highlights as being crucial for both the training and testing stages. The proposed system demonstrated significant effectiveness in classifying network traffic, achieving a high accuracy rate, although the specific accuracy percentage is not detailed in the provided contexts. The results indicate that the system not only excels in correct classification but also processes data efficiently, minimizing packet loss during execution. The research contributes to the existing body of knowledge by

comparing and contrasting traditional and machine learning-based IDS methodologies, highlighting the challenges and future directions for IoT security solutions.

Bian et al. [10] addresses the issue of non-technical losses in power distribution caused by abnormal electricity consumption behaviours. The authors suggest a brand-new detection model called PSO-Attention-LSTM, which combines Particle Swarm Optimization (PSO) with Long Short-Term Memory (LSTM) networks that have been improved by an attention mechanism. The model aims to accurately predict normal electricity consumption and identify anomalies by analyzing deviations from predicted values. The study compares the PSO-Attention-LSTM model's performance to that of a number of alternative algorithms, including CNN-LSTM, Attention-LSTM, LSTM, Gated Recurrent Unit (GRU), Support Vector Regression (SVR), Random Forest (RF), and Linear Regression (LR), using a dataset from the University of Massachusetts. The findings corroborate the PSO-Attention-LSTM model's superior anomaly detection capabilities by showing that it performs better than the other approaches in terms of detection accuracy, obtaining a greater positive rate and a lower false positive rate. However, the study also notes that the model struggles with detecting "burr" points, which are influenced by significant noise and randomness, indicating areas for future improvement in detection accuracy and practical application.

Arık and Pfister [11] introduces TabNet, a novel deep learning architecture specifically designed for tabular data learning, which employs a sequential attention mechanism to select relevant features at each decision step. This approach enhances interpretability and efficiency by focusing the model's capacity on the most salient features. The researchers utilized the Adam optimization algorithm and Glorot uniform initialization for training, and they conducted experiments across various datasets for both classification and regression tasks, employing standard loss functions such as softmax cross-entropy and mean squared error . The results demonstrated that TabNet outperforms existing tree-based models, such as random forests and XGBoost, across multiple benchmarks, achieving significant performance improvements, particularly when leveraging unsupervised pre-training techniques. The study highlights the model's capacity to offer both local and global interpretability, which is critical for applications in delicate domains where comprehension of model decisions is critical, such as healthcare and finance. With its combination of excellent performance and interpretability, TabNet is an important development in the field of tabular data learning and a useful tool for both data scientists and decision-makers.

Nguyen et al. [12] offers a unique method for creating a lightweight intrusion detection system (IDS) that is tailored for Internet of Things (IoT) gateways by using the TabNet attention-based methodology. The research highlights the increasing vulnerability of IoT devices to cyber-attacks due to their limited computing resources, which prevents the direct use of traditional antivirus software. The authors conducted experiments using two datasets: BOT-IoT and UNSW-NB15, achieving impressive accuracy rates of 98.53% and 99.43% for classifying 11 subcategories and 5 main categories, respectively, in the BOT-IoT dataset. Additionally, the model demonstrated a 97.47% accuracy on the UNSW-NB15 dataset. The results indicate that the proposed approach not only outperforms existing methods, such as MidSIoT and IMIDS, in terms of F1-score for most classes but also maintains a lightweight characteristic suitable for deployment on resource-constrained devices like the Raspberry Pi 4. This innovative application of TabNet in the intrusion detection field marks a significant advancement in enhancing the security of IoT networks.

Nader and Bou-Harb [13] investigates the detection of malware-infected Internet-of-Things (IoT) bots operating behind Network Address Translation (NAT) gateways, utilizing large-scale one-way darknet data. The researchers employed an attentive interpretable tabular transformer model, specifically TabNet, which is known for its self-attention mechanism that enhances

interpretability while maintaining high performance. The study evaluated TabNet's performance against a number of conventional machine learning techniques, such as Multi-layer Perceptron (MLP), Random Forest (RF), and Logistic Regression. The results demonstrated that TabNet significantly outperformed these models, achieving an impressive accuracy of 93% on the October 20, 2021, dataset and 91% on the November 2021 dataset. Additionally, the study highlighted the identification of approximately 4 million Mirai-infected NATed IoT bots and 16,871 unique NATed IP addresses, underscoring the evolving nature of malware threats in the IoT landscape. The findings emphasize the importance of addressing attention and interpretability in machine learning approaches for better security in IoT networks.

Roh et al. [14] presents a novel self-supervised fault diagnosis method utilizing a TabNet architecture, specifically designed for multivariate time-series process data without the need for labelled data. The research focuses on the Tennessee Eastman Process (TEP), a simulated industrial chemical process, where the temporal information of the data is compressed using a Long Short-Term Memory (LSTM) structure. Even with a small amount of labeled data, the suggested approach achieves improved fault diagnostic accuracy and shows notable performance gains over conventional supervised learning techniques. The results indicate that the self-supervised approach can effectively leverage abundant unlabelled data, leading to a substantial enhancement in performance, particularly when labelled data is scarce. The study highlights the interpretability of the model, as the tree-based nature of TabNet allows for the identification of salient features influencing fault diagnosis outcomes. The overall results indicate that the approach performs better than current supervised learning models, demonstrating the promise of self-supervised learning in industrial applications, even though exact accuracy criteria are not specified in the settings given.

The reviewed studies demonstrate the significant advancements in leveraging the TabNet model and Attention Mechanisms to enhance machine learning and intrusion detection capabilities.

Table 1 shows a summary of the studies conducted on TabNet and the attention mechanism. Attention mechanisms, as shown in multiple studies, improve feature prioritization, leading to higher accuracy and lower false negative rates in cybersecurity applications. Integrations with models like CNN, BiLSTM, and LSTM further bolster the ability to analyze both spatial and temporal data characteristics.

Table 1: Summary of TabNet and Attention Mechanism studies.

Ref	Focus	Dataset	Key Findings
[10]	LSTM with Attention and dimensionality reduction for IDS	NSL-KDD	Accuracy of 99.09% (binary) and 98.49% (multiclass). Reduced false negatives.
[11]	DDoS detection with CNN, BiLSTM, and self-attention	CIC-IDS 2017, CIC-DDoS2019	Achieved 95.67% accuracy and high precision using Random Forest for feature selection.
[12]	CNN-BiLSTM-Attention model for spatiotemporal feature extraction	NSL-KDD, UNSW-NB15, CIC DDoS2019	Accuracy of 99.79%, low false positive rate (0.17%), and better detection for minority classes.
[13]	Attention-based CNN for intrusion detection	CSE-CIC-IDS2018	Over 96% accuracy and efficient real-time performance (>500 samples/sec).
[14]	LSTM with Attention for IoT intrusion detection	NSL-KDD, CICIDS-2017	Achieved 99.70% and 99.60% accuracies for the datasets, addressing limitations of traditional IDS.
[15]	Attention-based feed-forward neural network for IoT IDS	Not specified	High classification accuracy and efficiency, though specific percentages are not provided.
[16]	PSO-Attention-LSTM for detecting abnormal electricity consumption	University of Massachusetts	Superior anomaly detection accuracy but challenges with noisy data ("burr" points).
[17]	TabNet for tabular data learning with interpretability	Multiple datasets	Outperformed tree-based models like XGBoost and demonstrated strong interpretability.
[18]	Lightweight TabNet IDS for IoT gateways	BOT-IoT, UNSW-NB15	Accuracy of 98.53% and 99.43% (BOT-IoT) and 97.47% (UNSW-NB15), outperforming existing IoT detection methods.
[19]	TabNet for detecting malware-infected IoT bots	Darknet data	91-93% accuracy; identified 4 million Mirai-infected NATed IoT bots.
[20]	Self-supervised TabNet for fault diagnosis in industrial processes	Tennessee Eastman Process (TEP)	Enhanced performance in low-labeled data environments; utilized interpretability for industrial fault analysis.

TabNet has also proven effective in handling tabular data with its sequential attention approach, offering both high interpretability and competitive performance, as highlighted by its applications in intrusion detection and IoT security. The reviewed works underscore the adaptability of these techniques across datasets, showcasing their potential to address challenges like class imbalance and computational efficiency.

In conclusion, the studies validate the effectiveness of combining TabNet and Attention Mechanisms with other machine learning frameworks, marking a robust pathway for future research to improve intrusion detection systems, especially when dealing with complex and imbalanced datasets.

3. THE PROPOSED MODEL

The suggested model for improving Network Intrusion Detection System (NIDS) performance is presented in this section. To enhance the detection of different network assaults, both frequent and uncommon, the model combines an attention mechanism with the deep learning architecture TabNet. The attention method helps the model detect cyberattacks more accurately and

efficiently by concentrating on the most pertinent elements in the input. The model also addresses the class imbalance problem using SMOTE and under-sampling, ensuring that all types of attacks, including underrepresented ones, are effectively detected. This chapter outlines the model's design, components, and the methodologies used to build a more robust and accurate intrusion detection system.

3.1. Proposed Model

In this thesis, the proposed model aims to address the challenges in Network Intrusion Detection Systems (NIDS) by leveraging deep learning techniques, specifically **TabNet** with an integrated attention mechanism. This model seeks to improve the detection accuracy of various network attacks, including **Analysis**, **Backdoor**, **Denial of Service (DoS)**, **Exploits**, **Fuzzers**, **Generic**, **Reconnaissance**, **Shellcode**, and **Worms**, which are often missed by traditional NIDS. The following section outlines the key components, design principles, and methodologies that contribute to the proposed model. and the figure 1 shows the steps followed in the proposed model.

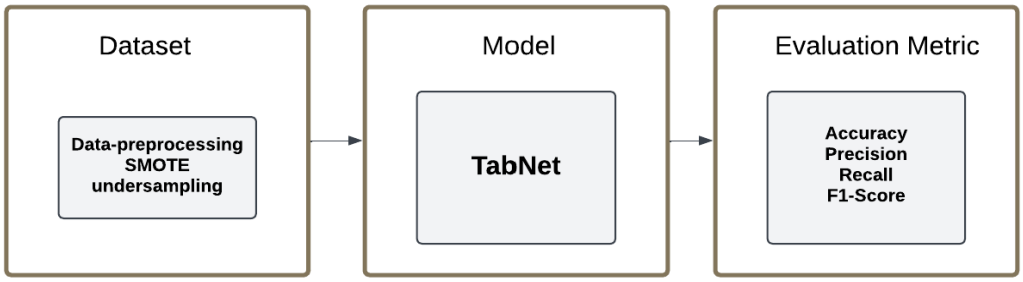

Figure 1: Steps followed in the proposed model.

3.2. Model Overview

The proposed model is built around TabNet, a deep learning architecture specifically designed for tabular data. In contrast to conventional neural networks, TabNet uses attention layers, which let it to focus on important patterns in the data and automatically choose pertinent features without requiring a lot of manual feature engineering. This capability makes TabNet particularly effective for detecting network intrusions, where identifying relevant patterns can be highly complex and vary significantly across different attack types.

To address the challenge of imbalanced attack distributions in the dataset, the model integrates the Synthetic Minority Oversampling Technique (SMOTE) and under-sampling. SMOTE ensures that underrepresented attack types are adequately represented during training, improving the model's ability to detect these attack types accuratelyIn the meantime, under-sampling is used on the majority class to keep it from dominating during training. This keeps the distribution of all classes balanced and improves the model's overall detection performance.

An attention method improves the model's capacity to dynamically rank features according to their significance at every decision stage, making the intrusion detection system more precise and comprehensible. The model is trained and tested on the CIC UNSW-NB15 Augmented Dataset, a comprehensive dataset widely used in NIDS research, containing diverse types of network attacks and benign traffic. The following figure shows the proposed model. the figure 2 presents the proposed model.

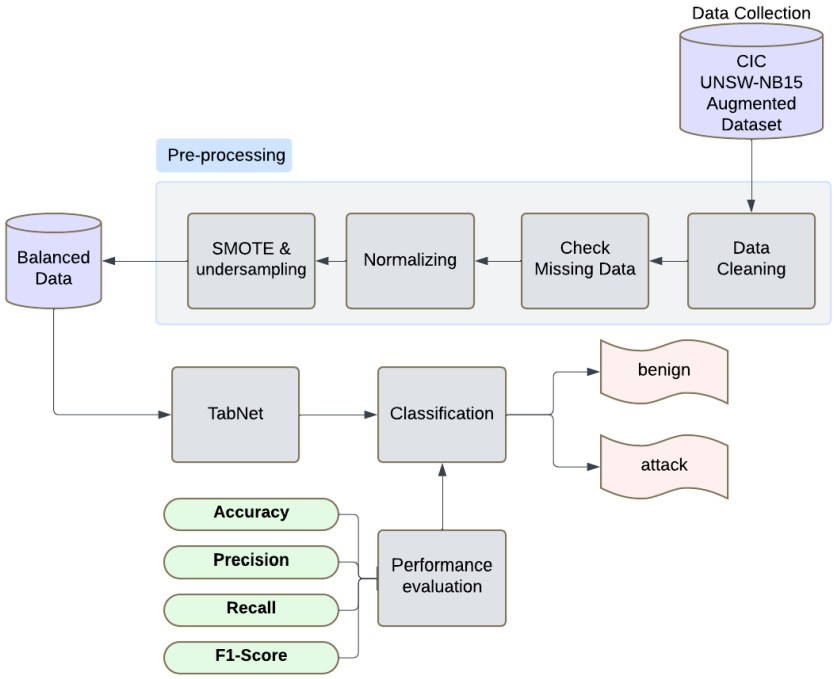

Figure 2: the proposed model.

By combining SMOTE for class balancing and TabNet's attention mechanism, the proposed model provides a robust framework for identifying complex attack patterns while addressing data imbalance issues.

3.3. Key Components of the Proposed Model

3.3.1. Attention Mechanism

The attention mechanism is at the heart of the TabNet model. It allows the model to focus on specific parts of the input data, ensuring that it considers the most relevant features for predicting network attacks. The attention layers dynamically adjust the importance of features during the learning process. This enables the model to handle complex, heterogeneous data effectively and make more precise decisions on attack classification.

Key features of the attention mechanism in the TabNet model include:

- Feature-wise attention: By giving distinct features varying degrees of importance, the model enhances its capacity to identify attacks by focusing on the most instructive features.
- Sparse attention: TabNet encourages sparse activation of attention, which reduces computational complexity and leads to faster learning.

By focusing on the most relevant features, the attention mechanism improves the model’s ability to distinguish between benign and malicious network traffic, particularly when dealing with imbalanced datasets or rare attacks.

3.3.2. Handling Class Imbalance with SMOTE

One of the significant challenges in network intrusion detection is the class imbalance problem. In most real-world datasets, benign traffic vastly outnumbers the malicious attack samples, leading to a bias in the model's predictions. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is applied.

SMOTE uses the feature space of existing samples to create artificial examples of the minority class (malicious attacks). In addition to ensuring that the model is trained on a more equitable representation of both malicious and benign traffic, this helps balance the dataset.The advantages of using SMOTE in this context include:

- Increased detection of rare attacks: SMOTE reduces the likelihood of the model being biased towards benign traffic, which is crucial for detecting various attack types, such as Reconnaissance, Shellcode, Exploits, and Worms.
- Improved model generalization: By balancing the data, the model learns better generalizations, making it more robust in detecting new and unknown attacks.

3.3.3. TabNet Model Architecture

- The TabNet architecture consists of several key components that make it a powerful tool for intrusion detection:
- Input Layer: This layer is responsible for feeding the dataset into the model, which includes both attack and benign traffic data from the CIC UNSW-NB15 Augmented Dataset.
- Attention Blocks: The core of TabNet consists of multiple attention blocks, where each block computes a set of attention weights to decide which features to focus on during processing. These blocks are structured to process the data efficiently by creating a sparse representation of the input.
- Decision Trees: After the attention mechanism, the model uses decision trees to classify the input data. These trees are built on the features that the attention mechanism has identified as important.
- Output Layer: The output layer produces the final classification result, predicting whether the given input corresponds to a benign or malicious network packet.

3.3.4. Data Preprocessing and Feature Engineering

A crucial first step in creating a successful intrusion detection model is data preprocessing. The model's performance is greatly impacted by the caliber of the data that is fed into it. For this thesis, the following preprocessing techniques are applied:

• Normalization and Encoding

Since the dataset contains features with varying ranges, Min-Max Scaling will be applied to scale all features between 0 and 1. This will prevent features with large ranges (e.g., Flow Duration) from dominating features with smaller ranges (e.g., Flag Counts). To make them appropriate for the deep learning model, any categorical features—like protocol types or flag counts—will be encoded using One-Hot Encoding.

3.3.5. Model Training

The model is trained using the prepared dataset, which is split into training and testing sets. In order to minimize the loss function during the training phase, a gradient-based optimization technique (like Adam) is used. The TabNet model is trained over several epochs, with adjustments to the model's weights made after each iteration based on the error between predicted and actual values.

Several metrics, including accuracy, precision, recall, and F1-score, are used to track the training process. Special attention is given to the detection of rare attacks, where the model's performance on these attack types is tracked separately.

3.3.6. Evaluation Metrics

To evaluate the performance of the model, several key metrics are used:

- Accuracy: Measures the overall correctness of the model, calculating the percentage of correctly classified instances [15].
- Precision: Measures the proportion of correctly predicted attacks out of all predicted attacks [15].
- Recall: evaluates the model's capacity to identify all real threats, including uncommon ones like exploits, backdoors, fuzzers, and denial-of-service assaults [15].
- F1-Score: An equilibrium between recall and precision that is particularly helpful when working with unbalanced datasets [15].

These metrics are used to evaluate the model's overall efficacy in identifying both common and uncommon attacks in NIDS.

3.4. Proposed Model's Contribution

The proposed model contributes to the field of network intrusion detection in several keyways:

- Improved Detection of Multiple Attack Types: The model can concentrate on the characteristics that are most important for identifying a variety of threats, including analysis, backdoor, denial-of-service, exploits, fuzzers, generic, reconnaissance, shellcode, and worms, by integrating the attention mechanism.
- Handling Imbalanced Data: The use of SMOTE and under-sampling helps address the challenge of class imbalance, ensuring that the model does not become biased towards benign traffic and that it has sufficient training data for detecting various attacks.
- Efficiency and Interpretability: Because of its efficiency and interpretability, the TabNet model can be used in real-world settings where it is essential to comprehend how the model makes decisions.
- Enhanced Generalization: By using advanced feature selection techniques and data balancing methods, the model generalizes well to new attack patterns, offering a robust solution for NIDS.

3.5. Model Deployment and Real-World Applications

Once the model is trained and evaluated, it can be deployed as part of a real-time intrusion detection system. The model will be able to monitor network traffic and flag potential attacks in real-time, providing cybersecurity professionals with actionable insights. Additionally, the interpretable nature of the TabNet model allows network administrators to understand why

certain network traffic is classified as malicious, aiding in decision-making and incident response.

4. EXPERIMENT AND EVALUATION

Fur In this section, I developed and evaluated a TabNet-based approach for classifying network traffic as either benign or one of several attack categories using the CIC UNSW-NB15 Augmented Dataset. The objective was to leverage TabNet's attention mechanism to improve detection accuracy, especially for rare and complex attack types, which traditional NIDS models often struggle to classify accurately.

4.1. Data Acquisition and Preprocessing

The dataset, which included network traffic data labeled with ten distinct attack categories in addition to benign traffic, was obtained from the CIC UNSW repository. The dataset contained 76 features representing various network flow characteristics such as packet counts, byte sizes, and flag metrics, with an initial high imbalance skewed towards benign traffic as show in figure 3. For example, benign samples constituted around 80% of the dataset, complicating accurate detection of minority attack classes.

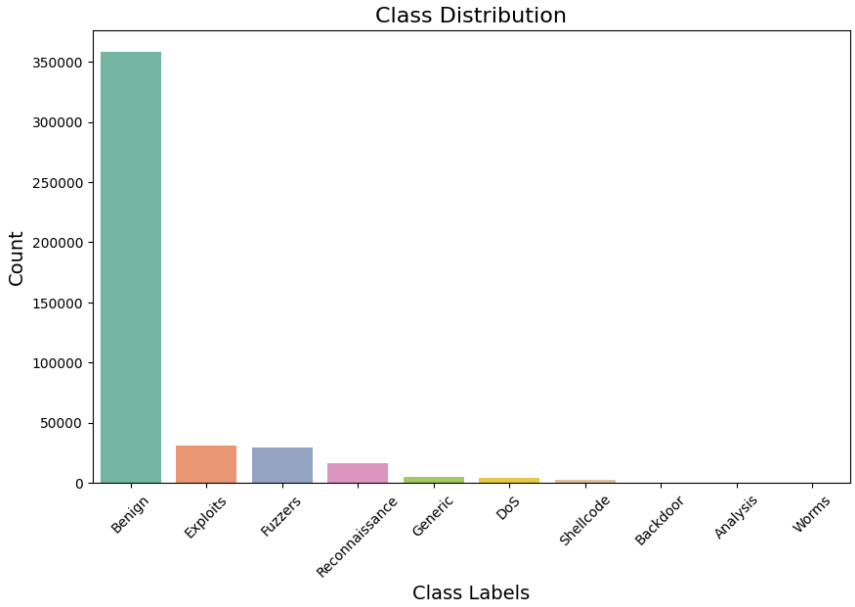

Figure 3: class Distribution.

Data preprocessing involved:

1. The significant class imbalance was addressed in two stages. First, I under-sampled the majority benign class to ensure a more balanced dataset, followed by synthetic oversampling using SMOTE for minority attack classes. This hybrid approach was essential to enhance model performance across all classes without sacrificing benign detection. Figure 4 shows the class distribution before and after SMOTE and Under-sampling.
2. normalized the feature space using Min-Max scaling to ensure uniform feature ranges, thereby preventing features with larger numerical ranges from disproportionately influencing the model's learning.

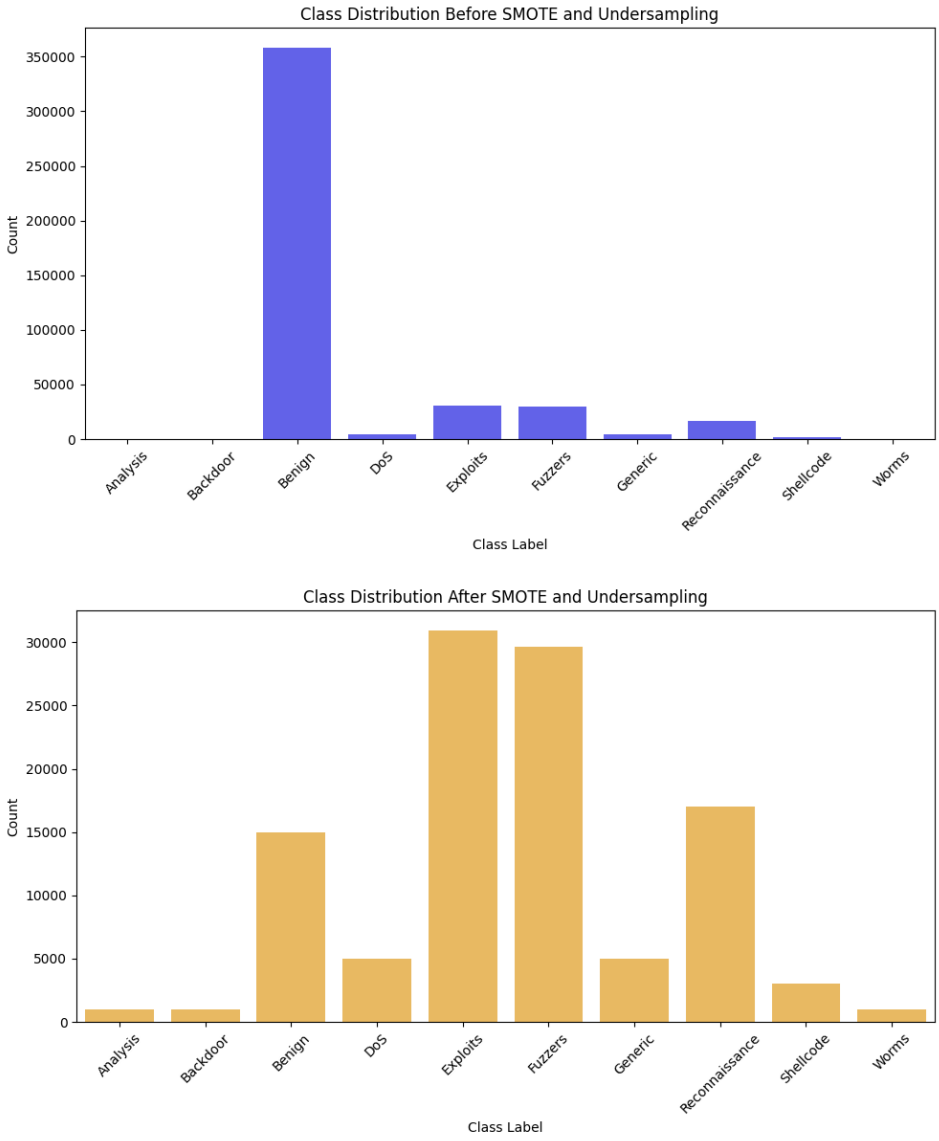

Figure 4: Class Distribution before/after SMOTE and Under-sampling.

4.2. Model Training and Evaluation

TabNet was selected due to its attention-based architecture, which dynamically focuses on the most relevant features for each sample. This dynamic feature selection is crucial for intrusion detection, where different attacks may reveal distinct patterns across various features.

4.2.1. Experiment 1: TabNet Model on the 10-Class Dataset

In the experiment, we trained TabNet on the 10 original classes. Using a batch size of 2048 and a learning rate of 0.0001, the model was trained with early stopping to prevent overfitting. I tracked the model's performance across epochs, noting improvements up to epoch 147, where early stopping occurred with a validation accuracy of 74%. The classification report in table 2, indicated strong detection rates for benign traffic but significant challenges in identifying minority classes, such as "Backdoor" and "Worms".

Table 2: Classification Report.

Label	Precision	Recall	F1-Score	Support
Analysis	0.18	0.99	0.3	77
Backdoor	0.04	0.6	0.08	90
Benign	1	0.97	0.99	71666
DoS	0.2	0.23	0.22	894
Exploits	0.9	0.46	0.61	6190
Fuzzers	0.74	0.43	0.55	5923
Generic	0.5	0.6	0.55	927
Reconnaissance	0.77	0.66	0.71	3347
Shellcode	0.18	0.47	0.26	420
Worms	0.01	0.88	0.02	49
macro avg	0.69	0.55	0.57	21713
weighted avg	0.74	0.74	0.72	21713
accuracy	0.74%			

Despite achieving high overall accuracy, the f1-scores for rarer attacks were lower, As shown in the figure 5. highlighting limitations in detecting infrequent classes within a highly imbalanced setting. This suggested that even with SMOTE oversampling, the complex, multi-class nature of the dataset limited TabNet's efficacy for certain attack types.

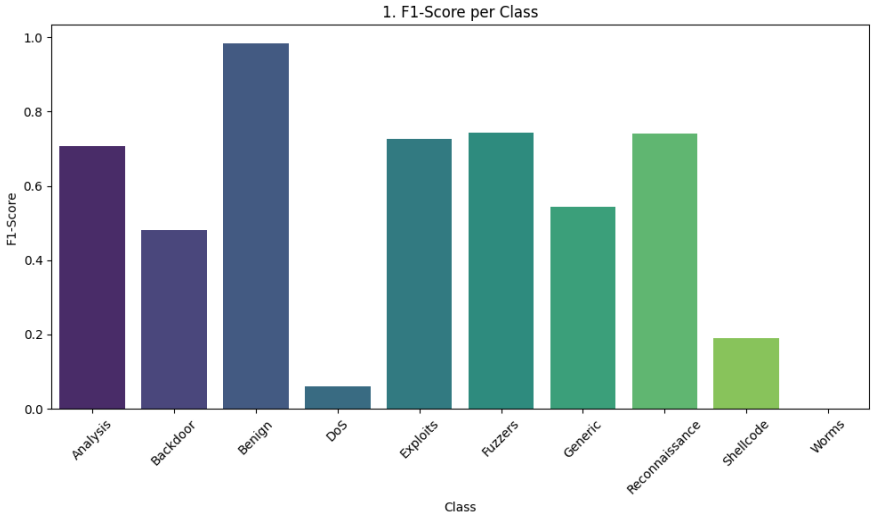

Figure 5: F1-Score per Class.

4.3. Observations and Results

The results from the TabNet model, after applying SMOTE and under-sampling techniques, reveal several key insights into its performance across the different attack categories:

- Benign Class: The model achieves perfect precision (1.00) and high recall (0.97), indicating it is very effective at identifying benign traffic. This is crucial, as false

positives in benign class predictions can lead to significant operational disruptions, which the model handles well.

- Reconnaissance and Generic: These classes also show relatively strong performance, with high precision (0.88 and 0.73 respectively) and reasonable recall (0.65 and 0.52). This suggests the model can detect these types of attacks fairly well, although there is still room for improvement in recall to capture all possible instances.
- Exploits and Fuzzers: For these categories, the model demonstrates good precision (0.72 and 0.64) and high recall (0.75 and 0.90), which means it is quite effective at detecting these attacks with fewer false negatives, though the precision for Fuzzers could still be optimized.
- DoS, Shellcode, and Worms: The performance on these attacks is less impressive. The precision and recall for DoS (0.53 and 0.04), Shellcode (0.53 and 0.14), and Worms (0.65 and 0.38) are much lower. This suggests that the model has trouble with these kinds of attacks, particularly when it comes to recall, when it fails to detect a sizable portion of real-world occurrences.
- This is a critical issue since DoS and Shellcode attacks can have severe impacts on network operations and require better detection capabilities.
- Backdoor: The model has the lowest performance in detecting Backdoor attacks (precision: 0.59, recall: 0.40), meaning that the model tends to miss many Backdoor instances while still correctly identifying some. This attack type requires more focus to improve detection accuracy.

4.4. Overall Performance

- The accuracy of the model is 74%, which means that about three-quarters of all predictions are correct. However, this figure can be misleading, as it doesn't reflect the performance on the rare attack categories where the model underperforms.
- The macro average precision (0.69) and recall (0.55) indicate that, although the model does well in certain classes, it has significant issues in others, leading to an imbalance in performance across all classes.
- The weighted average (precision: 0.74, recall: 0.74, F1-score: 0.72) reflects the overall class distribution and shows that the model is more balanced when considering the prevalence of each class, but it still struggles with rare or complex attack types.

4.5. Conclusion

While the TabNet model performs well in certain attack categories (especially benign traffic and reconnaissance), it has notable weaknesses in detecting attacks like DoS, Shellcode, and Backdoor, which are critical for network security. In order to achieve a more balanced and dependable detection system across all categories, the model's capacity to identify uncommon or complicated attack types needs to be enhanced. other tuning and the potential integration of other methodologies are also required.

5. CONCLUSIONS AND FUTURE WORK

This thesis explored the use of the TabNet deep learning model for developing an intrusion detection system (NIDS) to identify various network attacks. By applying the CIC UNSW-NB15 Augmented Dataset, the study addressed challenges such as data preprocessing, class imbalance, and model evaluation. Techniques like SMOTE were used to improve data quality and ensure effective model training. The results showed that TabNet achieved 74% accuracy overall, with

strong performance in detecting benign traffic and reconnaissance attacks, but challenges in identifying rare attacks like DoS, Shellcode, and Backdoor.

Future work will focus on improving the model's accuracy and generalization to handle a broader range of attacks, especially rare and complex ones. Hyperparameter tuning, which involves modifying the learning rate and batch size to optimize the model's learning process, is one of the main areas for development.

Addressing class imbalance with more advanced oversampling methods or hybrid sampling approaches can improve detection of infrequent attack patterns. Expanding the dataset to include a wider variety of real-world attack scenarios will further strengthen the model's adaptability to emerging cyber threats. These enhancements aim to create a more accurate, reliable, and interpretable intrusion detection model suited to the evolving landscape of cybersecurity.

AUTHOR

Alsuhaimi.A received a B.Sc. degree in Information Systems from the College of Computer Science and Engineering at Taibah University, Madinah, Saudi Arabia. She is currently pursuing an M.Sc. degree in Cybersecurity at Taif University, Taif, Saudi Arabia.

Dr. Janbi.J. is an assistant professor in Computer and Information Technology college in Taif University (TU). She is mastering substantial teaching and administrative skills. She became a fellow of the higher education academy. She occupied several administrative positions and became a member of several university level committees. Now, she is holding the position of vice-dean of the college of Computer and Information Technology in TU. In research, her main interest is in Pattern Recognition, dealing with image analysis and processing. In her PhD dissertation, she worked on encoding Arabic digital font's design characteristics into a number composed of several digits to enhance manipulating and searching fonts based on their appearance. Currently, she is working on medical image analysis using deep learning.

REFERENCES

[1] Redscan Ltd., "NIDS | Network Intrusion Detection System | Redscan," Redscan, Aug. 14, 2024. Available: https://www.redscan.com/services/nids/

[2] "Papers with Code - TabNet Explained." Available: https://paperswithcode.com/method/tabnet

[3] "UNSW-NB15 Augmented Dataset | Datasets | Research | Canadian Institute for Cybersecurity | UNB." Available: https://www.unb.ca/cic/datasets/cic-unsw-nb15.html

[4] F. Laghrissi, S. Douzi, K. Douzi, and B. Hssina, "IDS-attention: an efficient algorithm for intrusion detection systems using attention mechanism," Journal of Big Data, vol. 8, no. 1, Nov. 2021, doi: 10.1186/s40537-021-00544-5. Available: https://doi.org/10.1186/s40537-021-00544-5

[5] J. Zhao, Y. Liu, Q. Zhang, and X. Zheng, "CNN-AttBiLSTM Mechanism: A DDoS Attack Detection Method Based on Attention Mechanism and CNN-BiLSTM," IEEE Access, vol. 11, pp. 136308–136317, Jan. 2023, doi: 10.1109/access.2023.3334916. Available: https://doi.org/10.1109/access.2023.3334916

[6] W. Dai, X. Li, W. Ji, and S. He, "Network Intrusion Detection Method Based on CNN, BiLSTM, and Attention Mechanism," IEEE Access, vol. 12, pp. 53099–53111, Jan. 2024, doi: 10.1109/access.2024.3384528. Available: https://doi.org/10.1109/access.2024.3384528

[7] Z. Wang and F. A. Ghaleb, "An Attention-Based Convolutional Neural Network for Intrusion Detection Model," IEEE Access, vol. 11, pp. 43116–43127, Jan. 2023, doi: 10.1109/access.2023.3271408. Available: https://doi.org/10.1109/access.2023.3271408

[8] S. E. V. S. Pillai, K. Polimetla, C. S. Prakash, P. K. Pareek, and P. P. Pawar, "IoT Security Detection and Evaluation for Smart Cyber Infrastructures Using LSTMs with Attention Mechanism," Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), vol. 136, pp. 1–5, Apr. 2024, doi: 10.1109/icdcece60827.2024.10548639. Available: https://doi.org/10.1109/icdcece60827.2024.10548639

[9] R. Anandh, V. S. Rane, S. S. Rane, S. Vijayakumar, T. P, and N. Gopinath, "Modelling a Novel Linear Transformed Attention Mechanism for Intrusion Detection Using Learning Approach," Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT), pp. 462–466, Aug. 2024, doi: 10.1109/ic2sdt62152.2024.10696639. Available: https://doi.org/10.1109/ic2sdt62152.2024.10696639

[10] J. Bian, L. Wang, R. Scherer, M. Wozniak, P. Zhang, and W. Wei, "Abnormal Detection of Electricity Consumption of User Based on Particle Swarm Optimization and Long Short Term Memory With the Attention Mechanism," IEEE Access, vol. 9, pp. 47252–47265, Jan. 2021, doi: 10.1109/access.2021.3062675. Available: https://doi.org/10.1109/access.2021.3062675

[11] S. Ö. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687, May 2021, doi: 10.1609/aaai.v35i8.16826. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16826

[12] T.-N. Nguyen, K.-M. Dang, A.-D. Tran, and K.-H. Le, "Towards an Attention-Based Threat Detection System for IoT Networks," in Communications in computer and information science, 2022, pp. 301–315. doi: 10.1007/978-981-19-8069-5_20. Available: https://doi.org/10.1007/978-981-19-8069-5_20

[13] C. Nader and E. Bou-Harb, "An attentive interpretable approach for identifying and quantifying malware-infected internet-scale IoT bots behind a NAT," The 19th ACM International Conference on Computing Frontiers, May 2022, doi: 10.1145/3528416.3530995. Available: https://doi.org/10.1145/3528416.3530995

[14] H. R. Roh, J. M. Lee, and School of Chemical and Biological Engineering, Seoul National University, "TabNet-based Self-supervised Fault Diagnosis in Multivariate Time-series Process Data without Labels," Jul. 2024.

[15] GeeksforGeeks, "Evaluation Metrics in Machine Learning," GeeksforGeeks, Jul. 03, 2024. Available: https://www.geeksforgeeks.org/metrics-for-machine-learning-model/