International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Survey on: Machine Learning Based Intrusion Detection System

Prof. Jyotsna Nanajkar[1], Aanchal Singh[2], Gaurav Bhoi[3], MohdSajeed Shaikh[4], Suryapratim Das[5]
Department of Information Technology at Zeal College of Engineering and Research Pune,
Savitribai Phule Pune University, Pune – Maharashtra.

Abstract - The growing prevalence of network attacks is a well-known problem which can impact the availability, confidentiality, and integrity of critical information for both individuals and enterprises. In this paper, we propose a intrusion detection approach using a supervised machine learning technique. Our approach is simple and efficient, and can be used with many machine learning techniques. We applied different well-known machine learning techniques to evaluate the performance of our IDS approach. Our experimental results show that the Support Vector Machines (SVM) and K-Nearest Neighbour technique can outperform the other techniques. Therefore, we further developed a intrusion detection system (IDS) using the SVM and KNN Algorithms to classify on-line network data as normal or attack data. We also identified 12 essential features of network data which are relevant to detecting network attacks using the information gain as our feature selection criterions. Our can distinguish normal network activities from main attack types (Probe and Denial of Service (DoS)) with a detection rate higher than 98% within 2 s. We also developed a new post-processing procedure to reduce the false-alarm rate as well as increase the reliability and detection accuracy of the intrusion detection system.

Keywords: - Intrusion Detection, Support Vector Machine, K- Nearest Neighbour, LCCDE Ensemble Framework, Naive Bayes, Machine Learning

Data Sets: - 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.

CICIDS2017 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.

I. Introduction

Internet services have become essential to business commerce as well as to individuals. With the increasing reliance on network services, the availability, confidentiality, and integrity of critical information have become increasingly compromised by remote intrusions. Enterprises are forced to fortify their networks against malicious activities and network threats. Therefore, a network system must use one or more security tools such as a firewall, anti-virus software or an intrusion detection system to protect important data/services from hackers or intruders.

Relying on a firewall system alone is not sufficient to prevent a corporate network from all types of network attacks. This is because a firewall cannot defend the network against intrusion attempts on open ports required for network services. Hence, an intrusion detection system (IDS) is usually installed to complement the firewall. An IDS collects information from a network or computer system, and analyses the information for symptoms of system breaches. As shown schematically in Fig. 1, a network IDS monitors network data and gives an alarm signal to the computer user or network administrator when it detects antagonistic activity on an open port. This signal allows the recipient to inspect the system for more symptoms of unauthorized network activities.

Network intrusion detection systems can be classified into two types which are host-based and network-based intrusion detection. Host-based detection captures and analyses network data at the attacked system itself while the network-based detection captures and inspects online network data at the network gateway or server, before the attack reaches the end users. In addition, network intrusion detection systems can operate in two modes which are off-line detection and on-line detection. An off-line network intrusion detection system periodically analyses or audits network information or log data to identify suspected activities or intrusions. In an on-line network intrusion detection

system, the network traffic data has to be inspected as it arrives for detecting network attacks or malicious activities.

In this paper, we focus on network and host based intrusion detection where the incoming network data is captured on-line and the detection result is reported instantaneously or within a fraction of a minute, so that the network administrator is notified and can stop the on-going attack. Our approach also could be applied as host-based detection. We designed our intrusion detection system (IDS) using a misuse detection technique. The misuse detection approach can classify attacks into categories. In contrast, the anomaly detection approach can only differentiate between normal activity and abnormal/attack activity. Although, there are many possible features of network data that could serve as input to an IDS, we propose to consider only 12 features of network traffic data extracted from the headers of data packets. We show that these 12 features are effective in identifying normal network activity and classifying main attack activities into two attack types namely Port Scanning (PS or probing) and Denial of Service (DoS). Using a small number of features reduces the complexity of data analysis and thus can increase the detection speed and reduce computer resource (CPU and memory) consumption.

II. 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 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. Libelled 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 designedby 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 Borah and Indra Kanta Maitra		Detection System	(SVM) and Naive Bayes
Springer Link	Nasrin Sultana , Naveen Chilamkurti , Wie Peng , Rabei Alhadad	2019	Survey on SDN based network intrusion detection system using machine learning approaches	KNN , SVM and PCA
Springer Link	Inadyuti Dutt , Samarjeet Borah and Indra Kanta Maitra	2018	Real Time Hybrid Intrusion Detection System Using Machine Learning Techniques	Support Vector Machine (SVM) and Naive Bayes
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

III. Proposed System

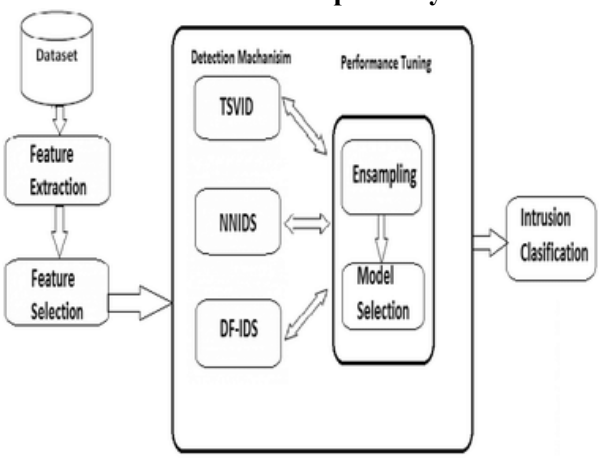

Fig 1. System Architecture

A] Machine Learning Techniques:

Intrusion detection is fundamentally a problem of classification. An IDS must classify a given set of network packets as normal or attack. Various characteristics or features of the network data stream provide input to the classification. The number of features used impacts the complexity of computation and the amount of computer resources required. Thus, we would like to identify the smallest possible set of relevant or effective features. 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. For training purpose, semi supervised learning uses fewer amounts of labelled data and huge amounts of unlabelled data. Trial and error method is used in Reinforcement Learning in which the actions yield the best rewards. Classification, regression and prediction are used. Agent, environment and actions are the three primary components used in this type of learning. The goal is that, the agent has to select those actions, which exploit the predictable reward. By applying good policy, the agent able to reach the goal much faster.

B] IDS process and algorithm:

In this section, we present the experimental results and performance evaluation of proposed IDS. We first present the network data used in the experiment. We then describe our experimental design and performance metrics used for evaluating the IDS. Finally, we present the experimental results.

A distinct existence of intrusion can steal or eliminate information from computer or network systems in limited duration. Hence intrusion is one of the major issues in network security. System hardware also gets harm due to intrusion. Various techniques of intrusion detection are performed; however, accuracy is one of the major problems. Detection rate and false alarm rate plays an essential role for the analysation of accuracy. Intrusion detection must be enriched to reduce false alarms and to increase the detection rate. 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.

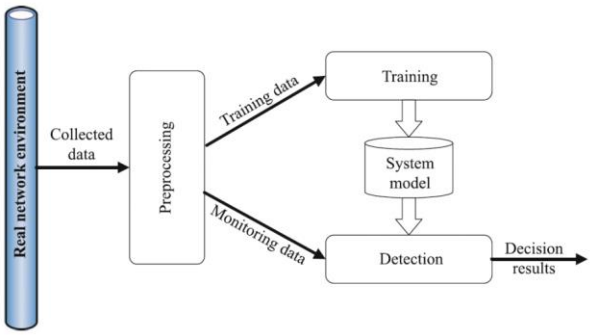

Fig 2. Block Diagram

Methodology:

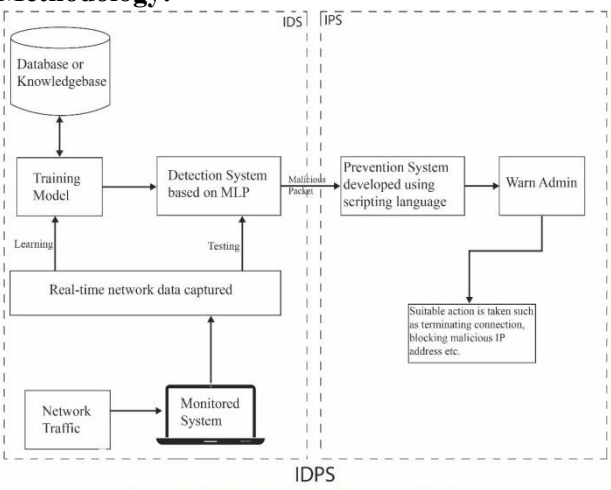

Figure 1 Intrusion Detection and Prevention System Architecture

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.

B] K-Nearest Neighbour:

K-Nearest Neighbours 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

C] Naïve Bayes: -

Bayesian classifiers are statistical classifiers. They are capable to forecast the probability that whether the given model fits to a particular class. It is based on Bayes' theorem. It constructed on the hypothesis that, for a given class, the attribute value is independent to the values of the attributes. This theory is called class conditional independence.

D] 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.

Expected Result:

The first attempt was a machine learning model with decision tree algorithm but the accuracy was only about 74% at around 100 epochs which was customized into another Supervised Machine Learning model namely SVM (Support Vector Machine) with linear kernel and random state 1. The newly generated SVM model was tested at 100 epochs and accuracy of 83% was obtained. To make the model more accurate machine learning MLP (Multilayer Perceptron) model with sci-kit learn high level API was selected consisting of 2 dense layers having activation functions such as relu and softmax. The loss used is sparse categorical cross entropy with Adam optimizer with a batch size of 16 that has been run for about 100 epochs. In the end an accuracy of 91.4% has been got and the model for intrusion detection system was finalized. A Wireshark tool was used to obtain real time network packet data and was exported into a csv file consisting of values needed for our model. Features like the IP address and the port number were considered for preventing the user from further malicious activities by using the administrative privileges and was established through a script.

IV. Conclusion

In this paper, we presented a practical and network and host-based intrusion detection system which detects the model and can be used with existing well-known machine learning algorithms. Our model consists of three phases: the pre-processing phase, the classification phase, and the postprocessing phase. We also presented how we preprocess the network packet header data into records of 12 essential features. The essential features were identified with the information gain method, to see. Intrusion detection and Intrusion prevention are needed in current trends. As our regular events are mainly dependent on networks and information systems, intrusion detection and intrusion prevention are very vital. Many approaches have been applied in intrusion detection systems. Among them machine learning plays a vital role. This analysis deals with machine learning algorithms like SVM and Naïve Bayes. It proposes while dealing with 19,000 instances SVM outperforms Naïve Bayes.

V. References

[1] H.Wang,J.Gu,andS.Wang,"An effective intrusion detection framework based on SVM with feature augmentation," Knowl.-Based Syst., vol. 136, pp. 130–139, Nov. 2017.

[2] 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.

[3] 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.

[4] 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

[5] 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

[6] "Multi Layer Perceptron (MLP) Models On Real World Banking Data". Medium, 2020, 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.

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

[8] 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

[9] 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 ACCESS, Survivability Strategies for Emerging Wireless Networks, Volume.6,May 2018,pp.33789-33795.

[10] BuseGulAtli1, YoanMiche,AapoKalliola, IanOliver, SilkeHoltmanns, AmauryLendasse; "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

[11] 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

[12] 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