National College of Ireland

ENHANCING INTRUSION DETECTION SYSTEMS (IDS) USING MACHINE LEARNING TECHNIQUES: A COMPARATIVE STUDY OF DEEP LEARNING AND CLASSICAL MACHINE LEARNING METHODS FOR IMPROVED DETECTION ACCURACY AND SPEED

MSc Research Project
MSc Cybersecurity

Egwu Nelson Chinedu
Student ID: x23258608

School of Computing
National College of Ireland

Supervisor: MICHAEL PANTRIDGE

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Chinedu Nelson Egwu
Student ID:	X23258608
Programme:	MSc Cybersecurity
Year:	2025
Module:	MSc Research Project
Supervisor:	Michael Pantridge
Submission Due Date:	29/01/2025
Project Title:	ENHANCING INTRUSION DETECTION SYSTEMS (IDS) USING MACHINE LEARNING TECHNIQUES: A COMPARATIVE STUDY OF DEEP LEARNING AND CLASSICAL MACHINE LEARNING METHODS FOR IMPROVED DETECTION ACCURACY AND SPEED
Word Count:	4975
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Chinedu Nelson Egwu
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	Q
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	Q
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	Q

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

ABSTRACT

This research proposes a comprehensive analysis on the approach of utilizing machine learning and deep learning systems to detect intrusion in a network based environment. As information technology advances, the use of malicious intent to attack networks also increases as well, therefore putting much infrastructure at the mercy of these hackers. This research examines some machine learning and deep learning algorithms which are decision trees, random forest, logistic regression, artificial neural network (ANN) and convolutional neural network (CNN), and their strengths in the detection of this intrusion in these network based systems. The key findings show that the Random Forest and the decision trees classifier, with their tree base estimator and accuracy optimization, achieved the best results with an accuracy of 98% for the both models.

Keywords: Decision trees, random forest, logistic regression, artificial neural network (ANN) and convolutional neural network (CNN), intrusion, machine learning, deep learning.

TABLE OF CONTENTS

1.0. INTRODUCTION

The Statistics Report of the World Internet states that the internet grew by 1,114% between 2000 and 2019 because more than two quintillion bytes of data are created daily (Basheer & Ranjana, 2022). This implies that the expansion of data from numerous sources is occurring at an incredibly rapid rate, and that hacking tools and techniques are likewise developing at that same rate (Bharati and Podder, 2022). Data analysis and information security are therefore necessary to safeguard the data against intrusion (Kumari *et al.*, 2019). Traditional detection systems are unable to identify intrusions quickly because of the size and speed of data generation (Lim *et al.*, 2020). Due to increased time and resource consumption, the typical data handling system becomes even more complex as data grows exponentially and because data can be so complicated, handling it requires advanced intelligence algorithms and powerful technologies (Kodheli *et al.*, 2020). In order to identify any suspicious behavior or attack from hackers, the intrusion detection system (IDS) will be required to monitor network traffic (Khraisat *et al.*, 2019).

An intrusion detection system (IDS) is used to find loopholes or malicious activities and secure a computer network serving as a network-level defense system (Barua *et al.*, 2022). Network defects, such as buffer overflows and inadequate security standards are targeted by malicious users to violate the security of the network (Mazhar *et al.*, 2023). These malicious users may be hackers, who are regular internet users who want to steal or corrupt sensitive data from the victim's system or they may be users with limited access who want to have more access by force. There are two types of intrusion detection methods namely anomaly detection and signature detection (Khraisat *et al.*, 2019). By keeping an eye on network packet flow, the signature-based detection system compares it to previously established signatures of known threats (Chatterjee and Ahmed, 2022). Many facets of human life, including commerce and industry, rely on computer networks therefore, creating a dependable network is the responsibility for managers in the information technology sector (Musarat *et al.*, 2022). On the other hand, the rapid advancement of information technology has led to a number of challenges in creating reliable networks, which is an extremely challenging task (Chowdhury *et al.*, 2020). The integrity of computer networks are under risk from a variety of threats such as Denial of service (DOS) attacks, which is among the most popular and dangerous types of attack and is designed to briefly stop end users from using a number of services (Tsiknas *et al.*, 2021).

Intrusion detection systems can be either network-based (NIDS) or host-based (HIDS) as well (Ahmad *et al.*, 2020). To keep an eye on and examine activity on a specific computer, network administrators use host-based intrusion detection systems. The fact that encrypted data can be viewed while moving over a network gives HIDS a common edge over NIDS. The network intrusion detection systems (NIDS) are hardware or software-based devices that are strategically placed throughout networks to passively watch network traffic passing through their host equipment. There are usually two interfaces for the network-based detection system, one which is used for network conversation listening, and the other is utilized for reporting and control (Dwivedi *et al.*, 2020). One advantage of network-based detection systems is that they are generally impervious to many attackers, making them secure against attacks (Khaliq *et al.*, 2022). However, monitoring a vast network may necessitate a few well-suited network-based detection systems. However, the limitation of the network-based system is that they have trouble identifying an attack when traffic is heavy (Porambage *et al.*, 2021).

The majority of the early intrusion detection systems were signature-based, meaning that intrusion detection relied on pre-defined and registered known attack signatures (DeMedeiros, Hendawi and Alvarez, 2023) which is a limitation because it requires constant updating of the database of known attack signatures, because hackers frequently figure out ways to take advantage of network activity. The emergence of machine learning enabled intrusion detection, which compares incidents that occurred in an environment with previously observed incidents to identify unknown anomalies (Djenna, Harous and Saidouni, 2021). Several machine learning techniques have been used over time to increase the prediction accuracy of intrusion detection systems (IDS), decrease false positives which can be detrimental especially in life threatening conditions, and improve detection rate. Therefore this study proposes an empirical approach to identify intrusion using both machine learning and deep learning algorithms.

1.1. SIGNIFICANCE OF THE STUDY

By studying and implementing a detection system to identify intrusions in network traffic, this study contributes to developing more accurate and responsive systems, protecting private and organizational information from exposure. It also plays a critical role in identifying and preventing unauthorized access, potential threats, and malicious activities within a network.

1.2. AIM AND OBJECTIVES

The aim of this study is to enhance cybersecurity using machine learning tools to identify and mitigate vulnerabilities such as intrusion in network infrastructure with the following objectives:

- To design and implement robust deep learning and machine learning solutions capable of detecting intrusion threats.
- Comparison of the capabilities of various supervised machine learning techniques in the detection of network vulnerabilities.
- Model optimization to determine the effectiveness and computational efficiency of each model, selecting the most suitable for the detection.

1.3. RESEARCH QUESTIONS

- Can machine learning and deep learning algorithms be implemented to mitigate cybersecurity risks of intrusion?
- Which model is more suited for the identification of network vulnerabilities?
- Can machine learning understand the patterns of network interactions?

1.4. LIMITATIONS

- Availability of data: high quality data are essential for the development of this software tool which may pose challenges because acquiring labeled, real-time and updated data is challenging.
- Data imbalance: cybersecurity datasets often have a significant class imbalance where there are fewer malicious cases. This may be detrimental to the model prediction.
- Resource availability: complex deep learning models, may require more computational resources for training and real time processing and will be a limiting factor for infrastructures who have cost deficiencies.

2.0. LITERATURE REVIEW

Since detecting and preventing intrusion in a network system is critical to improving a safe space in network and data infrastructure, this literature review will provide insight on retrospective methods employed in the detection of intrusion.

2.1. MACHINE LEARNING BASED INTRUSION DETECTION.

Intrusion detection is a very important aspect of cybersecurity which detects unauthorized access within network systems. With the rapid rate at which cyber-attacks are occurring, traditional rule-based and signature-based systems have proven unreliable with the sophisticated attacks. With the emergence of machine learning, the detection of intrusion has been enhanced since it has the ability to analyze large voluminous data, identify patterns in a network and detect anomalies in the network. This section provides comprehensive analysis of the literature on intrusion detection using machine learning, supervised learning to be exact.

To detect intrusion in network systems, Ugochukwu et al. (2019) proposed a predictive approach using machine learning. In this research, the authors utilized the Bayes net, J48, random forest, and random tree models which are machine learning models. The models were developed using the KDDCup99 dataset which was readily available online. The researchers utilized the WEKA open source machine learning scripting software. In this research, there were five unequal classes in the label. After the whole training and testing process was completed, the random forest had the higher precision of 97.7%, the Bayes net had the highest recall and the random tree had the highest F-measure score. Generally, the random forest and random tree were the most efficient models for the classification. This research was limited to a few algorithms, therefore the authors suggested an implementation of a wider range of algorithms for the future recommendations.

Hassan et al. (2020) also performed an experiment using the same predictive approach as Ugochukwu et al. (2019), in that the authors utilized a dataset, the NSL-KDD dataset for network intrusion to distinguish between normal and anomalous activity in networks. The models utilized by the authors were random tree, J48 and Naive Bayes algorithms. After the testing, the random forest and J48 achieved the most accuracy of 99.7 and 99.8 respectively but the false detection rate of the both were 0.2 and 0.3 respectively so the authors concluded that the random tree was the

best performing model. For future recommendations, the authors suggested applying several other machine learning algorithms to reduce the rate of false positives.

2.2. DEEP LEARNING BASED INTRUSION DETECTION.

Due to the architecture of the deep learning models, the neural network has proven to be a very efficient tool in solving complex problems relating to anomaly detection, prediction and further decision making. Because of its feature extraction prowess, it has been employed in several tasks such as image classification and segmentation, audio and video classification and as far as medical imaging. This section delves into the use of deep neural networks such as CNN, ANN and RNN for the detection of intrusion given their state-of-the-art results in such areas.

Ashiku et al. (2021) presented a predictive approach to detect intrusion in network based systems using deep learning architectures. In this research, the authors utilized the UNSW-NB15 dataset which reflects real world modern communication patterns but the attacks were synthesized artificially because getting actual data on intrusion would require some breaches. The proposed model in this paper was the 1D convolutional neural network (CNN) implemented using tensorflow. After the training and testing process, the proposed model achieved an accuracy of 94.4% on the original dataset. It should be noted that the authors stated an issue with class imbalance which caused a model downgrade, therefore undersampling was implemented. The authors went further to partition the UNSW-NB15 dataset with some custom definitions and augmentation which achieved an accuracy of 95.6% after testing. It was stated that transfer learning would play a significant role in the future development of this research to serve as a baseline for the proposed dataset.

Yin et al. (2017) proposed a method similar to Ashiku et al. (2021) to detect intrusion. In the research, the authors performed a comparison between the recurrent neural network (RNN) to the J48, artificial neural network (ANN), random forest, and support vector machine amongst other machine learning algorithms that were reviewed from different literature. The dataset utilized in this paper is the NSL-KDD which consists of both train and test sets with four different attack records. This dataset is so unique in that it has some attack records present in the test set which was purposely removed from the training set so as to imitate the performance of a real world intrusion. The training and testing was performed on different conditions and in both conditions,

the RNN achieved the highest performance with an accuracy of 83.28 and 81.29 on the KDDTest and 68.55 and 64.67 on the KDDTest_21. The authors recommended the usage of GPU to reduce training time and increase performance.

2.3. UNSUPERVISED LEARNING FOR NETWORK BASED INTRUSION.

Unsupervised learning is a branch of machine learning which does not make use of labeled data. Its objective is to understand the hidden patterns and structures within a data distribution, which makes it a very useful tool in the detection of unknown intrusions in a given system. The way these systems work is by grouping similar data points into a cluster thereby identifying outliers which may be classified as an anomaly or intrusive attack. This section focuses on the application of such unsupervised learning techniques such as Kmeans, principal component analysis (PCA) and Gaussian mixture methods.

Verkerken et al. (2020) proposed a method that automatically identifies intrusion in networks using four unsupervised learning techniques, two of which are self-supervised. In this research, the dataset utilized was the CIC-IDS-2017 which contains different classes of attack type. Being an unsupervised learning system, this dataset first underwent a series of preprocessing stages before it was fed into the machine learning algorithms, steps such as feature selection which removes redundant features, parsing of the data into the correct format and incomplete rows were dropped and the data was scaled before transformation using PCA as an unsupervised feature reduction method. The four machine learning models employed in this study were principal component analysis (PCA), isolation forest, one class support vector machine and auto-encoders. After the training and testing was concluded, the auto-encoder which is a deep learning algorithm and also an unsupervised learning technique was seen to be the best performing model with an AUROC score of 97.75. As this unsupervised learning process proved resilient, the authors suggested testing in a real world environment as a future recommendation.

Samriya et al. (2020) proposed a novel hybridization approach to detect intrusion in cloud based computing networks which uses a fuzzy artificial neural network (ANN). The authors proposed that the hybridization approach would overcome the iterative classification and the selection of fuzzy clusters by updating the fit value automatically, being unsupervised. Just as the last step in Verkerken et al's paper was dimensionality reduction, this research proposed a spider-monkey

optimization (SMO) technique which reduces the dimension of the dataset before it's passed into the neural network. The dataset utilized in this research was the NSL-KDD data. The authors presented 4 models, the proposed hybrid FCM-SMO model, the fuzzy c-means + ANN, fuzzy c-means + SVM and lastly the ANN. After the experiment, the authors found the hybrid model to be the most efficient model for the detection of anomalies.

Chen et al. (2020) presented a similar novel hybrid approach but this time, the researchers used k-means as the baseline model upon which the hybrid model which they referred to as QALO-K would stem from. In this model, the researchers combined k-means with a quantum-inspired ant lion optimization, reinforcing the k-means with quantum computing and swarm intelligence. The dataset utilized in this experiment was the KDDCup 99 large dataset for intrusion detection as well as other datasets from the UCI machine learning repository. In the result comparison, the proposed QALO-K was compared against the baseline k-means and ALO-K algorithms and the proposed model performed better than the compared models.

3.0. RESEARCH METHODOLOGY

This section contains the steps taken in the development of the proposed intrusion detection models from the data input. It encompasses techniques from data collection, model selection, training and evaluation.

3.1. DESIGN AND IMPLEMENTATION SPECIFICATIONS

The system design for this machine learning approach for an intrusion detection system from data collection stage to inference and evaluation of test results is shown in the flowchart below:

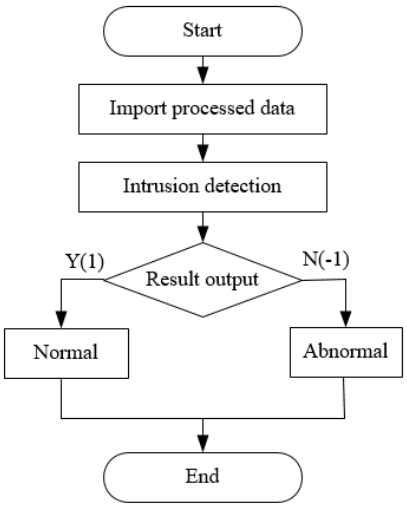

Figure 1: Flowchart of the design step

3.1.1. THE DATASET

The dataset utilized in this research is the WSN-DS dataset obtained from kaggle. It is a dataset used for intrusion detection systems in wireless sensor networks. The WSN stands for Wireless Sensor Networks and they have become increasingly one of the hottest research areas in computer science due to their wide range of applications. To ensure the security and effectiveness of WSN services, an Intrusion Detection System (IDS) should be in operation. In this research, the WSN dataset was utilized to help better detect and classify four types of Denial of Service (DoS) attacks:

Black Hole, Grayhole, Flooding, and Scheduling attacks. The dataset consists of 19 columns and 374661 rows.

3.2. MATERIALS AND EQUIPMENT

In this study, various materials and equipment were utilized. These materials are divided into the hardware, software and libraries. The components of each class are listed below:

- Hardware: Intel core i5 processor and a 16gb ram device.
- Software and Libraries: Python (version 3.11), Pandas and numpy for data handling and manipulation, matplotlib and seaborn for visualization, sci-kit learn, tensorflow and keras for data processing, machine learning and deep learning model development.

3.3. DATA PREPROCESSING TECHNIQUES

In this dissertation, it was mandatory that the dataset utilized was properly cleaned and preprocessed to ensure a smooth flow of operation from the cleaning stage to the model training stage which involved the creation of both the machine and deep learning models. The major steps involved are shown below:

- Filling in missing data: There were no missing values in the dataset utilized in this research. There are methods used to fill in missing data information in a scenario where some values are missing such as filling in with the mean, mode or median of the feature distribution. But where the percentage of missing values are high and wrong inputs would affect the prediction, the missing rows will be dropped off.
- Scaling the data: When dealing with datasets relating to intrusion or anomaly detection, its custom to use the Standard Scaler or MinMax Scaler, but in a machine learning application there is room for trial and error to see which best suits the problem. In this research, using the Standard Scaler or MinMax Scaler resulted in overfitting therefore the Normalizer was used instead. The Normalizer is a technique used to scale data independently by its L2 norm, making it suitable for problems where the feature magnitudes are important. The Standard Scaler adjusts the mean and standard deviation of the given features while the

Normalizer just transforms individual samples so the norm becomes unity, making the vector length equal to 1.

- Label Encoding: This is a preprocessing technique used to transform non-numerical features into numerical features. This is a necessary component of machine learning data preprocessing because machine learning algorithms require input data in numerical format and not in object or string format.
- Selection of important features: The redundant features in the dataset were dropped as they contributed little to no meaningful insight to the overall performance of the machine learning training and testing.
- Balancing of the dataset: In this research, the level of class imbalance was so high. The normal scenario had a 90.8% appearance while the other labels fit into the 9.2%, which would result in a skewed view of the machine learning model on the data. The sampling technique employed in this research was the SMOTE over sampler which handles cases of data imbalance where the majority class is over represented compared to the minority class. It tackled this problem by duplicating the samples from the minority class until they matched the majority class. This enabled the models to learn from a more represented distribution of classes.
- Splitting of the data: The train, test and validation sets were split in the ratio 60:20:20.
- CPU and Time tracking: The training time and CPU consumption for each model was calculated using the measurement for threading. The efficiency of each model was measured as it's essential for machine learning models to produce efficient results and at the same time consume less computational hardware and less time. The psutil library aided the detailed system and process-level information.

3.4. MACHINE LEARNING MODELING

In this section, the machine learning and deep learning models employed in this intrusion detection system would be comprehensively discussed, staying how and why each model was developed. The models utilized were Decision trees classifier, Logistic regression, Random forest classifier, Artificial Neural Network (ANN) and Convolutional Neural Network (CNN).

3.4.1. DECISION TREES CLASSIFIER

A decision tree is a supervisor learning model used for both classification and regression (Varol, Omurlu and Türe, 2024). Pictorially and functionally, it resembles a branching tree where each node in the tree represents a test, and each branch shows a result that leads to a lead node which stands for the class label. It works by recursions where the dataset is split into subsets based on features that provide the most information gain and minimal impurity (Paul and Das, 2023). This model was chosen because of its interpretative nature and can capture non - linear relationships like intrusion patterns.

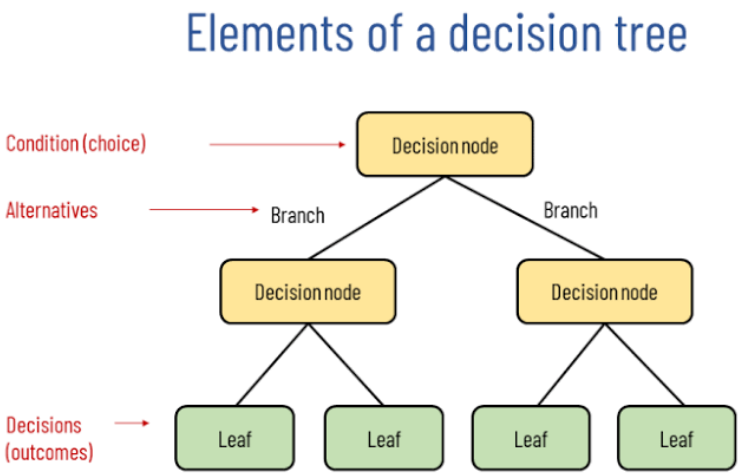

Figure 2: understanding the decision tree (Tales, 2023)

3.4.2. LOGISTIC REGRESSION

This is also a supervised learning model used mainly for binary classification tasks because of its sigmoid output nature, although it can be extended to fit multiclass classifications (Farahani *et al.*, 2024). A sigmoid model is one which has a categorical output, where one of only two outcomes is possible i.e. 1 or 0 (Nwafor, Nwafor and Brahma, 2024). The model was chosen because of its simple and probabilistic output which is very interpretable and provides confidence scores.

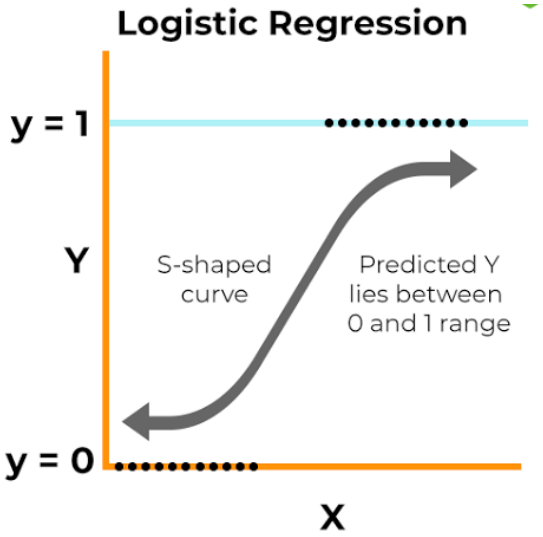

Figure 3: logistic regression (Kanade, 2024)

3.4.3. RANDOM FOREST CLASSIFIER

This is a very powerful ensemble model used for both classification and regression tasks. The building blocks of the random forest is a combination of several decision trees in which their results from training are aggregated to form a final probabilistic output prediction (Shabbir *et al.*, 2024). This model was developed based on the bagging concept and is known for its durability and ability to handle multidimensional data structures (Hussein *et al.*, 2024). Due to its combined nature of decision trees, it tends to produce a higher accuracy than the single decision trees algorithm (Sapkota *et al.*, 2024). This model was chosen for this purpose because it helps reduce overfitting, handle large datasets properly, robust to outliers and imbalances such as intrusion in a dataset.

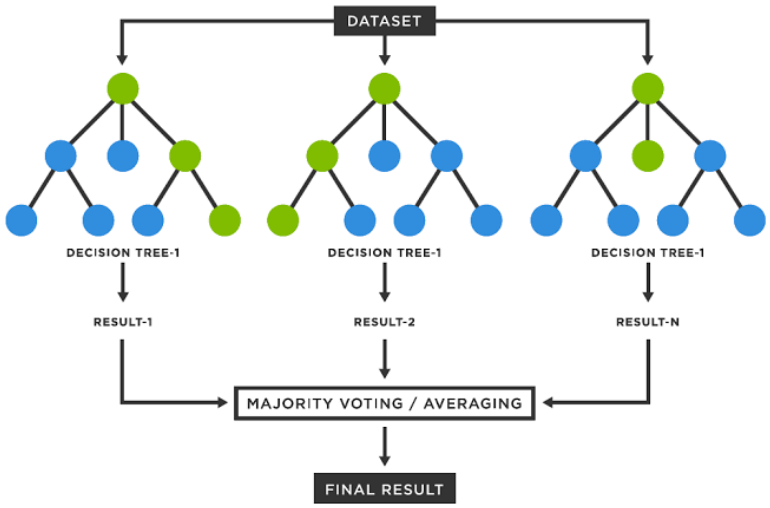

Figure 4: random forest (Gunay, 2023)

3.4.4. ARTIFICIAL NEURAL NETWORK (ANN)

This is a model developed from the inspiration of the sensory and motor neurons of the human brain (Krauhausen *et al.*, 2024). This is a core deep learning model which can understand extremely complex relationships between input and output of a dataset (Siam *et al.*, 2024). It is a multipurpose model used for classification, regression and clustering analysis. Since it works well for high dimensional data, it was chosen for this reason to detect intrusion in network systems.

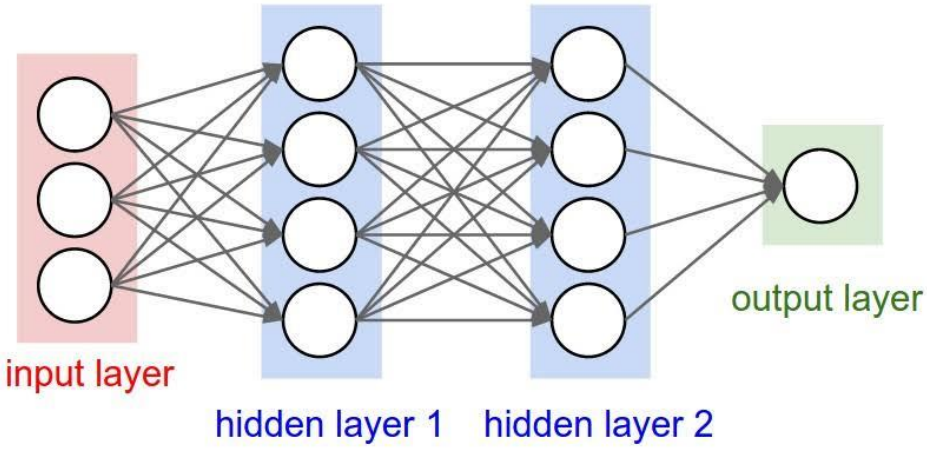

Figure 5: artificial neural network (Dormehl, 2022)

3.4.5. CONVOLUTIONAL NEURAL NETWORK (CNN)

A Convolutional neural network (CNN) is a deep learning model developed after the artificial neural network or multi-layer perceptron to process structured image and video data in grid format (Madani *et al.*, 2024). The main difference between the CNN and the ANN is the presence of the convolution layer which plays a very important role in image processing (Zou *et al.*, 2024). This model is used in computer vision tasks but can be applied to tabular data such as intrusion detection because this model being more powerful than the artificial neural network, automatically learns hierarchical patterns in data, reduces the number of parameters unlike the ANN and detects features regardless of their position in a data such as hidden intrusion in a data structure.

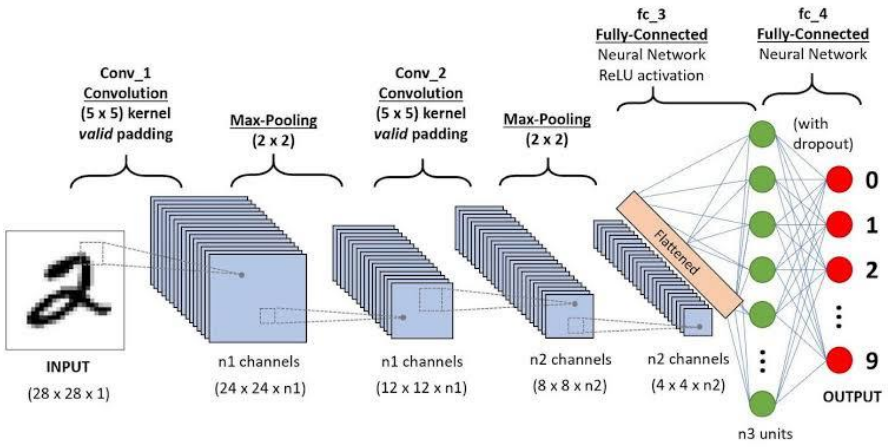

Figure 6: Convolutional neural network (Dandekar, 2023)

4.0. EVALUATION

This study proposes five (5) machine learning and deep learning models for the detection of intrusion in a network system. The results obtained from the testing of the models are shown below:

MODEL	ACCURACY	PRECISION (0)	PRECISION (1)	PRECISION (2)	PRECISION (3)	PRECISION (4)
Decision Trees	98%	67%	92%	92%	100%	77%
Logistic Regression	79%	14%	9%	0%	0%	10%
Random Forest	98%	68%	91%	90%	100%	95%
ANN	92%	65%	54%	28%	100%	89%
CNN	95%	66%	38%	54%	100%	69%

Table 4.1: Accuracy and Precision of the models

MODEL	RECALL (0)	RECALL (1)	RECALL (2)	RECALL (3)	RECALL (4)
Decision Trees	95%	95%	66%	99%	92%
Logistic Regression	30%	16%	0%	85%	82%
Random Forest	94%	99%	68%	100%	92%
ANN	93%	100%	56%	93%	92%
CNN	95%	94%	64%	97%	50%

Table 4.2: Recall of the models

MODEL	F1-SCORE (0)	F1-SCORE (1)	F1-SCORE (2)	F1-SCORE (3)	F1-SCORE (4)
Decision Trees	79%	93%	77%	100%	84%
Logistic Regression	19%	12%	0%	92%	17%
Random Forest	79%	95%	77%	100%	93%
ANN	77%	70%	37%	97%	91%
CNN	78%	54%	59%	98%	58%

Table 4.3: F1-score of the models

Table 4.1 shows the results obtained from the comparative analysis of the proposed machine learning and deep learning models in the research, highlighting the accuracy and precision of the models across the five (5) classes of intrusion labels. From the table, according to their accuracies, decision trees classifier and random forest are the best performing models with an accuracy of 98% and their precisions across all the classes exhibiting a fantastic performance in the classification of the intrusion. The convolutional neural network and the artificial neural network achieved similarly high accuracies but slightly lower than the best two. In contrast, the Logistic regression was the worst performing model in terms of accuracy and also precision. This result was expected because the logistic regression excels better in the field of binary classification because of the nature of its architecture which is a sigmoid output.
In this research, the random forest and the decision trees are the most efficient models to be used for the detection of intrusion in network based systems.

Table 4.2 also shows the recall values for the proposed models across the classes in the dataset, with the random forest and decision trees having the best values which confirms the claim of them being the best models from table 4.1. The convolutional neural network also performs almost as

the decision trees in terms of class performance. The logistic regression also performs poorly when measured using this metric of performance.

Table 4.3 shows the F1 score of the models with random forest and decision trees being up above the others. Following the best models is the convolutional neural network and lastly, the logistic regression with the worst performance.

4.1. DISCUSSIONS

In this research, machine learning and deep learning models were proposed for the detection of intrusion in a network based system. This section examines the results obtained from the research and compares them with existing papers in the reviews section in chapter 2.

From the previous research, it can be noted that one of the limitation encountered was the lack of diverse machine learning models which could give a broad view on how the models performed with the dataset so in this research, that limitation was tackled by the proposition of five (5) models, both machine and deep learning models. Also, Ugochukwu et al. (2019) stated in their paper that the random forest had the highest precision of 97.7% and from this research the random forest achieved a higher accuracy with a score of 98% which shows an improvement in literature. Hassan et al. (2020) also stated in his paper that the random forest performed well with a 99% accuracy, so therefore it can be seen that they outperformed our best model but with a point difference, giving room for overfitting. Our research also took into consideration the research from Ashiku et al. (2021) utilizing the 1D CNN which had an accuracy of 94.4%, our research outperformed theirs by improving on the accuracy of the convolutional neural network (CNN) to 95%. Lastly, comparing our research to that of Yin et al. (2017), their best performing model had an accuracy of 83.28%, whereas our proposed deep learning models both outclassed this performance. From these results and comparison, we can conclude by saying we have improved and presented an optimized system for the detection of intrusion in network systems using both machine learning and deep learning approaches.

5.0. CONCLUSION AND FUTURE RECOMMENDATIONS

In this dissertation, five machine learning and deep learning models were implemented using a comparative approach to detect intrusion on a network based system, using the WSN-DS dataset. After the testing phase of this research, the random forest and decision trees classifier were the best performing models. It is safe to say that the enhancement of any detection system must implement these models for effective capturing of the network intrusion. From this research, the worst performance was observed in the logistic regression model because of its sigmoid output architecture.

Prior to the research, the deep learning models were the models with the most probable outcome but we can see the random forest outperformed the deep learning models even if it's not by a huge margin. In this research, the limitation can be seen in the authenticity of the dataset which may not be obtained from a real world network system.

Future work should consider incorporating more data retrieval techniques to obtain a more robust and sophisticated dataset. Also, implementing feature selection on the dataset obtained would help improve efficiency and effectiveness of the machine learning models.

REFERENCE

Ahmad, Z. *et al.* (2020) 'Network intrusion detection system: A systematic study of machine learning and deep learning approaches,' *Transactions on Emerging Telecommunications Technologies*, 32(1).

Ashiku, L. and Dagli, C., 2021. Network intrusion detection system using deep learning. Procedia Computer Science, 185, pp.239-247.

Barua, A. *et al.* (2022) 'Security and Privacy Threats for Bluetooth Low energy in IoT and wearable devices: A Comprehensive survey,' *IEEE Open Journal of the Communications Society*, 3, pp. 251–281.

Basheer, L. and Ranjana, P., 2022, May. A comparative study of various intrusion detections in smart cities using machine learning. In 2022 International Conference on IoT and Blockchain Technology (ICIBT) (pp. 1-6). IEEE.

Bharati, S. and Podder, P. (2022) 'Machine and deep learning for IoT security and privacy: applications, challenges, and future directions,' *Security and Communication Networks*, 2022, pp. 1–41.

Chatterjee, A. and Ahmed, B.S. (2022) 'IoT anomaly detection methods and applications: A survey,' *Internet of Things*, 19, p. 100568.

Chen, J., Qi, X., Chen, L., Chen, F. and Cheng, G., 2020. Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection. Knowledge-Based Systems, 203, p.106167.

Chowdhury, M.Z. *et al.* (2020) '6G Wireless Communication Systems: applications, requirements, technologies, challenges, and research directions,' *IEEE Open Journal of the Communications Society*, 1, pp. 957–975.

Dandekar, I. (2023) 'Introduction to Convolutional Neural Networks: Part 1,' *Medium*, 13 February.

DeMedeiros, K., Hendawi, A. and Alvarez, M. (2023) 'A survey of AI-Based anomaly detection in IoT and sensor networks,' *Sensors*, 23(3), p. 1352.

Djenna, A., Harous, S. and Saidouni, D.E. (2021) 'Internet of Things meet Internet of Threats: New concern Cyber security Issues of critical cyber infrastructure,' *Applied Sciences*, 11(10), p. 4580.

Dormehl, L. (2022) *What is an artificial neural network? Here's everything you need to know*.

Dwivedi, Y.K. *et al.* (2020) 'Setting the future of digital and social media marketing research: Perspectives and research propositions,' *International Journal of Information Management*, 59, p. 102168.

Farahani, M.A. *et al.* (2024) 'Time-series classification in smart manufacturing systems: An experimental evaluation of state-of-the-art machine learning algorithms,' *Robotics and Computer-Integrated Manufacturing*, 91, p. 102839.

Gunay, D. (2023) 'Random Forest - Deniz Gunay - Medium,' *Medium*, 14 September.

Hassan, E., Saleh, M. and Ahmed, A., 2020. Network intrusion detection approach using machine learning based on decision tree algorithm. Journal of Engineering and Applied Sciences, 7(2), p.1.

Hussein, H.M. *et al.* (2024) 'Comparative Study-Based Data-Driven Models for Lithium-Ion Battery State-of-Charge Estimation,' *Batteries*, 10(3), p. 89.

Kanade, V. (2024) *Everything you need to know about logistic regression*.

Khaliq, A.A. *et al.* (2022) 'A secure and privacy-preserved parking recommender system using elliptic curve cryptography and local differential privacy,' *IEEE Access*, 10, pp. 56410–56426.

Khraisat, A. *et al.* (2019) 'Survey of intrusion detection systems: techniques, datasets and challenges,' *Cybersecurity*, 2(1).

Kodheli, O. *et al.* (2020) 'Satellite Communications in the New Space Era: A survey and future challenges,' *IEEE Communications Surveys & Tutorials*, 23(1), pp. 70–109.

Krauhausen, I. *et al.* (2024) 'Bio-inspired multimodal learning with organic neuromorphic electronics for behavioral conditioning in robotics,' *Nature Communications*, 15(1).

Kumari, A. *et al.* (2019) 'Verification and validation techniques for streaming big data analytics in internet of things environment,' *IET Networks*, 8(3), pp. 155–163.

Lim, W.Y.B. *et al.* (2020) 'Federated Learning in Mobile Edge Networks: A Comprehensive survey,' *IEEE Communications Surveys & Tutorials*, 22(3), pp. 2031–2063.

Madani, S.S. *et al.* (2024) 'Recent progress of deep learning methods for health monitoring of Lithium-Ion batteries,' *Batteries*, 10(6), p. 204.

Mazhar, T. *et al.* (2023) 'Analysis of Cyber Security Attacks and Its Solutions for the Smart grid Using Machine Learning and Blockchain Methods,' *Future Internet*, 15(2), p. 83.

Musarat, M.A. *et al.* (2022) 'Health and Safety Improvement through Industrial Revolution 4.0: Malaysian Construction Industry Case,' *Sustainability*, 15(1), p. 201.

Nwafor, C.N., Nwafor, O. and Brahma, S. (2024) 'Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach,' *Scientific Reports*, 14(1).

Paul, R. and Das, K.N. (2023) 'Trends of Optimization Algorithms from Supervised Learning Perspective,' *Journal of Computational and Cognitive Engineering*[Preprint].

Porambage, P. *et al.* (2021) 'The roadmap to 6G security and privacy,' *IEEE Open Journal of the Communications Society*, 2, pp. 1094–1122.

Samriya, J.K. and Kumar, N., 2020, October. A novel intrusion detection system using hybrid clustering-optimization approach in cloud computing. In Materials Today: Proceedings (Vol. 2, No. 1, pp. 23-54). Elsevier.

Sapkota, D.B. *et al.* (2024) 'An artificial neural network based approach for harmonic component prediction in a distribution line,' *Energy Reports*, 12, pp. 3861–3873.

Shabbir, N. *et al.* (2024) 'Comparative Analysis of Machine Learning Techniques for Non-Intrusive Load Monitoring,' *Electronics*, 13(8), p. 1420.

Siam, S.I. *et al.* (2024) 'Artificial Intelligence of Things: a survey,' *ACM Transactions on Sensor Networks* [Preprint].

Tales, T.& (2023) 'Navigating the path to understanding decision trees in machine learning,' *Medium*, 9 September.

Tsiknas, K. *et al.* (2021) 'Cyber Threats to Industrial IoT: A survey on attacks and countermeasures,' *IoT*, 2(1), pp. 163–186.

Ugochukwu, C.J., Bennett, E.O. and Harcourt, P., 2019. An intrusion detection system using a machine learning algorithm. LAP LAMBERT Academic Publishing.

Varol, B., Omurlu, İ.K. and Türe, M. (2024) 'The effect of regularized regression and Tree-Based missing data imputation methods on classification performance in high dimensional data,' *Black Sea Journal of Engineering and Science* [Preprint].

Verkerken, M., D'hooge, L., Wauters, T., Volckaert, B. and De Turck, F., 2020, October. Unsupervised machine learning techniques for network intrusion detection on modern data. In 2020 4th Cyber Security in Networking Conference (CSNet) (pp. 1-8). IEEE.

Yin, C., Zhu, Y., Fei, J. and He, X., 2017. A deep learning approach for intrusion detection using recurrent neural networks. Ieee Access, 5, pp.21954-21961.

Zou, X. *et al.* (2024) 'Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook,' *Information Fusion*, 113, p. 102606.