JOURNAL OF DIGITAL SYSTEM DEVELOPMENT

https://e-journal.uum.edu.my/index.php/jdsd

How to cite this article:
Hwung, T.C. & Yuhanis Yusof (2024). Comparative of Data Mining Models for Intrusion Detection. *Journal of Digital System Developments*, *2*(1), 140-153. https://doi.org/10.32890/jdsd2024.2.1.9

COMPARATIVE OF DATA MINING MODELS FOR INTRUSION DETECTION

[1]Teoh Chun Hwung & [2]Yuhanis Yusof
[1&2]School of Computing, Universiti Utara Malaysia

[1]Corresponding author: yuhanis@uum.edu.my

Received: 17/3/2024 Revised: 2/4/2024 Accepted: 16/4//2024 Published: 30/4/2024

ABSTRACT

In the ever-evolving landscape of network security, the role of Intrusion Detection Systems (IDSs) is critical. These systems serve as the guardians of digital networks, defending against a relentless tide of cyber threats. However, the efficacy of IDSs in accurately detecting and preventing these threats remains a critical concern. The emergence of new attack vectors and the growing sophistication of cyberattacks underscore the need for innovative approaches to intrusion detection. Nonetheless, there is an information vacuum about the efficacy of different data mining approaches in intrusion detection. This study investigates the deployment of seven data mining models, which include Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and two types of Recurrent Neural Networks (RNN) on five (5) IDS benchmark datasets. Based on the undertaken experiments, it is evident that the RNN is the best classifier as it obtained 100% accuracy for all datasets. Due to its strength in modelling time-dependent and sequential data problems, RNN has become powerful in predicting future attacks. As our digital world evolves, so must our cyber defences; hence, this study strives to equip network security professionals with the knowledge and tools needed to fortify their networks, ensuring resilience against the ever-adaptive landscape of cyber threats.

Keywords: Intrusion Detection, cyber threats, cyberattack, data mining, network security

INTRODUCTION

Networks need Intrusion Detection Systems (IDSs) to be safeguarded against cyberattacks. However, existing IDSs often struggle to detect new and emerging attack types accurately, leading to false positives and security breaches (Chou & Jiang, 2022). This is a significant problem as cyberattacks are becoming increasingly sophisticated and targeted (Zipperle et al., 2022). Data mining methods can enhance the effectiveness of IDSs by discerning network traffic patterns that signal potential malicious behaviour (Duraz et al., 2023). Data mining can also be used to develop models that can classify different types of network traffic, such as legitimate, malicious, and suspicious traffic. Nevertheless, there is a shortfall of studies on the efficacy of various data mining approaches for intrusion detection despite the potential of these techniques (Chou & Jiang, 2022).

This study assessed the effectiveness of various data mining models on IDS benchmark datasets, including the deep learning model for intrusion detection. The study also examined how data provenance and quality affect how well data mining-based IDSs work. In particular, we focus on supervised data mining models deployed on historical labelled data where the analysis was on Recurrent Neural Network (Tsantekidis et al., 2022; Yin et al., 2017), Decision Tree (Fan, 2021), (Lei et al., 2021) Random Forest, Support Vector Machines (SVM) (Schlag et al., 2021), Logistic Regression (Zhang & Qin, 2022), and Naïve Bayes(Guo, 2022). This article is structured into four sections, beginning with literature studies in the related area, followed by the study's methodology. This is followed by a presentation and discussion of the experimental results, and the study ends with a conclusion.

RELATED WORK

Intrusion detection systems (IDSs) are crucial for defending networks against assaults via the internet. IDSs monitor network traffic for suspicious activity and alert security analysts to potential threats. Several studies (Chou & Jiang, 2022; Cihan et al., 2021; Mehta et al., 2023; Yin et al., 2017; Zhao & Gan, 2021) have investigated the application of data mining methods for intrusion detection. For instance, they assessed the effectiveness of different deep learning and machine learning models for intrusion detection. Using the NSL dataset, the study discovered that deep learning techniques performed better than machine learning algorithms in terms of accuracy (CİHAN et al., 2021). The deep learning models, nevertheless, require more computational cost.

Liu, Huang, and Zhu emphasised the limitations of traditional IDSs in combating modern cyberattacks (Liu et al., 2022). They proposed an IDS based on machine learning that leverages historical attack data to identify novel and emerging threats. The research illustrates the efficacy of the suggested intrusion detection system (IDS) in attaining elevated detection precision for known and new threats through the utilisation of diverse machine learning models, such as random forests, decision trees, and support vector machines (SVMs) (Liu et al., 2022). In response to the specific challenges in intrusion detection within Controller Area Networks (CANs), studies have introduced DT-DS (Mehta et al., 2023), an IDS that employs decision tree ensembles to learn CAN network behaviour and identify anomalies indicative of attacks. The DT-DS extracts timing and frequency data from CAN traffic and trains decision tree models for real-time operation. The proposed IDS (Mehta et al., 2023) exhibits high accuracy in detecting known and unknown attacks while maintaining efficient cost, making it suitable for deployment in real-time CAN environments.

The Random Forest algorithm is a comprehensive ensemble learning strategy known for its reliability and capacity to handle classification and regression tasks well. It integrates many decision trees, each trained on a separate subset of the data and attributes, to create an effective ensemble-based classifier. Using an ensemble technique, overfitting is lessened, generalisation is strengthened, and the model's overall accuracy and dependability are improved. Additionally, Random Forest provides feature importance scores, identifying critical attributes contributing to intrusion detection. Yang, Cai, Duan, and Yang proposed an intrusion detection algorithm that combines approximate information entropy with random forest classification(Yang et al., 2019). The approach reduces dimensionality and eliminates noise in training data, enhancing classification accuracy and reducing time complexity. Experiment outcomes demonstrated that their method achieved high accuracy on the KDD-CUP99 dataset and maintained computational efficiency, contributing to developing effective and efficient intrusion detection techniques.

The Support Vector Machine, or SVM, is a widely recognised and robust machine learning algorithm. It is particularly suitable for binary and multiclass classification tasks, making it relevant for intrusion detection. To successfully separate various classes in high-dimensional feature spaces, SVM seeks an ideal hyperplane that optimises the margin between them. This characteristic is beneficial when dealing with complex and nonlinear patterns in network traffic data. SVM's ability to handle high dimensional data and its potential to generalise well are key factors in its inclusion in this study. Shah et al. (2021) introduced a semi-supervised IDS that employs support vector machines (SVMs) and random forests (RFs) (Shah et al., 2021). This IDS addresses the challenge of acquiring large, labelled datasets by leveraging semi-supervised learning. The incoming network traffic is then categorised as normal or malicious using RF, after which the SVM model is expanded with unlabeled data. The SVM model was initially trained with limited labelled data. The semi-supervised IDS exhibits high accuracy for known and unknown attacks on benchmark datasets.

Another classifier deployed in IDS is the Naïve Bayes classifier, a probabilistic algorithm grounded in Bayes' theorem (Ahsan et al., 2022). Despite its simplicity, Naïve Bayes operates on the feature independence assumption, simplifying calculations, but it might not always hold in real-world scenarios. However, in the real world, the properties are rarely independent, necessitating the development of classifiers that consider this restriction. Naive Bayes classifiers (NBCs) are based on Bayes' theorem, with the assumption that the value of one predictor (x) does not affect the probability of a given class (c), given the values of the other predictors.

Logistic Regression is a fundamental algorithm in binary and multiclass classification. Because it forecasts the probability of a binary outcome, it has significant potential for intrusion detection, where the goal is often to classify network data as either genuine or malicious. Logistic regression employs a linear combination of input features to determine the likelihood of an event occurring. Despite its simplicity, it can yield interpretable results and is computationally efficient. Regression analysis emerges as a valuable tool for intrusion detection, as evident in its application in modelling network behaviour and identifying anomalies suggestive of attacks (Nagaraja et al., 2021). The work demonstrates the high accuracy of regression analysis in detecting attacks.

Despite the traditional machine learning classifiers, recent Deep Learning technology has also attracted IDS researchers' attention. The Recurrent Neural Networks (RNNs) is an example of a deep learning model well suited for processing sequential data. RNNs can incorporate information from prior inputs, which enables them to capture temporal dependencies in the data, in contrast to typical feedforward neural networks, which handle each input individually. This makes RNNs ideal for a variety of tasks, including natural language processing (NLP), speech recognition (Shewalkar et al., 2019; Wang, 2023), and time series forecasting (Yu et al., 2021). Another example of using RNNs is extracting meaningful patterns and insights from agricultural datasets (Mohamed & Mohamud, 2022).

Based on the strength of data mining models reported in the literature, this study intends to compare the model's performance on a list of IDS benchmark datasets. Furthermore, the analysis will be based on the performance between traditional models (i.e. Decision Tree, Logistic Regression, SVM, Random Forest and Naïve Bayes) and the recent technology of Deep Learning. Comparing and contrasting traditional data mining and deep learning models is essential to comprehending their relative advantages and disadvantages while classifying cybersecurity activities. Making well-informed decisions about which model is best for intrusion detection, maximising performance, and providing alternative AI-based solutions are made possible by this kind of study.

METHODOLOGY

Research procedures outline the methods and procedures used to collect, analyse, and validate data to guarantee the accuracy and dependability of research. This study's methodology is structured into four phases: Data Acquisition, Data Cleaning, Model Implementation, and Model Evaluation. Each is designed to facilitate the comparative analysis of various data mining models for intrusion detection.

Data Acquisition

In this initial phase, five benchmark intrusion detection datasets, namely the NSL KDD (CİHAN et al., 2021), KYOTO (Song et al., 2011), CIC-DDoS 2019 (Sharafaldin et al., 2019), UNSW-NB15 (Meftah et al., 2019), and CSE-CIC-IDS2018 (Herrera et al., 2022), are obtained from the relevant source. These datasets encompass many network traffic scenarios, enabling a diverse and comprehensive analysis. Table 1 depicts the summary of the deployed datasets.

Table 1

Summary of IDS Dataset

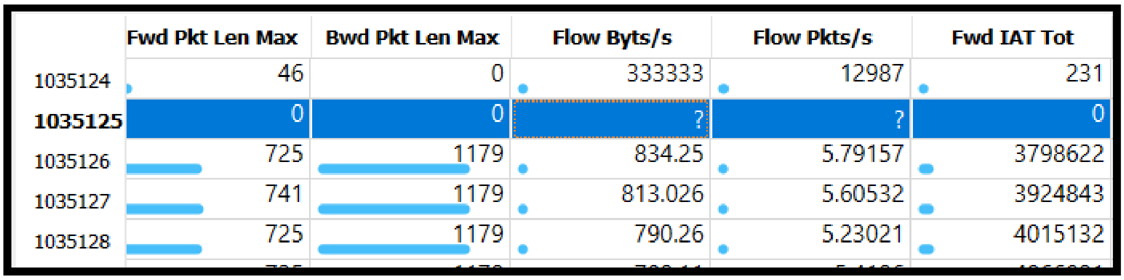
Dataset	# Instances	# Features	Types of Attack
NSL KDD	117682	38	DoS, R2L, U2R, Probe
CIC-DDoS2019	225691	42	DDoS
CSE-CIC-IDS2018	1035193	42	Brute force, Portscan, Botnet, Dos, DDoS, Web, Infiltration
UNSW-NB15	690216	31	DoS, Analysis, Backdoors, Exploits, Generic, Reconnaissance, Shellcode, Worms
KYOTO	304680	10	Normal and Attack sessions

Data Cleaning

The quality and reliability of the collected data are crucial for meaningful analysis. Therefore, this second phase of the study focused on data cleaning to ensure the datasets are free from inconsistencies and missing values. To address missing values, instances with missing values were removed to maintain data integrity. This process ensures that each dataset is ready for subsequent analysis without the interference of incomplete or missing data points. Figure 1 depicts the column with question marks representing missing values.

Figure 1

Sample of Missing Values in Dataset

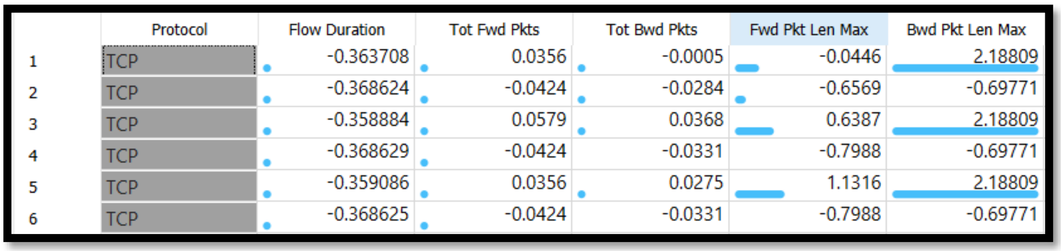


	Fwd Pkt Len Max	Bwd Pkt Len Max	Flow Byts/s	Flow Pkts/s	Fwd IAT Tot
1035124	46	0	333333	12987	231
1035125	0	0	?	?	0
1035126	725	1179	834.25	5.79157	3798622
1035127	741	1179	813.026	5.60532	3924843
1035128	725	1179	790.26	5.23021	4015132

Data normalisation was deployed on continuous attributes to address the inconsistencies and bring the data to a consistent scale. In this study, data normalisation is realised by deducting the value under analysis from the average value and dividing it by the standard deviation; then, the values are substituted with standardised values. Attribute zero-based are not affected by this standardisation. This normalisation is applied consistently across all five datasets, ensuring that the models built on these datasets are not biased by varying feature scales. Figure 2 contains examples of data after normalisation. Combining these data preprocessing techniques ensures the datasets are cleaned, standardised, and prepared for model implementation.

Figure 2

Sample of Normalized Data



	Protocol	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	Fwd Pkt Len Max	Bwd Pkt Len Max
1	TCP	-0.363708	0.0356	-0.0005	-0.0446	2.18809
2	TCP	-0.368624	-0.0424	-0.0284	-0.6569	-0.69771
3	TCP	-0.358884	0.0579	0.0368	0.6387	2.18809
4	TCP	-0.368629	-0.0424	-0.0331	-0.7988	-0.69771
5	TCP	-0.359086	0.0356	0.0275	1.1316	2.18809
6	TCP	-0.368625	-0.0424	-0.0331	-0.7988	-0.69771

Model Implementation

This study focuses on implementing six data mining models that include Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, and Recurrent Neural Network (RNN). For the Decision Tree model, a binary tree was induced with a minimum of two instances in leaves. Subsets smaller than five were not split, and the maximal tree depth was limited to 100. Classification stopped when the majority reached 95%. The Random Forest model consists of ten trees, and subsets smaller than five were not split during the construction of each tree. On the other hand, Logistic Regression employed Ridge (L2) regularisation with a strength parameter (C) set to 1. At the same time, the SVM model was configured with a cost (C) of 1.00, a regression loss epsilon (ε) of 0.10, and utilising RBF kernel. Optimisation parameters included a numerical tolerance of 0.001 and an iteration limit of 100.

The RNN was built using the "keras_model_sequential" function, featuring a single Long Short-Term Memory (LSTM) layer followed by a dense layer with "Sigmoid" and "Relu" activations. This architecture was chosen to capture sequential patterns inherent in intrusion detection datasets. On the other hand, the Naïve Bayes, being a probabilistic model, requires no specific configuration settings.

Model Evaluation

A consistent approach was adopted for all models for testing and scoring results. The training involved random sampling with a repeat train/test of 20 times. All datasets were divided into two portions: 70% for training and 30% for testing. This standardised methodology aimed to facilitate a fair and comparable assessment of the data mining models across diverse datasets. The performance of each model was assessed using evaluation metrics that include Area Under Curve (AUC), Classification Accuracy (CA), F1 Measure (F1), Precision (Prec), Recall, and Matthews Correlation Coefficient (MCC). The larger the value of these metrics, the better the model is. Equations 1 to 6 denote how to calculate the mentioned metrics.

$$AUC = \int_a^b f(x).dx \tag{1}$$

$$\text{CA} = \frac{True\ Positive\ (TP) + True\ Negative(TN)}{True\ Positive\ (TP) + True\ Negative(TN) + False\ Positive(FP) + False\ Negative\ (FN)} \tag{2}$$

$$\text{Prec} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive(FP)} \tag{3}$$

$$\text{Recall} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative(FN)} \tag{4}$$

$$\text{F1} = 2\ x\ \frac{Prec\ x\ Recall}{Prec + Recall} \tag{5}$$

$$\mathrm{CC} = \frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP+FP)(TP+\ FN)(TN+FP)(TN+FN)}} \tag{6}$$

RESULTS

Discussion on the results is presented in two phases: 1) a comparison between traditional data mining models (i.e. five models) and 2) a comparison between traditional models and deep learning models (i.e. two types of RNN). The outcome of Phase 1 is presented in Table 2, which depicts evaluation metrics for training and testing datasets. Column 'Tr' denoted the outcome from the training dataset, while the ones from the testing dataset are depicted under the 'Ts' column.

Table 2

Result of Phase 1

Model	Dataset	Evaluation Metrics											
		AUC		CA		F1		Prec		Recall		MCC	
		Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts
Decision Tree	NSL KDD	0.980	0.982	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.966	0.969
	CIC-DDoS2019	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.999	0.999
	CSE-CIC-IDS2018	0.999	0.999	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.994	0.993
	UNSW-NB15	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	KYOTO	0.927	0.932	0.879	0.880	0.874	0.875	0.883	0.884	0.879	0.880	0.715	0.717
Random Forest	NSL KDD	**1.000**	**1.000**	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.994	0.996
	CIC-DDoS2019	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	CSE-CIC-IDS2018	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	UNSW-NB15	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	KYOTO	0.938	0.938	0.883	0.883	0.878	0.877	0.888	0.888	0.883	0.883	0.725	0.724

SVM	NSL KDD	0.964	0.998	0.971	0.990	0.971	0.990	0.971	0.990	0.971	0.990	0.872	0.956
	CIC-DDoS2019	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	CSE-CIC-IDS2018	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	UNSW-NB15	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	KYOTO	0.605	0.732	0.648	0.666	0.625	0.613	0.616	0.614	0.648	0.666	0.109	0.092
Logistic Regression	NSL KDD	**1.000**	**1.000**	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.987	0.988
	CIC-DDoS2019	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	CSE-CIC-IDS2018	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	UNSW-NB15	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	KYOTO	0.919	0.919	0.870	0.868	0.862	0.861	0.876	0.874	0.870	0.868	0.692	0.688
Naïve Bayes	NSL KDD	0.999	0.999	0.988	0.988	0.988	0.988	0.989	0.989	0.988	0.988	0.949	0.949
	CIC-DDoS2019	**1.000**	**1.000**	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.988	0.988
	CSE-CIC-IDS2018	**1.000**	**1.000**	0.991	0.991	0.991	0.991	0.991	0.991	0.991	0.991	0.971	0.971
	UNSW-NB15	**1.000**	**1.000**	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.997	0.997
	KYOTO	0.859	0.858	0.856	0.852	0.847	0.843	0.862	0.858	0.856	0.852	0.657	0.649

In the evaluation of the Kyoto Dataset, the Decision Tree model displayed remarkable consistency between its performance on the training and testing sets, with AUC values of 0.927 and 0.932, respectively. This marginal difference underscores the model's reliability in discerning different classes, affirming its effectiveness in intrusion detection scenarios. Similarly, the Random Forest model exhibited stability, slightly outperforming the Decision Tree on the testing set with an AUC of 0.938 and maintaining parity with the training set at 0.938, showcasing its robust generalisation capabilities. Conversely, the Support Vector Machine (SVM) model demonstrated a notable performance gap, with an AUC of 0.605 on the training set and 0.732 on the testing set, indicating potential overfitting and limiting its suitability for deployment on the Kyoto Dataset. In contrast, Logistic Regression consistently excelled with an AUC of 0.919 for both sets, emphasising its reliability in intrusion detection. The Naïve

Bayes model also proved reliable, achieving AUC values of 0.859 for training and 0.858 for testing, highlighting its ability to apply learned patterns to new instances effectively. In summary, while Decision Tree, Random Forest, Logistic Regression, and Naïve Bayes demonstrated consistent performance, the SVM model exhibited a performance gap, urging caution in its deployment for intrusion detection on the Kyoto Dataset.

Furthermore, in the evaluation of the DDoS Evaluation Dataset (CIC-DDoS2019), all machine learning models demonstrated outstanding and consistent performance, achieving perfect AUC scores of 1.000 on both the training and testing sets. This remarkable consistency underscores the robustness of the models in effectively detecting intrusions associated with the unique characteristics of DDoS attacks. The Decision Tree model exhibited flawless performance, highlighting its adaptability and reliability in accurately classifying instances. Similarly, the Random Forest model showcased impeccable performance, leveraging the collective strength of multiple decision trees for robust intrusion detection. The Support Vector Machine (SVM) model demonstrated uniform excellence in creating effective decision boundaries between normal and malicious network traffic. Logistic Regression maintained perfect AUC scores, emphasising its reliability in intrusion detection, and even with a marginal decline, the Naïve Bayes model still exhibited outstanding performance. The overall training and testing sets reinforce the suitability of these models for robust security measures in scenarios represented by the DDoS Evaluation Dataset, providing confidence in their accuracy and effectiveness in intrusion detection tasks.

In the assessment of the IDS 2018 Intrusion CSVs (CSE-CIC-IDS2018) dataset, machine learning models demonstrated exceptional performance on both the training and testing sets, consistently achieving high AUC scores, affirming their efficacy in intrusion detection tasks specific to the dataset's characteristics. The Decision Tree model exhibited remarkable AUC scores of 0.999 on both subsets, indicating its robust ability to distinguish between normal and intrusive network activities, making it a reliable choice for intrusion detection within the IDS 2018 dataset. The Random Forest model showcased impeccable performance with a perfect AUC of 1.000 across both sets, emphasising its collective strength in providing accurate intrusion detection. The Support Vector Machine (SVM) model maintained perfect AUC scores, highlighting its effectiveness in creating decision boundaries for intrusion detection within the dataset. Logistic Regression demonstrated flawless performance with perfect AUC scores on both sets, emphasising its accuracy and reliability in generalising to new instances. Despite a slight decline, the Naïve Bayes model maintained commendable performance on both sets, reaffirming its robustness in identifying potential intrusions within the IDS 2018 dataset.

In the evaluation of the UNSW_NB15 Dataset, the machine learning models showcased exceptional and consistent performance on both the training and testing sets, achieving perfect scores of 1.000 across all evaluation metrics, including AUC, CA, F1, Precision, Recall, and MCC. Commencing with the Decision Tree model, its flawless performance on both sets underscores the robustness and reliability of the model in accurately detecting and classifying intrusions within the UNSW_NB15 Dataset. The Random Forest model mirrored this excellence, emphasising the collective strength of the ensemble in providing accurate intrusion detection, with perfect scores of 1.000. Similarly, the Support Vector Machine (SVM) model demonstrated impeccable performance, achieving perfect scores on both sets and reinforcing its suitability for intrusion detection tasks. Logistic Regression maintained flawless performance, with perfect scores of 1.000 on both sets, emphasising its effectiveness in generalising to new instances. Even with a slight decline in performance, the Naïve Bayes model exhibited commendable results, underlining its robustness in identifying potential intrusions within the UNSW_NB15 Dataset. The detailed comparison underscores the consistent and high-level performance of these models in accurately detecting intrusions, with perfect scores of 1.000, reinforcing their suitability for effective security measures in scenarios represented by the UNSW_NB15 Dataset.

In the evaluation of the NSL Dataset, diverse machine learning models showcased varied performance characteristics in the context of intrusion detection. The Decision Tree model exhibited robust capabilities with consistent AUC scores of 0.980 for the training set and 0.982 for the testing set, along with high CA, F1, Precision, Recall, and MCC scores for the testing set, indicating its ability to strike a balance between precision and recall. The Random Forest model demonstrated exceptional and near-perfect performance across all metrics for both sets, highlighting the strength of ensemble learning in improving generalisation and achieving robust intrusion detection. The Support Vector Machine (SVM) model, while showing slightly lower scores on the testing set, maintained a high AUC, suggesting its capability to distinguish between normal and intrusive instances in the NSL Dataset. Logistic Regression consistently displayed strong performance on both sets, solidifying its reliability for intrusion detection. In contrast, the Naïve Bayes model, while achieving a high AUC, showed a decline in other metrics, indicating potential challenges in balancing precision and recall.

Overall, the evaluation of diverse machine learning models across different intrusion detection datasets has provided valuable insights into their performance characteristics. Notably, the Decision Tree, Random Forest, Logistic Regression, and Naïve Bayes models demonstrated consistent performance in accurately detecting intrusions, showcasing their reliability in various scenarios. These models exhibited robustness, achieving high AUC scores, underscoring their effectiveness in distinguishing between normal and malicious network activities. However, it is essential to exercise caution in the deployment of the Support Vector Machine (SVM) model, as indicated by its performance gap in the Kyoto Dataset.

As deep learning models are showing promising outcomes in data analytics, this study investigates the deployment of Recurrent Neural Networks (RNN) in intrusion detection. Phase 2 of the evaluation details the performance of RNN models as compared to the traditional models. Table 3 depicts the classification accuracy for all compared models. The “Tr” denotes the outcome of the training set, while the testing results are represented under the “Ts” column.

Table 3

Result of Phase 2

Dataset	Models													
	Decision Tree		Random Forest		SVM		Logistic Regression		Naïve Bayes		RNN (Sigmoid)		RNN (ReLu)	
	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts
NSL KDD	0.993	0.993	0.999	0.999	0.971	0.990	0.997	0.997	0.988	0.988	**1.000**	**1.000**	**1.000**	**1.000**
CIC-DDoS2019	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	**1.000**	**1.000**	**1.000**	**1.000**
CSE-CIC-IDS2018	0.998	0.998	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.991	0.991	**1.000**	**1.000**	**1.000**	**1.000**
UNSW-NB15	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**
KYOTO	0.879	0.880	0.883	0.883	0.648	0.666	0.870	0.868	0.856	0.852	**1.000**	**1.000**	**1.000**	**1.000**

The result of a 70% data training set of Recurrent Neural Networks (RNNs) across diverse datasets has yielded noteworthy results. In the NSL Dataset, the Decision Tree achieved an accuracy of 0.993, Random Forest excelled with an accuracy of 0.999, and SVM, Logistic Regression, and Naïve Bayes demonstrated competitive performances with accuracies of 0.971, 0.997, and 0.988, respectively. Notably, RNNs, utilising both Sigmoid and ReLu activation functions, outperformed traditional models with a perfect accuracy 1.000. Moving to the DDoS Evaluation Dataset (CIC-DDoS2019), all traditional models achieved perfect accuracies, emphasising their robustness. Similarly, RNNs with both Sigmoid and ReLu activations maintained the accuracies.

In the IDS 2018 Intrusion CSVs (CSE-CIC-IDS2018) Dataset, traditional models, including Decision Tree, Random Forest, SVM, and Logistic Regression, exhibited strong performances with perfect accuracies. Naïve Bayes showed resilience with an accuracy of 0.991. RNNs, employing both Sigmoid and ReLu, again demonstrated impeccable accuracy, highlighting their suitability for intrusion detection tasks. The UNSW_NB15 Dataset showcased the adaptability of traditional models, with Decision Tree, Random Forest, SVM, and Logistic Regression achieving perfect accuracies. Naïve Bayes maintained high accuracy at 0.999. RNNs with both Sigmoid and ReLu activations consistently demonstrated flawless accuracy, providing a compelling alternative for intrusion detection in this dataset.

In the Kyoto Dataset, where traditional models faced challenges, RNNs emerged as particularly effective. Decision Tree and Random Forest achieved accuracies of 0.879 and 0.883, respectively, while SVM lagged with an accuracy of 0.648. Logistic Regression and Naïve Bayes demonstrated respectable performances with accuracies of 0.870 and 0.856. Strikingly, RNNs with both Sigmoid and ReLu activations showcased perfect accuracies, underscoring their potential for handling the nuances of network security in this specific environment.The training set of Recurrent Neural Networks, especially with Sigmoid and ReLu activations, has demonstrated promising results in intrusion detection across diverse datasets. These findings contribute valuable insights into the potential of neural network approaches for enhancing the accuracy of intrusion detection systems.

The evaluation of model performance on the testing set reveals significant insights into the effectiveness of intrusion detection across diverse datasets. In the NSL dataset, traditional models such as Decision Tree, Random Forest, SVM, and Logistic Regression displayed robust accuracy levels, with SVM exhibiting improved performance on the testing set compared to the training set at 0.990. The RNNs, employing both Sigmoid and ReLu activations, continued to showcase impeccable accuracy, emphasising their consistency in accurately identifying intrusive network activities. In fact, the RNNs performed very well in all datasets.

The CIC-DDoS2019 reinforced the reliability of traditional models, with Decision Tree, Random Forest, SVM, and Logistic Regression achieving perfect accuracies. Naïve Bayes maintained a high accuracy of 0.997 on the testing set. In the CSE-CIC-IDS2018 dataset, all models except the Decision Tree and Naïve Bayes obtained 100 per cent accuracy. A similar pattern can be seen in the UNSW_NB15 dataset for all models except for Naïve Bayes (with 0.999).

The Kyoto dataset, where traditional models faced challenges, witnessed a similar trend in the training set. Decision Tree and Random Forest achieved accuracies of 0.880 and 0.883, respectively, while SVM exhibited a testing set accuracy of 0.666. Logistic Regression and Naïve Bayes demonstrated respectable performances, with accuracies of 0.868 and 0.852. However, the traditional models fail to outperform the RNNs as both Relu and Sigmoid RNN generated perfect accuracy.

The evaluation of model performance on the testing set reaffirms the consistency and reliability of traditional models. It underscores the robustness of RNNs, significantly when activated with Sigmoid and ReLu functions, in accurately identifying and classifying intrusions across diverse datasets. Activated with either Sigmoid or ReLu functions, RNNs consistently outshine other classification models, demonstrating perfect accuracy not only in the training but also in the testing set. Such a positive outcome underscores the unparalleled ability of RNN to understand and distinguish between normal and intrusive network activities. While other models exhibit commendable accuracy, RNN's unwavering perfection across datasets is essential for intrusion detection tasks.

CONCLUSION

The main concepts involve employing data mining models to identify recurring and valuable patterns in intrusion detection datasets and utilising pertinent attributes to create (inductively taught) classifiers capable of identifying intrusion activity. The comparative analysis highlighted the exceptional performance of Recurrent Neural Networks (RNNs) that are equipped with either Sigmoid or ReLu activation functions. These models consistently achieved remarkable performance across all datasets, showcasing their unparalleled ability to discern normal and intrusive network activities. While traditional models like Decision Tree, Random Forest, SVM, Logistic Regression, and Naïve Bayes demonstrated commendable performance, the consistent perfection of RNNs underscored their supremacy in intrusion detection.

In addressing the dataset specificity, it is essential to acknowledge that the study's findings might be confined to the characteristics of the selected intrusion detection datasets. While these datasets offer valuable insights, their unique features may limit the broader applicability of the models. To mitigate this limitation, future research should explore the performance of intrusion detection models across a more diverse range of real-world datasets, ensuring a comprehensive understanding of model generalisation in varied network environments. Additionally, the study utilised default hyperparameters for model training, a decision that might not optimise model performance to its full potential. Recognising the sensitivity of intrusion detection models to hyperparameter configurations, future research could thoroughly explore hyperparameter tuning. Such an approach may uncover optimal configurations for improved model robustness and accuracy. As cybersecurity threats evolve, the study suggests exploring ensemble approaches as a promising future direction for data mining in intrusion detection. Combining the strengths of multiple intrusion detection models through ensemble techniques may enhance overall performance and resilience.

To sum up, this research adds essential knowledge to the intrusion detection domain as it emphasises the prominence of RNNs and outlines key considerations for future research. The deployment of robust and adaptive intrusion detection systems is crucial in safeguarding network environments, and this research lays a foundation for continued advancements in the realm of cybersecurity.

REFERENCES

Ahsan, R., Shi, W., & Corriveau, J. P. (2022). Network intrusion detection using machine learning approaches: Addressing data imbalance. *IET Cyber-Physical Systems: Theory and Applications, 7*(1). https://doi.org/10.1049/cps2.12013

Chou, D., & Jiang, M. (2022). A Survey on Data-driven Network Intrusion Detection. *ACM Computing Surveys*, *54*(9). https://doi.org/10.1145/3472753

Cihan, Ş., Aydos, M., & Şimşek, N. Y. (2021). A Tree-Based Machine Learning and Deep Learning Classification for Network Intrusion Detection. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.889994

Duraz, R., Espes, D., Francq, J., & Vaton, S. (2023). Cyber Informedness: A New Metric Using CVSS to Increase Trust in Intrusion Detection Systems. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3590777.3590786

Fan, Z. (2021). The evaluation of bank credit is based on the improved decision tree model. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3482632.3487479

Guo, W. (2022). Applications of Logistic Regression and Naive Bayes in Commodity Sentiment Analysis. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3531232.3531265

Herrera, J. A., Camargo, J. E., & Torre, J. I. M. (2022). Network anomaly detection with machine learning techniques for SDN networks. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3535735.3535750

Lei, G., Su, S., & Liao, W. (2021). Classification of credit card holders based on random forest algorithm. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3453800.3453806

Liu, B., Huang, Z., & Zhu, Z. (2022). Intrusion Detection System Based on Machine Learning. *ICCSIE '22: Proceedings of the 7th International Conference on Cyber Security and Information Engineering*, pp. 121–125. https://doi.org/10.1145/3558819.3558840

Meftah, S., Rachidi, T., & Assem, N. (2019). Network-based intrusion detection using the UNSW-NB15 dataset. *International Journal of Computing and Digital Systems*, *8*(5). https://doi.org/10.12785/ijcds/080505

Mehta, J., Richard, G., Lugosch, L., Yu, D., & Meyer, B. H. (2023). DT-DS: CAN Intrusion Detection with Decision Tree Ensembles. *ACM Transactions on Cyber-Physical Systems*, *7*(1). https://doi.org/10.1145/3566132

Mohamed, M. A., & Mohamud, M. A. (2022). Data Analysis of Agriculture using Data Mining Techniques. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3549206.3549226

Nagaraja, A., Boregowda, U., & Radhakrishna, V. (2021). Regression analysis for network intrusion detection. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3460620.3460751

Schlag, S., Schmitt, M., & Schulz, C. (2021). Faster Support Vector Machines. *ACM Journal of Experimental Algorithmics*, *26*. https://doi.org/10.1145/3484730

Shah, S., Muhuri, P. S., Yuan, X., Roy, K., & Chatterjee, P. (2021). Implementing a network intrusion detection system using a semi-supervised support vector machine and random forest. *Proceedings of the 2021 ACMSE Conference - ACMSE 2021: The Annual ACM Southeast Conference*. https://doi.org/10.1145/3409334.3452073

Sharafaldin, I., Lashkari, A. H., Hakak, S., & Ghorbani, A. A. (2019). Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. *Proceedings - International Carnahan Conference on Security Technology*, *2019-October*. https://doi.org/10.1109/CCST.2019.8888419

Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance Evaluation of Deep neural networks Applied to Speech Recognition: Rnn, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, *9*(4). https://doi.org/10.2478/jaiscr-2019-0006

Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D., & Nakao, K. (2011). Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. *Proceedings of the 1st Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, BADGERS 2011*. https://doi.org/10.1145/1978672.1978676

Tsantekidis, A., Passalis, N., & Tefas, A. (2022). Recurrent neural networks. In *Deep Learning for Robot Perception and Cognition*. https://doi.org/10.1016/B978-0-32-385787-1.00010-5

Wang, S. (2023). Recognition of English speech - Using a deep learning algorithm. *Journal of Intelligent Systems*, *32*(1). https://doi.org/10.1515/jisys-2022-0236

Yang, L., Cai, M., Duan, Y., & Yang, X. (2019). Intrusion detection based on approximate information entropy for random forest classification. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3335484.3335488

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. *IEEE Access*, *5*. https://doi.org/10.1109/ACCESS.2017.2762418

Yu, W., Kim, I. Y., & Mechefske, C. (2021). Analysis of different RNN autoencoder variants for time series classification and machine prognostics. *Mechanical Systems and Signal Processing*, *149*. https://doi.org/10.1016/j.ymssp.2020.107322

Zhang, M., & Qin, X. (2022). Research on Satisfaction Evaluation Model Based on Ordered Logistic Regression. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3584748.3584794

Zhao, Y., & Gan, G. (2021). Research on Intrusion Detection Technology Based on Ensemble Learning. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3474198.3478139

Zipperle, M., Gottwalt, F., Chang, E., & Dillon, T. (2022). Provenance-based Intrusion Detection Systems: A Survey. *ACM Computing Surveys*, *55*(7). https://doi.org/10.1145/3539605