International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT22819

Network Intrusion Detection System Using Machine Learning

Shailaja Jadhav, Vinaya Bhalerao, Varsha Yadav, Snehal Kamble, Bhavana Shinde
Department of Computer Engineering, Marathwada Mitra Mandal's College of Engineering, Karve Nagar, Pune, Maharashtra, India

Article Info
Volume 8, Issue 1
Page Number : 74-81

Publication Issue :
January-February-2022

Article History
Accepted : 10 Jan 2022
Published : 23 Jan 2022

ABSTRACT

The "Network Intrusion Detection System Based on Machine Learning Algorithms" is a component of software that invigilate a network of computers detecting potentially hazardous activities like capturing sensitive secret data or corrupting/hacking network protocols. Today's IDS techniques are incapable of doing this cope with the many sorts of security cyber-attacks on computer networks that are dynamic and complex. The effectiveness of an intruder the precision of detection is crucial. Intrusion detection accuracy must be able to reduce the number of false alarms and raise the pace at which alerts are detected. Various methods have been used to escalate the performance.
In recent studies, approaches have been applied. The main function of this group is to analyze large amounts of network traffic data system for detecting intrusions to address this, a well-organized categorization system is necessary issue. Machine Learning methods like Support Vector Machine (SVM) and Naıve bayes are applied for evaluation of IDS. NSL-KDD knowledge discovery data set is used, their accuracy and misclassification rate get calculated.
Keywords : Support Vector Machine (SVM), Naive bayes, Dynamic Complex types of security.

I. INTRODUCTION

Intrusion detection systems (IDS) recognize and deflect attacks using either a network or a host-based technique. In either case, these products look for attack signatures that usually indicate malicious or suspicious intent. If IDS looks at these patterns in network traffic then it is network based. It is host-based when an IDS looks for attack signatures in log files. Various algorithms have been developed to identify various types of network intrusions; however, no methodology has been developed to validate the accuracy of their results. The exact efficiency of a network intrusion detection system's ability to identify malevolent sources cannot be reported unless detailed performance measurement is given. This is this study outlines the behavior pattern of machine learning for identifying intruders. This is quite beneficial for avoiding intrusions based on the specific type of attack. The model can also achieve real-time intrusion identification based on dimensionality reduction and on a simple classifier.

II. BACKGROUND

Intrusion Detection System (IDS):

An Intrusion is an unaccridited approach or spiteful practice of a computer system. An Intruder tries to acquire unaccredited data and carries out destruction to the spiteful task exists. Intrusion Detection System (IDS) is employed to find out those kinds of unaccredited task going on to the system. Hence IDS may be also be a security system that observes computer system. IDS are security systems finding various several activities the strike on the system and keep our systems shielded. Intrusion Detection System Functions are as follows the functioning of IDS is fulfilled in four stages that is data collection, feature selection, analysis, action.

Figure 1: Capabilities of IDS

1. Data Collection

This unit collects the data and handover it to IDS. Here the data is stored and it is examined during this stage.

2. Feature Selection

This unit selects a feature from the information which exist on the server or dataset.

3. Analysis

In this unit IDS are used for examining the data and observes the computer or network system nature.

4. Action

This unit explain about the threat or attack of the system.

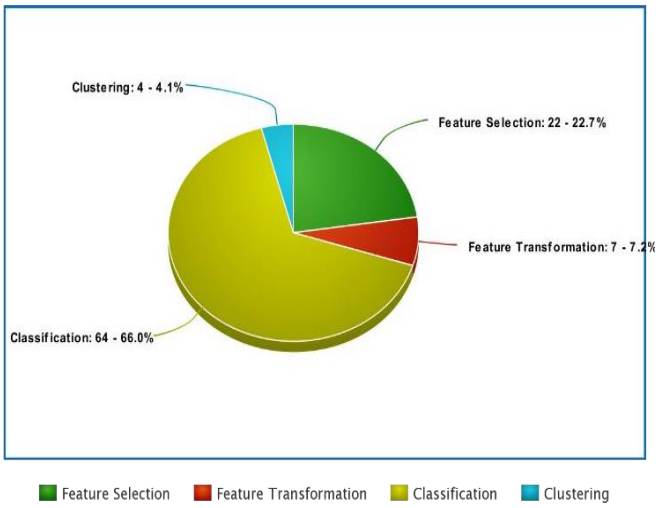


Figure 2 : Research focused area in IDS from 2016-2020

III. Machine Learning

This part describes Machine Learning and basic terminologies of machine learning. Now a days the utilization of Machine Learning has been increasing in corporation world Machine Learning is a branch of computing that allows systems to learn from their experiences without being explicitly programmed. Machine learning is concerned with the development of computer programs that will access data and use it to find out for themselves.

Machine learning algorithms are grouped into three categories as follow: (a) supervised learning, (b)unsupervised learning, and (c) reinforcement learning.

A. Supervised Machine Learning Algorithm: we have trained data with known labels. Machine algorithm trained on labelled data. during this type the data must be labelled accurately to figure. The foremost supervised learning technique is classification.

B. Unsupervised Machine Learning Algorithm: Throughout this learning experience, trained data with unidentified labels is available. And it discovers instantly from the data-based supported cosine similarity. Clustering is the most extensively

used unsupervised learning technique.

C. Reinforcement Learning: Throughout that learning experience, a computer was made available for the purpose of achieving a particular goal. It features self improving algorithm that learns from new situations through trial and error.Machine learning for IDS can solve a variety of issues, including speed and computational time, while also allowing for the development of accurate IDS.

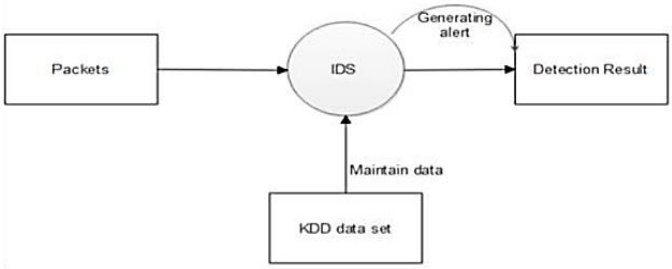


Figure 3. DFD0

Techniques for Machine Learning :

Machine learning algorithms are enacted for trying to solve ids issues supported configuration of single classifier. An Ids developed by one classifier or simple classifier.

A. Naive bayes:

Algorithm of naive bayes may also be a supervised classification algorithm. It is based on the bayes theorem with the independence assumption between all class values. There are three types of naive bayes algorithms, as demonstrated below:

1. gaussian: this is a supervised algorithm appropriate for categorical datasets with normal distribution.
The probability calculations are being done using the following equation:

P (ajx) = p (xja)p
(a) P (x)

P(a|x) represents the posterior probability. P(a) is the primary attack probability.

P (x |a) represents the likelihood, which is that the predictor's probability for a given class. P(x): is predictor's prior probability.

2. multi nominal naive bayes: this is a supervised algorithm for prolonged datasets that take compact tally. For example, in calculating the number of occurrences of words in text 3. Bernoulli naive bayes: this is a binomial model that is appropriate for both continuous and categorical data sources, but the selected features have to be in binary, that's also, zeros and one's. So, our set of data is a categorical model with two or quite two groups of threats, so the gaussian naive bayes algorithm is selected.

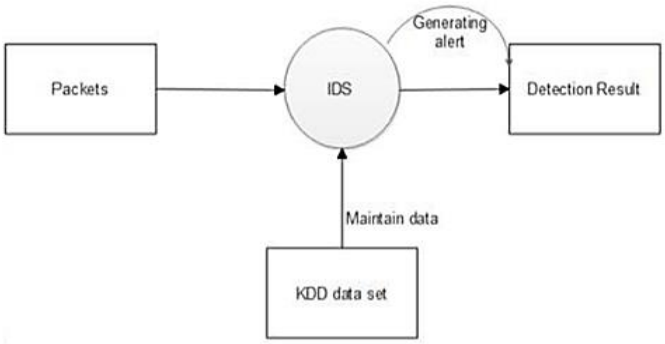


Figure 4. DFD 1

IV. LITERATURESURVEY

Ahmad et al [1] analyzed and compared well-known machine learning methods like support vector mechanism and extreme learning machine. The interruption detection system is evaluated using the datasets NSL and data mining datasets. In their analysis result, it's concluded that ELM is more precise than RF, SVM on complete data samples, and SVM more precise on partial samples, besides in quarter dataset SVM is better.

M. Al-Qatfet al. [4] proposed an IDS technique using self-taught learning (STL) which is an active deep learning technology for feature learning and dimensionality. This is done with the scant auto encoder device, which is an excellent unsupervised learning approach for rebuilding a unique feature illustration.The paper presently enhances SVM

categorization accurateness and faster training and testing times. It also displays accurate estimations in two and five-category classifications. When compared to other shallow classification methods such as J48, Naive Bayesian, RF, and SVM, this methodology achieves a greater accuracy rate in five-category classification.

Xu et al [5] introduced deep learning hypothesis for IDS uses feature extraction to construct a deep learning model. He proposed intrusion detection that comprises of a discontinuous neural system with gated recurrent units (GRU), multilayer perceptron (MLP), soft max module. The research was prepared on both the KDD dataset and NSL-KDD data sets. The combined results of BGRU and MLP for the KDD 99 and NSL-KDD datasets are superior, according to this article. Naseer et al [6] looked at a viable solution for anomaly-based IDS assembled on different deep neural networks, such as convolutional neural systemsautoencoders and periodic neural systems. These were trained on the NSLKDD dataset and estimated using NSLKDDTest+ and NSLKDDTest21 on a GPU-based testbed using Kera's no back end. In this, evaluations were done using the organization metrics viz. recipient working attribute, the area under the curve, precision-recall curve, mean average precision and accuracy of classification for deep as well as a conventional machine learning technique.

M.H. Ali et al [7] introduced an established knowledge model for fast learning net- work (FLN) supported particle swarm optimization (PSO) is planned. This is applicable to the detection of intruders and is backed by the notable dataset KDD99. The established system is correlated against a good vary of metaheuristic systems to tutor extreme learning system as well as FLN classifier. Much differentiation has been accomplished with a special variety of neurons within the unseen layer of FLN, and therefore the unique ELM that improve the FLN guidelines to boost the IDS accuracy.

Table 1: List of Recent Researches In NIDS

Reference	Dataset	Method	Accuracy
E. K.Viegas and L. S. Oliveira, "Towards reliable anomaly- based intrusion detection in real- world environments," Computer Networks, vol. 127, pp. 200–216, 2017.	TRAbID (Probe, DoS)	Decision Tree (DT) and Naïve Bayes (NB)	Probe; DT (98.42), NB (97.29) DoS; DT (99.90), NB (99.66)
T. Hamed, R. Dara, and S. C. Kremer, "Network intrusion detection system based on recursive feature addition And bigram technique," Computer Secure, vol. 73, pp. 137–155, 2018.	ISCX 2012	Recursive Feature Addition (RFA) with SVM	91.90
A. H. Hamamoto, L. F. Carvalho, L. D. H. Sampaio, T. Abrão, and M. L. Proença, "Network Anomaly Detection System using Genetic Algorithm and Fuzzy Logic," Expert Syst. Appl., vol. 92, pp. 390–	Real network traffic	Fuzzy Logic	96.50

402, 2018.			
W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," Expert Syst. Appl., vol. 67, pp. 296–303, 2017.	KDDCup99	Multi-level hybrid Support Vector Machine (SVM) and ELM	95.80
I. Sumaiya Thaseen and C. Aswani Kumar, "Intrusion detection model using fusion of chi- square feature selection and multi class SVM," J. King Saud Univ. - Computer. Inf. Sci., vol. 29, no. 4, pp. 462–472, 2017.	NSL-KDD	SVM-Radial Basis Function (RBF)	98.10
D. Papamartzivanos, F. Gómez Marmol, and G. Kambourakis, "Dendron: Genetic trees driven rule induction for network intrusion detection systems," Future Generation Computer Syst., vol. 79, pp. 558–574, 2018.	KDDCup99 NSL-KDD UNSW- NB15	Dendron (DT and GA)	KDDCup99 (98.90), NSlKDD (97.60), UNSQ-NB15 (84.30)

V. General Discussion

The KDD Cup dataset is used to assess the suggested model's performance. To train classifiers like SVM and ELM, a ten percent KDD training dataset with a high number of examples is used. Because using the complete dataset would cause various difficulties, just 10% of the KDD dataset was used. Protocol, service, and flag are all symbolic properties that may be altered or deleted. Finally, the cases are categorised into four groups: normal, dos, probe, and R2L. They used the Dataset to train SVM and ELM. They employed a multi-level model with a rectified KDD dataset for the testing procedure. Using the KDD Cup 1999 dataset, the suggested model's accuracy reached 95.75 percent and the false alarm rate was 1.87 percent.

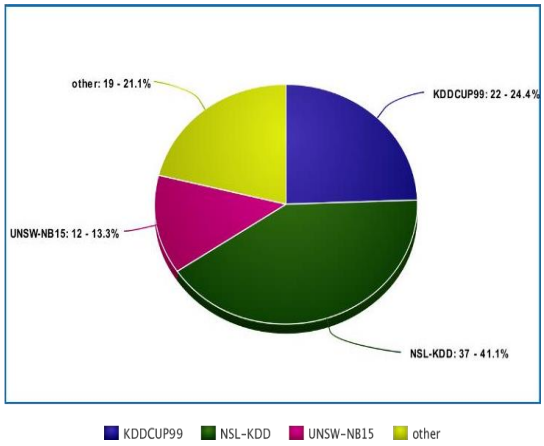
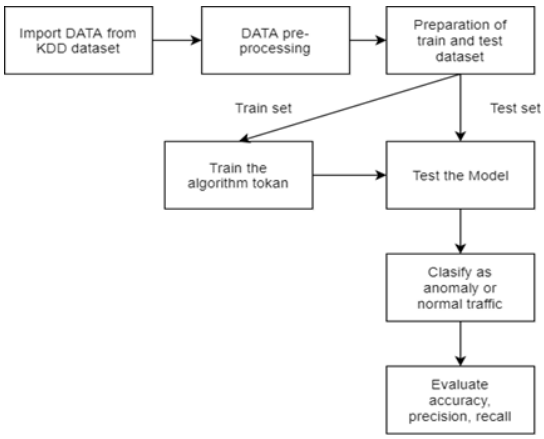


Figure 5: Research focused area of dataset in IDS from 2016- 2020

VI. Design Work



No	Feature Description	Data Type
1	TCP Packets	Integer
2	TCP Source Packets	Integer
3	TCP Fin Flag	Integer
4	TCP Destination Packets	Integer
5	TCP syn flag	Integer
6	TCP urget flag	Integer
7	UDP Packets	Integer
8	UDP Source Packets	Integer
9	UDP Destination port	Integer
10	ICMP Packets	Integer

Figure 6. System Design Architecture

Upload Dataset:

Collect suspicious traffic and ancillary data that defines or characterises it, identifying different network connections or relationships (connectionless traffic), and providing enough detail to aid criminal investigations and prosecutions. Detect incursions that are particular to a protected area. Service overloads, broadcast storms, and message floods are all examples of denial of service attacks.

Preprocessing:

Perform data reduction, ideally at the source, to lessen data load. Refine raw data to eliminate redundancy and false alarms. To create reports for following up on and taking corrective action on suspicious events or discovered vulnerabilities not getting immediate attention. Ability to open, track, and document the resolution of an intrusion event or vulnerability detected. An operator can utilise a site profile database to keep track of site-specific activity.

Feature Selection/Extraction:

A classification task normally requires training and testing data that contains a variety of data examples. The training set encompasses one "target value" (class labels) and multiple "attributes" for each instance (features). SVM's purpose is to create a model that predicts the target value of data instances in the testing set given just the characteristics. The following data is tested and shown on the output: The Attacker Profile output gives information on an attacker, whether an invader or a prober, an outsider or an insider. The Security Profile output gives you extensive security information about a network domain you've chosen. The System Profile shows which areas—addresses, components, and systems—are affected.

Table 2: Features In Preprocessing data

VII. Algorithm

A. Support Vector Machine:

Support Vector Machine (SVM) is a supervised learning approach that involves training numerous sorts of data from diverse disciplines. SVM generates a hyperplane or numerous hyperplanes in a high-dimensional space. A best hyperplane is one that divides the provided data into several classes with the principal division in the most efficient way possible. A non-linear classifier uses multiple kernel functions to evaluate the margins between hyperplanes. The basic goal of kernel functions like linear, polynomial radial basis, and sigmoid is to maximise margins between hyperplanes. Developers and researchers have built renowned applications as a result of the increased interest in SVMs. In image processing and pattern recognition applications, SVM plays a crucial role. A classification job usually entails splitting data into two sets: training datasets and testing datasets. Labels will be stated to as "target variables," and attributes will be referred to as "features" or "observed variables" in that class.

A. Naive Bayes:

Bayesian classifiers are statistical classifiers. They are proficient of calculating the likelihood that a given model would fit into a specific class. Its basis is Bayes' theorem. It is based on the assumption that the attribute value for a particular class is independent of the attribute values. Class conditional independence is the name given to this theory.

B. Naive Bayes

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)\ P(H)}{P(\mathbf{X})}$$

Bayesian classifiers are statistical classifiers. They are proficient of calculating the likelihood that a given model would fit into a specific class. Its basis is Bayes' theorem. It is based on the assumption that the attribute value for a particular class is independent of the attribute values. Class conditional independence is the name given to this theory.

VIII. Graphical Analysis

The performance of model is compared with single SVM and Random Forest classifiers. KDD1999 dataset is used for testing & training data. All the performance metrics are higher than the existing. The training and testing periods of the CWS IDS are minor than single SVM. Thus, the model is proficient compared to individual SVM. The performance metrics that are evaluated in the detection evaluation is equated with the SVM, random forest, and CWS IDS for training data and test data. Figure 7 and 8 represents the performance metrics after the calculation on training dataset and test dataset respectively.

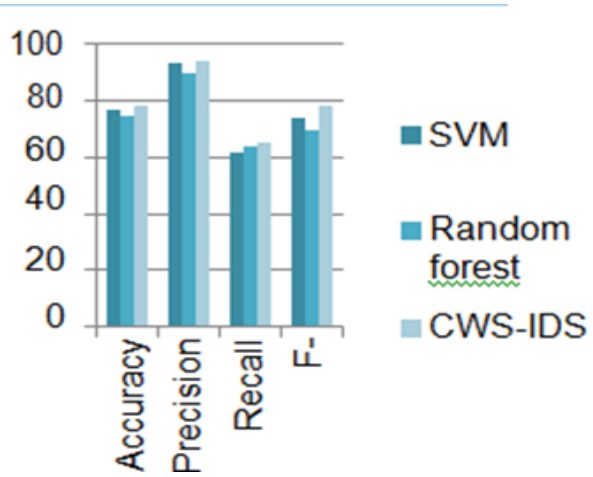


Figure. 7: Comparison of performance metrics on training dataset

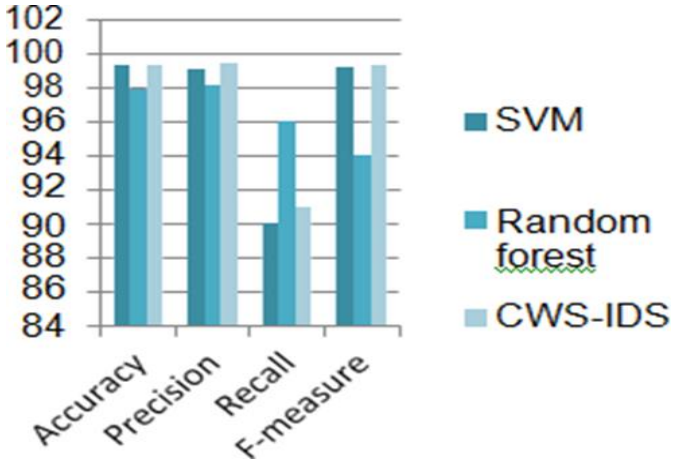


Figure 8: Comparison of Performance metrics on test dataset

IX. Conclusion

In this research, we focus on the use of machine learning methods and their applications to identify intrusion detection systems. The following are the four goals of this study review paper: I make recommendations for researchers who are new to the machine learning field and want to contribute to it; ii) present a state-of-the-art overview of machine learning; iii) provides further research directions required into intrusion detection system using machine learning

X. Future Work

Future work will cope with vast volumes of data, and a hybrid multilevel model will be built to improve accuracy. It is concerned with developing a simplified model backed by well-organized classifiers capable of categorizing new attacks with improved performance.

XI. REFERENCES

[1]. Shetty Akshada, Jadhav Shailaja et. al., "Detection of fake accounts in online social networks (OSN)" International Journal of Modern Trends in Engineering and Science-IJMTES 2017, Volume 4 -Issue 5 Pages 1-3.

[2]. Shailaja Jadhav, Minal Pokale, "Detecting attacks for security of information using Data Mining Technique" IJMTES 2017, Volume 4 Issue 05C. Chang and C. J. Lin, LIBSVM, "A Library for Support Vector Machines", the use of LIBSVM, 2009.

[3]. Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set and Support Vector Machine for Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, 2009.

[4]. Phurivit Sangkatsanee, Naruemon Wattanapongsakorn and Chalermpol Charnsripinyo, "Real-time Intrusion Detection and Classification", IEEE network, 2009.

[5]. Liberios Vokorokos, Alzebeta Kleniova, "Network Security on the Intrusion Detection System Level", IEEE network, 2004.

[6]. Thomas Heyman, Bart De Win, Christophe Huygens, and Wouter Joosen, "Improving Intrusion Detection through Alert Verification", IEEE Transaction on Dependable and Secure Computing, 2004.

[7]. T. Lin and C.J. Lin, "A study on sigmoid kernels for SVM and the training of non- PSD kernels by SMO-type methods", Technical report, Department of Computer Science, National Taiwan University, 2003.

Cite this article as :

Shailaja Jadhav, Vinaya Bhalerao, Varsha Yadav, Snehal Kamble, Bhavana Shinde, "Network Intrusion Detection System Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 1, pp. 74-81, January-February 2022. Available at doi : https://doi.org/10.32628/CSEIT22819
Journal URL : https://ijsrcseit.com/CSEIT22819