Machine Learning Based Intrusion Detection System

JYOTSNA NANAJKAR[1], AANCHAL SINGH[2], GAURAV BHOI[3], MOHD. SAJEED SHAIKH[4], SURYAPRATIM DAS[5]

[1] *Department of Information Technology at Zeal College of Engineering and Research Pune, Savitribai Phule Pune University, Pune – Maharashtra.*

[2, 3, 4, 5] *Student, Department of Information Technology at Zeal College of Engineering and Research Pune, Savitribai Phule Pune University, Pune – Maharashtra.*

Abstract*— The increasing prevalence of network assaults presents a well-recognized challenge that can jeopardize critical information's availability, confidentiality, and integrity for individuals and organizations alike. In this paper, we introduce an intrusion detection methodology employing supervised machine learning. Our approach is straightforward yet effective, adaptable to various machine learning techniques. We tested several established machine learning methods to assess the efficacy of our intrusion detection system (IDS). Our empirical findings indicate that Support Vector Machines (SVM) and K-Nearest Neighbour (KNN) techniques outperform others. Consequently, we proceeded to develop an IDS utilizing SVM and KNN algorithms to classify online network data as either normal or indicative of an attack. Furthermore, we identified 12 crucial features of network data essential for detecting network attacks, employing information gain as our feature selection criterion. Our system can discern normal network activities from primary attack types (Probe and Denial of Service (DoS)) with a detection rate exceeding 98% within a 2-second timeframe. Additionally, we devised a novel post-processing method to mitigate the false-alarm rate and enhance the reliability and precision of the intrusion detection system.

***Index Terms*— Intrusion Detection, Support Vector Machine, K- Nearest Neighbour, LCCDE Ensemble Framework, Machine Learning**

I. INTRODUCTION

Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs) are the most important defence tools against the sophisticated and ever-growing network attacks. Due to the lack of reliable test and validation datasets, anomaly-based intrusion detection approaches are suffering from consistent and accurate performance evolutions.

Off-line network intrusion detection systems periodically scrutinize network information or log data to identify suspicious activities or potential intrusions. On the other hand, in an on-line network intrusion detection system, based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files) are also available in the extracted features definition. Network traffic data is continuously monitored as it arrives, enabling real-time detection of network attacks or malicious activities.

As internet services have become indispensable for both business transactions and individual use, the reliance on network services has surged, exposing critical information to escalating risks from remote intrusions. Enterprises are compelled to bolster their networks against malicious activities and various network threats. Consequently, a network infrastructure must incorporate one or more security measures such as firewalls, antivirus software, or intrusion detection systems (IDS) to safeguard vital data and services from potential hackers or intruders.

However, solely relying on a firewall system proves inadequate in shielding a corporate network from the array of network attacks. This inadequacy stems from the firewall's inability to thwart intrusion attempts targeting open ports essential for network services. Therefore, an intrusion detection system (IDS) is typically deployed in tandem with the firewall. The IDS functions by gathering data from network or computer systems and scrutinizing it for indications of potential breaches.

Network intrusion detection systems are categorized into two main types: host-based and network-based intrusion detection. Host-based detection involves capturing and analysing network data directly on the targeted system. Conversely, network-based detection

operates at the network gateway or server, intercepting and examining online network data before potential attacks can reach end-users. Moreover, these systems can function in two distinct modes: off-line detection and on-line detection.

II. DATASETS

KDD99 dataset contains benign and the most up-to-date common attacks, which resembles the true real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files). Also available is the extracted features definition.

This paper concentrates on both network-based and host-based intrusion detection, wherein incoming network data is captured in real-time, and detection outcomes are promptly reported, enabling network administrators to intervene and halt ongoing attacks. Additionally, our approach can function effectively as a host-based detection system. We constructed our intrusion detection system (IDS) utilizing a misuse detection technique, which categorizes attacks into distinct types. In contrast, anomaly detection methods solely differentiate between normal and abnormal/attack activities.

While numerous features of network data could potentially serve as input for an IDS, we advocate considering only 12 features extracted from the headers of data packets. We demonstrate the effectiveness of these 12 features in discerning normal network behaviour and categorizing primary attack activities into Port Scanning (PS or probing) and Denial of Service (DoS). By employing a limited number of features, we mitigate data analysis complexity, thereby enhancing detection speed and reducing CPU and memory consumption.

III. LITERATURE REVIEW

In previous research, most researchers have concentrated on off-line intrusion detection using a well-known KDD99 benchmark dataset to verify their IDS development. The KDD99 dataset is a statistically pre-processed dataset which has been available from DARPA since 1999 [13]. There also exist a few on-line intrusion detection approaches

Protecting computer and network information of an organizations and individuals become an important task, because compromised information can cause huge loss. Hence, intrusion detection system is used to prevent this damage. To enrich the function of IDS, different machine learning approaches get developed. The main objective [2] is to address the problem of adaptability of Intrusion Detection System (IDS).The proposed IDS has the proficiency to recognize the well-known attacks as well as unknown attacks. The proposed IDS consist of three major mechanisms: Clustering Manager (CM), Decision Maker (DM), Update Manager (UM). CICIDS2017 dataset is applied to estimate the working of the proposed IDS. Both supervised and unsupervised techniques were taken.

In this section, we present the experimental results and performance evaluation of the proposed Intrusion Detection System (IDS). We begin by outlining the network data utilized in the experiment.

accompanied. The information received to the system is grounded on the education of an agent who disregards the correction proposals presented by IDS. This technique is applied on supervised mode. Both known and unknown traffics can be detected by the system, when they work under unsupervised mode. After updating recently arrived data from both supervised and unsupervised modes, the function of the system has been improved. Performance of the system gets improved.

By incorporating machine learning techniques like, [3] SVM and Extreme Learning Machine (ELM), a hybrid model get developed. Modified K-means is used to construct high quality dataset. It builds small dataset that denote overall original training datasets. By this step, the training time of the classifier gets reduced. KDDCUP 1999 is used for implementation. It shows accuracy of about95. 75 percentages Various machine learning techniques like SVM, Random Forest (RF) and ELM are examined to report this problem. ELM shows better result when compared to other techniques in accuracy. Datasets get divided into one fourth of the data samples, half of the dataset and full datasets.

However, SVM produces better results in half of the data samples and one-fourth samples of data. ELM is the best method to handle the huge amount of data of about two lakh instances and more.

Over the past few years, as the development and proliferation of infinite communication paradigm and massive increase in the number of networked digital devices, there is considerable concern about cyber security that attempts to maintain the system's information and communication technology. Attackers identify and create new attacks on a daily basis, so attacks need to be correctly designed by the intrusion detection systems (IDSs) and appropriate responses should be provided, that are the primary objective of IDPS. IDSs, which play a very important role in network security, comprise three main components: data collection, feature selection/conversion and decision engine.

Machine Learning is used to automate analytical model building. It is a technique of data analysis. It is one of the branches of Artificial Intelligence which works on the concept that a system gets trained, make decisions and learn to identify patterns with fewer interventions of humans. Supervised and Unsupervised learning are the two most extensively used machine learning techniques. Labelled examples like an input with preferred output are taken for training algorithms. Instances without historical labels get trained using unsupervised learning. To discover some structure within the data and to explore the data are the two main objective of unsupervised learning. Apart from these methods, approaches like Semi supervised learning and Reinforcement learning are used.

Protecting computer and network information of an organizations and individuals become an important task, because compromised information can cause huge loss. Hence, intrusion detection system is used to prevent this damage. To enrich the function of IDS, different machine learning approaches get developed. The main objective [2] is to address the problem of adaptability of Intrusion Detection System (IDS). The proposed IDS has the proficiency to recognize the well-known attacks as well as unknown attacks. The proposed IDS consist of three major mechanisms: Clustering Manager (CM), Decision Maker (DM), Update Manager (UM). CICIDS2017 dataset is applied to estimate the working of the proposed IDS. Both supervised and unsupervised techniques were accompanied. The information received to the system is grounded on the education of an agent who disregards the correction proposals presented by IDS. This technique is applied on supervised mode. Both known and unknown traffics can be designed by the system, when they work under unsupervised mode. After updating recently arrived data from both supervised and unsupervised modes, the function of the system has been improved. Performance of the system gets improved, when it runs in unsupervised mode

Correlation-based feature selection method which is a simple filter-based model is used in the proposed system. Datasets containing the features, highly correlated with the class, yet uncorrelated with the others are applied. By using CICIDS2017 and UNSW-NB15 datasets this approach get achieved 99 percentages of detection rate of anomalies and 0.01 percentages of false positive rate. A hybrid method for A-NIDS using AdaBoost algorithms and Artificial Bee Colony to obtain low false positive rate (FPR) and high detection rate (DR).

Publisher	Author	Year	Name of Paper	Methodology
IEEE Access 2020	Akhil Krishna , Ashik Lal M A , Athul Joe Mathewkutty , Dhanya Sarah Jacob and Hari M	2020	Intrusion Detection and Prevention System Using Deep Learning	Artificial Neural Network , Multilayer perceptron, Convolutional Neural Networks and Recurrent Neural Networks
Springer Link	Inadyuti Dutt ,	2019	Machine Learning Based Intrusion	Support Vector Machine
	Samarjeet		Detection	(SVM) and

	Borah and		System	Naive Bayes
	Indra Kanta			
	Maitra			
Springer	Nasrin	2019	Survey on	KNN , SVM
Link	Sultana ,		SDN based	and PCA
	Naveen		network	
	Chilamkurti ,		intrusion	
	Wie Peng ,		detection	
	Rabei		system	
	Alhadad		using	
			machine	
			learning	
			approaches	
Springer Link	Inadyuti Dutt , Samarjeet Borah and Indra Kanta	2018	Real Time Hybrid Intrusion Detection System	Support Vector Machine (SVM) and Naive Bayes
	Maitra		Using Machine Learning Techniques	
IEEE Access 2020	Amir Ali	2020	Novel Three-Tier Intrusion Detection and Prevention System in Software Defined Network	User Validation , Packet Validation and Flow Validation.
Springer Link	Ahmed Aleroud and George Karabatis	2017	Ahmed Aleroud and George Karabatis	Ahmed Aleroud and George Karabatis

IV. PROPOSED SYSTEM

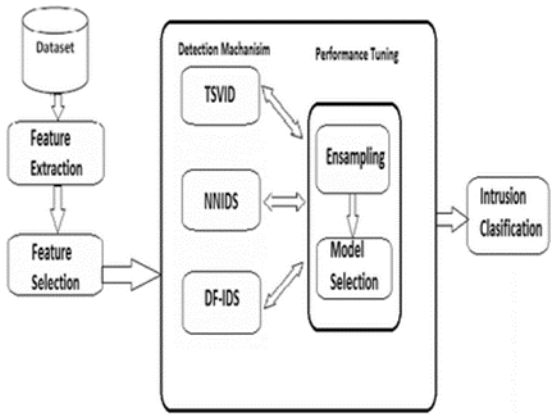

Fig 1. System Architecture

A] Machine Learning Techniques:

Intrusion detection fundamentally revolves around classification Intrusion, where an Intrusion Detection System (IDS) must categorize a given set of network packets as either normal or indicative of an attack. Various characteristics or features of the network data stream serve as input for this classification. The selection of features impacts computational complexity and the required computer resources, thus it's desirable to identify the smallest set of relevant or effective feature.

Machine Learning is employed to automate the process of building analytical models. It operates within the domain of data analysis and is a branch of Artificial Intelligence. Machine Learning operates on the principle that systems can be trained to make decisions and identify patterns with minimal human intervention. The two primary techniques used in Machine Learning are Supervised and Unsupervised learning. Supervised learning involves labelled examples, where inputs are paired with preferred outputs for training algorithms. Conversely, unsupervised learning operates on instances without historical labels, aiming to uncover structure within the data and explore its patterns.

Additionally, approaches such as Semi-supervised learning and Reinforcement learning are utilized. Semi-supervised learning employs a smaller amount of labelled data and a larger amount of unlabelled data for training purposes. Reinforcement learning employs a trial-and-error method, where actions lead to rewards, aiming to select actions that yield the best rewards. The primary components in reinforcement learning are the agent, environment, and actions. The goal is for the agent to select actions that exploit predictable rewards, ultimately reaching the goal more efficiently by applying a good policy.

B] IDS process and algorithm:

In this section, we present the experimental results and performance evaluation of the proposed Intrusion Detection System (IDS). We begin by outlining the network data utilized in the experiment. Subsequently, we detail our experimental design and the performance metrics employed to evaluate the IDS. Finally, we present the experimental findings.

The presence of intrusion poses a significant threat to computer and network systems, potentially resulting in the theft or destruction of information within a short timeframe. Consequently, intrusion stands out as a major concern in network security, as it can also inflict damage on system hardware. While various intrusion detection techniques are employed, accuracy remains a primary challenge. The detection rate and false alarm rate are pivotal factors in assessing accuracy. Therefore, enhancing intrusion detection capabilities is crucial for reducing false alarms and ensuring the effectiveness of the system.

Thus, Support Vector Machine (SVM) and Naïve Bayes are applied. Classification can be addressed by these algorithms. Apart from that, Normalization and Feature Reduction are also applied to make a comparative analysis A new hybrid classification algorithm on Artificial Bee Colony (ABC) and Artificial Fish Swam (AFS) is proposed [6]. Nowadays computer system is prone to different information thefts due to the widespread usage of internet, which leads to the emergence of IDS. Fuzzy C Means Clustering (FCM) and Correlation-based Feature Selection (CFS) is applied [6] for separating training datasets and to eliminate irrelevant features. If-then rues are generated by using CART technique, which is applied to differentiate normal and anomaly records. In the detection phase a deep learning model is created for detecting intrusions and any possible threats that may encounter in our network, this is done with a sequence of steps which make our model with maximum possible accuracy and negligible loss. In this system they allocate part of the network to a database node. To send some data to any node in the network first send it to the server and then the server will redirect the data.

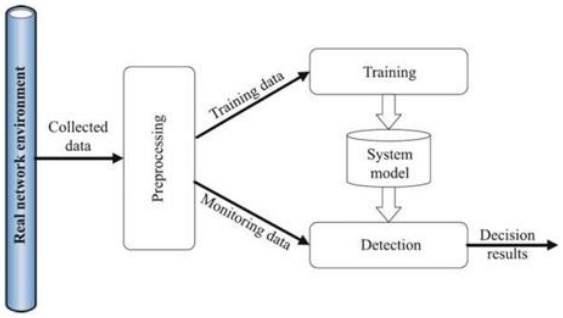

Fig 2. Block Diagram

V. METHODOLOGY

A] Support Vector Machines:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or

decision boundary that can segregate n- dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

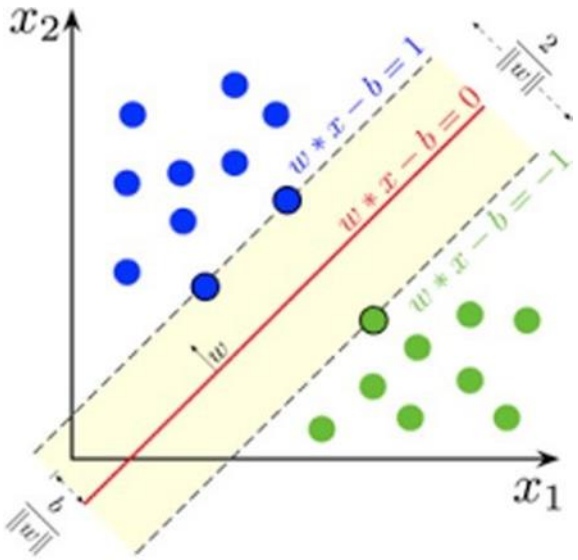

Fig no. 3

B] K-Nearest Neighbour

K-Nearest Neighbour is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

Distance Metrics Used in KNN Algorithm

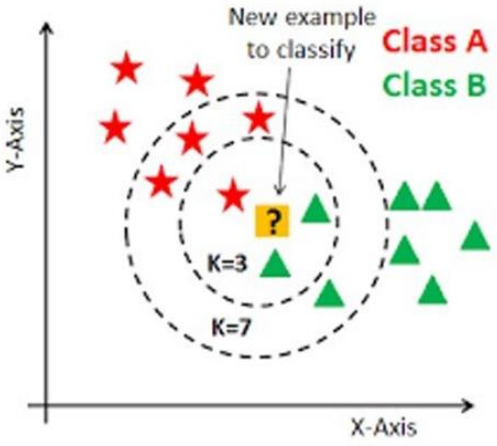

Fig no. 4

C] LCCDE (Leader Class and Confidence Decision Ensemble): Proposed Ensemble Algorithm: -

The performance of different ML models often varies on different types of attack detection tasks. For example, when applying multiple ML models on the same network traffic dataset, a ML model perform the best for detecting the first type of attack (e.g., DoS attacks), while another ML model may outperform other models for detecting the second type of attack (e.g., sniffing attacks). Therefore, this work aims to propose an ensemble framework that can achieve optimal model performance for the detection of every type of attack. Ensemble learning is a technique that combines multiple base ML models to improve learning performance and generalizability. The proposed ensemble model is constructed using XGBoost, LightGBM, and CatBoost, three advanced gradient boosting ML methods.

Accuracy Table :- Accuracy rate of attacks

Algorithm	Accuracy
LCCDE	99.44%
SVM	98.55%
KNN	98.29%

Table no. 1

CONCLUSION

In this search , the datasets are treated by the three algorithms known as LCCDE SVM AND KNN in which the KNN is eliminated due to weak result,

while LCCDE and SVM have shown good performance with the size of dataset or type of attacks it contained. This model will be optimized in future work in terms of processing time and also, we will work on its implementation on a firewall and real-time.

REFERENCES

[1] H. Wang, J. Gu, and S. Wang, ‘‘An effective ACCESS, Survivability Strategies for Emerging Wireless Networks, Volume.6, May 2018, pp.33789-33795.

[2] BuseGulAtli1, Yoan Miche, Aapo Kalliola, Ian Oliver, Silke Holtmanns, Amaury Lendasse; “Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space” SPRINGER, Cognitive Computation, June 2018, pp. 1-16

[3] Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R. Lyu; “A Feature Reduced Intrusion Detection System Using ANN Classifier”, ELSEVIER, Expert Systems with Applications, Vol.88, December 2017 pp.249-247

[4] Vajiheh Hajisalem, Shahram Babaie; “A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection”, ELSEVIER, Department of Computer Engineering, Vol. 136, pp. 37-50, May 2018 detection framework based on SVM with feature augmentation,’’ Knowl.-Based Syst., vol. 136, pp. 130– 139, Nov. 2017.

[5] Setareh Roshan, Yoan Miche, Anton Akusok, Amaury Lendasse; “Adaptive and Online Network Intrusion Detection System using Clustering and Extreme Learning Machines”, ELSEVIER, Journal of the Franklin Institute, Volume.355, Issue 4, March 2018, pp.1752-1779.

[6] Wathiq Laftah Al-Yaseen, Zulaiha Ali Othman, Mohd Zakree Ahmad Nazri; “Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System”, ELSEVIER, Expert System with Applications, Volume.66, Jan 2017, pp.296-303.

[7] Vasudeo, S. H., Patil, P., & Kumar, R. V. (2015). IMMIX-intrusion detection and prevention system. 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM). doi:10.1109/icstm.2015.7225396

[8] Ahmed, M., Pal, R., Hossain, M. M., Bikas, M. A. N., & Hasan, M. K. (2009). NIDS: A Network Based Approach to Intrusion Detection and Prevention. 2009 International Association of Computer Science and Information Technology - Spring Conference. doi:10.1109/iacsit-sc.2009.96

[9] "Multi-Layer Perceptron (MLP) Models on Real WorldBankingData”. Medium2020, https://becominghuman.ai/multi-layerperceptron-mlp- models-on-real-world-banking- dataf6dd3d7e998f?gi=7057648ed14f Network, Intrusion. "Intrusion Detection Using Artificial Neural Network - Docshare. Tips”. Docshare.Tips,2020, http://docshare.tips/intrusion-detection-using-artificial-neuralnetwork_584e6fd3b6d87f49628b524f.html.

[10] "An Introduction to IDS | Symantec Connect". Symantec.Com,2020, https://www.symantec.com/connect/ articles/introduction-ids.

[11] Sonali Rathore, Prof. Amit Saxena, and Dr. Manish Manoria. “Intrusion Detection System on KDDCup99 Dataset: A Survey.” IJCSIT) International Journal of Computer Science and Information Technologies, vol. 6, no. 4, 2015

[12] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, Aneel Raheem; “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection”, IEEE