Network Intrusion Detection Using AI/ML

Capstone project submitted as part of the fulfillment of the course curriculum for the third semester of the

Master of Science

in

Artificial Intelligence and Machine Learning

by

Divyanshu Mittal

(MSA23021)

under the guidance of

Dr. Abhinesh Kaushik

विद्या ददाति विनयं विनयाद् याति पात्रताम्

Abhinesh Kaushik

Indian Institute of Information Technology, Lucknow

2023-25

Declaration

I hereby declare that the project titled "**Network Intrusion Detection Using AI/ML**" is the result of my own work, conducted with dedication and under the valuable guidance of Dr. Abhinesh Kaushik at the Indian Institute of Information Technology, Lucknow. This project has not been submitted, either wholly or in part, for any other degree or academic credit at this or any other institution. All sources of information used have been acknowledged.

Divyanshu Mittal
Roll No: MSA23021

Certificate

This is to certify that the capstone project titled "**Network Intrusion Detection Using AI/ML**" submitted by Divyanshu Mittal (Roll No.: MSA23021) has been carried out under my guidance and supervision. This project has been conducted as part of the fulfillment of the course curriculum for the third semester of the Master of Science degree in Artificial Intelligence and Machine Learning. The work presented in this report is, to the best of my knowledge, a result of the student's independent efforts and it meets the academic standards required for this degree program. I hereby endorse the project and recommend it for evaluation.

Dr. Abhinesh Kaushik
(Department of Information Technology)
Indian Institute of Information Technology, Lucknow

Date: ___________ Signature: _________________

Acknowledgment

I would like to express my heartfelt gratitude to my supervisor, Dr. Abhinesh Kaushik, for his invaluable guidance, encouragement and continuous support throughout the development of this project. His insightful feedback, expert knowledge and mentorship were instrumental in shaping this work and helped me overcome numerous challenges during the research process.

I am also thankful to the esteemed faculty and dedicated staff at the Indian Institute of Information Technology, Lucknow, for providing an inspiring academic environment and the resources necessary for the completion of this project. Their assistance and support have been invaluable in helping me achieve my goals.

Finally, I would like to extend my appreciation to my peers and friends who offered their assistance and encouragement, making this journey both rewarding and memorable.

Contents

Chapter 1

Introduction

1.1 Introduction

Wireless Sensor Network (WSNs) is a network that consists of tiny and multiple nodes or sensing devices that are used to collect or gather the information from the observable or physical environment such as temperature, pressure, body movement etc. These tiny sensing devices are low in sizes and have not very high processing capability. To create a WSN efficiently, there are very challenges and issues. WSNs are used in many fields like science, military, security, health, to monitor traffic, to explore oceans, transportation, agriculture and many more. Security is a major issue in WSN. WSNs can be attacked by many ways. These attacks are called intrusions. Any type of unwanted or unapproved system activity is known as intrusion. Hence dealing with intrusions is very important and so intrusion detection is a prominent research field and a development topic also for those applications that require integrity, confidentiality and should be available for all time.

In 2014, according to Forbes research, "there were the intrusions include cyber attack stealing personal records of the users of eBay", "intrusion to Montana Health Department" and many more. It is very important to have the intrusion detection as it plays a very crucial role important for good security policy. "Intrusion detection is the second line of defense, the first line of security is intrusion prevention." These are the two main approaches for security management. Intrusion detection serves as an alarm mechanism for a network to secure it. IDS defined as a device that monitor the network to detect attacks or intrusions. For WSN, IDS added for detecting the physical damages done to the sensing nodes. Detecting sensor damage is important for fault tolerance and for ensuring availability of the network.

Enhancing the security of the network against the intrusion requires the upgradation of the IDS. Various researches have done the great job in the field of IDS from time to time then various researches have focused on ML based systems for IDS to protect the network by intrusions. ML based algorithms are mostly used for the purpose of classification, prediction and clustering. ML has a very effective role in the IDS in WSN using "Support Vector Machine (SVM), Random Forest, Logistic Regression, k-nearest neighbors and Gaussian Naive Bayes". With the help of these algorithms, ML can detect any type of attacks in the network. ML techniques have gained attention and credibility for their wonderful performance in the field of intrusion detection.

1.2 Problem Statement

In this modern digital era, cybersecurity threats have become increasingly sophisticated, posing significant challenges to traditional Network Intrusion Detection Systems (NIDS). These conventional systems often rely on static rules and signatures, making it difficult to detect new and emerging threats. This project focuses on utilizing AI and machine learning techniques to enhance the effectiveness of NIDS. The goal is to develop a system that detects the intrusions clearly. By integrating machine learning algorithms, the system aims to increase detection accuracy, lower false alarm rates and provide a more dynamic and proactive approach to network security.

1.3 Objective

To accomplish the project's purpose, the following particular objectives have been established:
i. Dataset collection and pre-processing.
ii. Machine-learning model selection and development.
iii. Model training and evaluation.
iv. Comparison of the results of different models.

1.4 Methodology

This project follows a systematic approach to develop a machine learning model for intrusion detection, leveraging NSL-KDD dataset. The following steps outline the methodology:

- **Data Collection:** The NSL-KDD dataset was loaded and explored to understand the structure and features from the [1] research paper. This research paper contains the information about the NSL-KDD dataset, its features.

- **Feature Extraction:** Relevant features were selected and the dataset was preprocessed including missing values, scaling numerical features using RobustScaler and encoding categorical variables.

- **Dataset Segmentation:** Divide the dataset into training and testing sets for model validation.

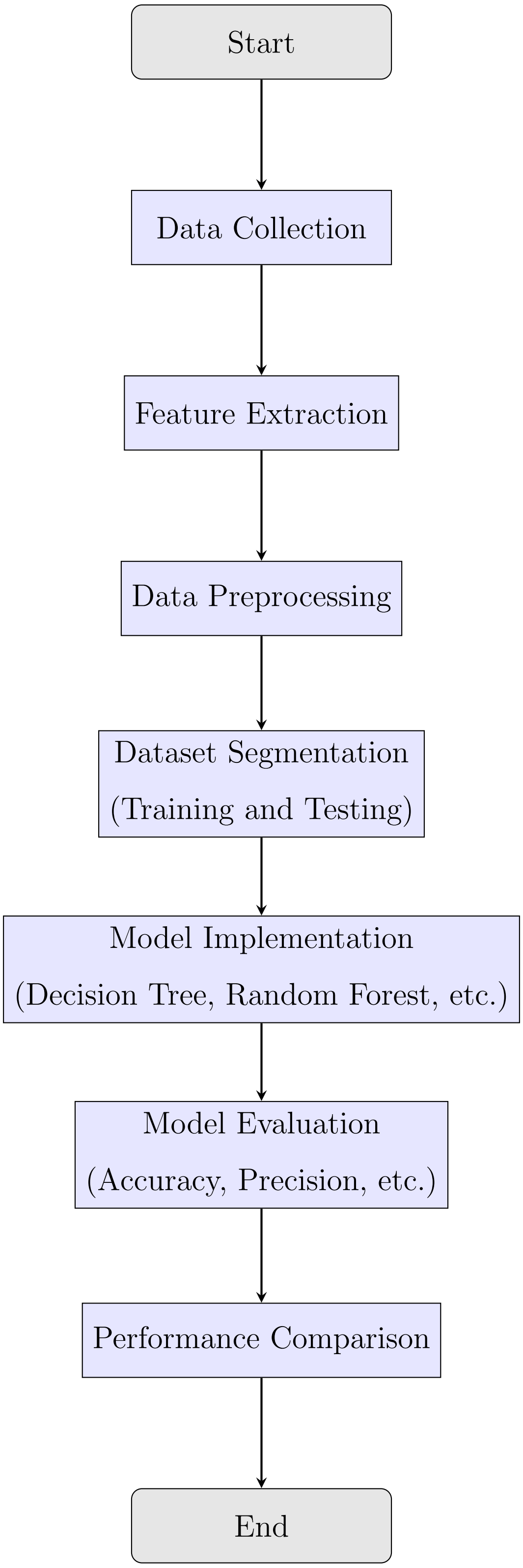
- **Model Implementation:** Implement various machine learning, including Logistic Regression, k-Nearest Neighbors, Support Vector Machine, Gaussian Naive Bayes, Decision Tree, Random Forest.

- **Evaluation Metrics:** Develop code to evaluate the performance of models based on accuracy and other metrics.

- **Model Comparison:** Compare the trained models to determine the most effective approach for intrusion detection.

Flowchart

The flowchart below visually represents the step-by-step process of the methodology.



Chapter 2

Literature Survey

In 2014 the authors in [2] proposed an IDS based on the kNN classifier algorithm. This system classifies the abnormal nodes from the normal nodes on the basis of their abnormal behaviour. The system classifies the abnormal nodes by calculating the distance of that suspecting node from its 'k' nearest neighbor and based upon the class majority, it assigns that respective class to the suspected node. These systems are efficient and fast performing. These systems are very important in the cases of flooding attacks in WSN. But these systems have some limitations:

- The computational cost of these systems is high.
- These systems perform slow on large datasets.
- These systems are sensitive to the choice of the value of 'k'. This should not be as the whole kNN algorithm depends upon it.
- These systems don't perform well if the features of the data are high.

In 2017 to overcome these limitations of the kNN algorithm systems, the authors in [3] proposed a few-shot deep learning approach with the algorithms like SVM and kNN. This few-shot deep learning approach improves the earlier algorithm's limitations. The approach of this few-shot deep learning algorithm is that it uses first a CNN model to extract the feature from the input data. This CNN model has multiple layers, it extracts outputs from multiple layers. Then it uses a SVM algorithm and kNN classifier with k=1 for intrusion detection.

Few-shot in this few-shot deep learning approach means the training samples are limited. Few-shot learning refers to those type of learning algorithms where the training samples to train the model are limited. This method also give better results to imbalanced datasets by the techniques of oversampling and under sampling. Experimental results also proved that this algorithm outperforms the previous algorithms trained on the "KDD 99" and "NSL-KDD" datasets. But this algorithm has some limitations:

- Due to the CNN used in this algorithm, it requires very large amount of data as deep learning requires more data.
- The computational requirements are high for this algorithm as it uses a CNN model and then SVM and 1-NN.
- Sometimes algorithm overfits on the training data.

In 2019, the authors in [4] proposed another deep learning model for IDS which reduces the limitations of previous algorithm very much. The proposed framework introduced using deep learning for IDS is called "Scale-Hybrid-IDS-AlertNet (SHIA)". This framework is very scalable, it can process very large network level and also the host level events to classify and detect the intrusions accurately. It do use CNN rather it uses deep neural networks i.e. deep ANNs to analyze real time data. Due to the deep neural networks, SHIA outperforms the previous algorithms in terms of results.
The advantage of using deep neural network is that it automatically learn and extract the features even from high dimensional data so it performs very well. But this algorithm has also some limitations:

- Due to the deep neural networks, this algorithm requires very large dataset. The datasets available do not fulfill the security and privacy concerns leading to the private datasets which are not trust worthy.
- Due to the deep neural networks, there are more parameters to train the algorithm resulting in the greater time to train the algorithm.
- The computational limitation is same as the previous algorithm. It requires high computational power to process.

In 2020, the authors in [5] proposed a fuzzy logic based algorithm called FzMAI which outperforms all the previous fuzzy algorithms. Because of this algorithm many researchers shift towards the fuzzy algorithms for IDS. This algorithm works in the three stages: “feature extraction, membership value computation and fuzzy rule applicator”. This algorithm categorizes the nodes in three different colors: red, orange and green. This algorithm used WSN-DS dataset. Using the fuzzy rules, this algorithm after categorization identify the malicious activity of the nodes and categorize the nodes based on the level of its trustworthiness.
The accuracy achieved by this algorithm is 98.29%. This accuracy is very high in comparison of the previous algorithms like: HNDMBFE with 61% accuracy and FBAHIDS with 98.20%. But this algorithm has some limitations:

- This algorithm mainly focuses on the intrusion prevention not the intrusion detection.
- The fuzzy logic is difficult in forming the formula to compute the logic.
- This algorithm uses a complex formula to compute which is very difficult to understand.

Gradually the researches come back to the machine learning approaches because of the difficulties in the reasoning of the fuzzy logic. In 2021, the authors in [6] proposed a “lightweight intelligent intrusion detection system”. This system consists of the combination of the kNN algorithm and the “sine cosine algorithm (SCA)”. This system has been tested on the “NSL-KDD” and “UNSW-NB15” datasets. This system also introduces the concept of compact sine cosine algorithm (CSCA) and an improved version also called PM-CSCA to increase the overall accuracy. The system is also suitable to deploy in the cloud computing and in the fog computing for real-time usage and for energy saving. The system combines the kNN and SCA algorithms for a lightweight intelligent IDS. The kNN algorithm has been used in this system as it is easy to implement and give better results than other classification algorithms. SCA is a heuristic optimization algorithm used to optimize the results obtained from the kNN algorithm. The combination of these two algorithms gives higher accuracy with the high rate of detecting the intrusions and low false alarm rate. The SCA algorithm optimizes the value of 'k' of kNN algorithm by the hyperparameter tuning techniques. But this algorithm has some limitations:

- The compact mechanism used in the SCA algorithm may reduce the accuracy of the algorithm.
- The kNN algorithm do not perform well on the higher dimensional data.
- The combination of kNN and SCA algorithms improve the accuracy but the trade-off between data transmission and sensor energy is still the issue.

Recently in 2024 itself, the authors in [7] proposed a new firefly algorithm called the FA-ML algorithm which enhances the security also with the intrusion detection. This algorithm consists of the combination of the "firefly algorithm" and the "SVM algorithm". The algorithm specifically handles the issues of limited resources and high dimensional data. The algorithm has been tested on the "NSL-KDD dataset" achieving the accuracy of 99.34%. It outperforms the previous algorithms and classifies the intruded nodes with a computational time of 6.3 seconds. The algorithm is uniquely suitable for the constraints of WSN-IOT intrusion detection systems. But this algorithm also has some limitations:

- The computational time is still very high as in this time the attack can cause much information loss or may be the software loss.
- The FA-ML algorithm is not very scalable. It is not useful for every intrusion detection.

Chapter 3

Model Development

3.1 Data Extraction

The collection of the NSL-KDD dataset was sourced from the open-source service **Kaggle**. The dataset can be accessed here: `https://github.com/d01mittal/Capstone-Project--Network-Intrusion-Detection/tree/main/nsl-kdd`.
The dataset contains 125972 rows with total 43 features.

3.2 Model Training

In this step, various machine learning are trained on the dataset to classify the requests as either normal or attack. Each model is evaluated based on its accuracy and robustness to ensure reliable intrusion detection.

Selection of Best Model

The model with the highest performance metrics, including accuracy, precision and recall is chosen. This model will serve as the primary classifier in the intrusion detection.

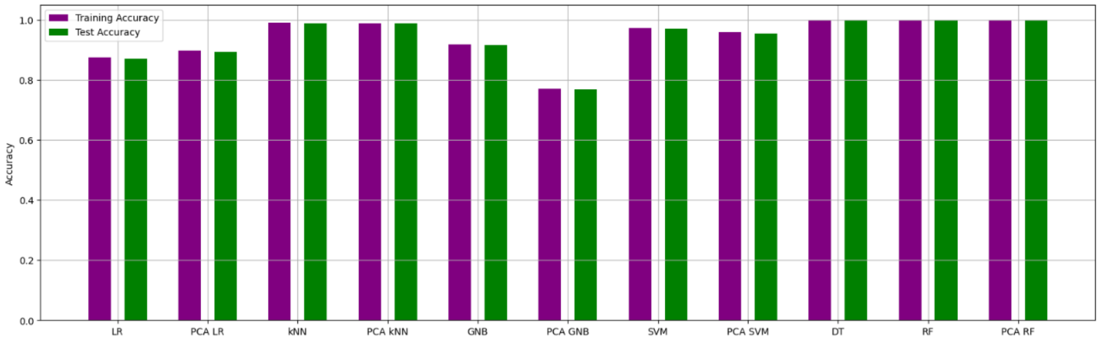
3.3 Result

The performance of the models varied:

- Random Forest Classifier showed high accuracy of 99.87% in identifying intrusions with a balanced trade-off between false positives and true positives.

- Decision tree with max_depth of 3 showed accuracy of 99.87% but were prone to overfitting compared to ensemble methods.
- Logistic Regression gave good results with accuracy of 87.17%.
- Support Vector Machine gave better results with accuracy of 97.04%.
- kNN gave better results with the n_neighbors value of 20 with accuracy of 98.94%.
- Gaussian Naive Bayes also gave good results with the accuracy of 91.61%.

The full code is uploaded on GitHub and can be accessed here: `https://github.com/d01mittal/Capstone-Project--Network-Intrusion-Detection` The comparison between the respective results based on the accuracy is as follows:



Chapter 4

Conclusion and Future Work

WSN plays a very crucial role in many fields, protecting them is very important. IDS acts as a line of defense which prevents WSN from the attacks or intrusions. Machine Learning (ML) algorithms like "SVM, kNN, Naive Bayes, Logistic Regression, Random Forest" and Neural Networks are very important in IDS for their capabilities of classification, clustering and prediction.
From kNN to deep learning neural networks like deep ANNs and CNNs is the journey for IDS in WSN. From time to time there is an improvement in the terms of accuracy, computational time, efficiency and scalability. There are many challenges also like computational cost, the complexities of the algorithms and to deal with the larger datasets.

In future, we can implement the intrusion detection model in real time using Kafka and other real-time data processing frameworks which will predict the requests behaviour at the time of hitting.

References

[1] S. S. Panwar and Y. Raiwani, “Data reduction techniques to analyze nsl-kdd dataset,” *Int. J. Comput. Eng. Technol*, vol. 5, no. 10, pp. 21–31, 2014.

[2] W. Li, P. Yi, Y. Wu, L. Pan, J. Li *et al.*, “A new intrusion detection system based on knn classification algorithm in wireless sensor network,” *Journal of Electrical and Computer Engineering*, vol. 2014, 2014.

[3] M. M. U. Chowdhury, F. Hammond, G. Konowicz, C. Xin, H. Wu, and J. Li, “A few-shot deep learning approach for improved intrusion detection,” in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. IEEE, 2017, pp. 456–462.

[4] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, “Deep learning approach for intelligent intrusion detection system,” *Ieee Access*, vol. 7, pp. 41 525–41 550, 2019.

[5] N. Singh, D. Virmani, and X.-Z. Gao, “A fuzzy logic-based method to avert intrusions in wireless sensor networks using wsn-ds dataset,” *International Journal of Computational Intelligence and Applications*, vol. 19, no. 03, p. 2050018, 2020.

[6] J.-S. Pan, F. Fan, S.-C. Chu, H.-Q. Zhao, and G.-Y. Liu, “A lightweight intelligent intrusion detection model for wireless sensor networks,” *Security and communication Networks*, vol. 2021, pp. 1–15, 2021.

[7] M. Karthikeyan, D. Manimegalai, and K. RajaGopal, “Firefly algorithm based wsn-iot security enhancement with machine learning for intrusion detection,” *Scientific Reports*, vol. 14, no. 1, p. 231, 2024.