A RESEARCH ON INTRUSION DETECTION SYSTEM USING MACHINE LEARNING APPROACH USING NSL KDD-DATASET

[1] SHWETHA T P, [2] MEENAKSHI R

LECTURER
Department of Computer science and Engineering,
Government Plytechnic College Bellary,Ballari,India

ABSTRACT : Intrusion detection systems (IDSs) are currently drawing a great amount of interest as a key part of system defence. IDSs collect network traffic information from some point on the network or computer system and then use this information to secure the network. Recently, machine learning methodologies are playing an important role in detecting network intrusions (or attacks), which further helps the network administrator to take precautionary measures for preventing intrusions. In this paper, we propose to use ten machine learning approaches that include Decision Tree (J48), Bayesian Belief Network, Hybrid Naïve Bayes with Decision Tree, Rotation Forest, Hybrid J48 with Lazy Locally weighted learning, Discriminative multinomial Naïve Bayes, Combining random Forest with Naïve Bayes and finally ensemble of classifiers using J48 and NB with AdaBoost (AB) to detect network intrusions efficiently. We use NSL-KDD dataset, a variant of widely used KDDCup 1999 intrusion detection benchmark dataset, for evaluating our proposed machine learning approaches for network intrusion detection. Finally, Experimental results with 5-class classification are demonstrated that include: Detection rate, false positive rate, and average cost for misclassification. These are used to aid a better understanding for the researchers in the domain of network intrusion detection.

Key Words— Intrusion detection, Machine Learning, Cost Matrix.

1. INTRODUCTION

In this modern era, the internet has been playing an essen- tial role in everyone's daily life because it provides useful information on a wide range of topics, including business, education, and entertainment. Network attacks are also onthe upsurge because of the widespread use of the internet. Intrusion detection systems and firewalls are just a few of the methods that have been proposed to counter these attacks. Firewall filters all incoming and outgoing packets based on predefined rules, while IDS just examines the network and delivers an alert message to the network administrator if any harmful activities are detected [1]. When compared to fire- walls, intrusion detection system is more secure and performs better [2].

To determine whether an intrusion attack has occurred or not, IDS depends on few approaches. First is signature-based approach, where known intrusion attack signature is stored in the IDS database to match with current system data. When theIDS finds a

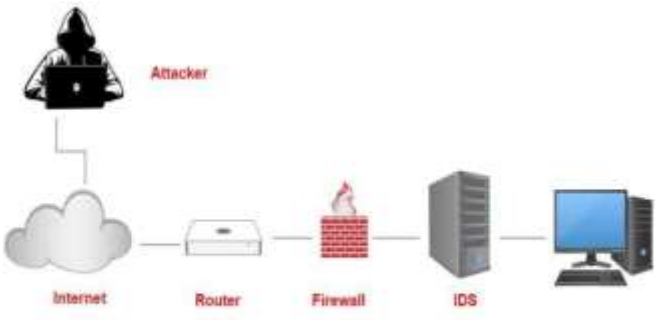
match, it will recognise it as an intrusion. This approach provides a fast and accurate detection. However, the drawback of this is to have periodic update of the signature database. In addition, the system could be compromised before the newest intrusion attack can be updated.

The second approach is anomaly-based, or behaviour-based, where IDS will determines an attack when the system operates out of the norm. This approach can detect both known and unknown attacks. However, the drawback of this approach is low accuracy with high false alarm rate.

Lastly, hybrid-based approach uses both signature-based and anomaly-based approaches. This approach uses signature- based approach to detect known attacks, and anomaly-based approach to detect unknown attacks. Combining both approaches can ensure a more effective detection, but may increase computational cost.

Machine Learning (ML) uses statistical modeling approach to learn past data pattern, and then predicts the most likely outcome using new data. Therefore, ML algorithm has been applied to IDS using anomaly-based approach. As stated above, the challenge here is to build a model that can give high accuracy with low false alarm rate.

Therefore, this study aims to analyse recent researches in IDS using ML approach; with specific interest in dataset, ML algorithms and metric. Dataset selection is very important to ensure model build is suitable for IDS use. In addition, dataset structure can affect effectiveness of ML algorithm. Thus, ML



Intrusion Detection System

algorithm selection is dependent on the structure of the selected dataset. After that, metric will provide a quantitative evaluation of ML algorithms towards specific dataset.

1.1 MACHINE LEARNING

ML algorithm can be categorized into 11 categories. This is shown in Fig. 1. Bayesian category uses Bayes Theorem of probability, which determines the probability of specific outcome to come true. The most popular algorithm in this category is Naïve Bayes. Decision tree has a tree like structure that starts from root nodes, which is the best predictor. Then progresses through its branches until it reach a leave node. This is the decision outcome.

Fig. 1 Category of ML algorithms adopted

Dimensional reduction is to find features that are important to the outcome. This will removes irrelevant and redundant features. It is mostly performed during the pre-processing phase. The most popular algorithm is Principal Component Analysis (PCA)

Instance-based is also known as memory-based learning. This category of algorithm finds the most similar instances, or training data, that matches the new data to make prediction. The most popular algorithm in this category is k-Nearest Neighbour (kNN).

Clustering is grouping of data points that are close together to form its own group. This category of algorithm works well in unsupervised learning approach, which do not require labelled data. The most popular algorithm in this category is k-Means.

Regression algorithm try to build model that can represent the relationship between variables. It is derived from statistical analysis. The most popular algorithm in this category is Logistic Regression.

Neural network is inspired by the brain cell called neuron that forms the biological neural network. This category finds patterns from the data to make its prediction. Normally it would require large amount of data to produce a good prediction. The most popular algorithm in this category is Perceptron.

Ensemble is a method of combining the result of several algorithms before producing the final outcome. There are typically 2 methods, bagging and boosting.

1.2 NETWORK ATTACKS AND THEIR TYPES

Network attacks are defined as an attempt to gain or perform an unauthorized action to an organizational network with the goal of looting data or carrying out other harmful activities. Network attacks are divided into two categories: passive attack and active attack . During passive attack, the attackers intercept the network and monitor or obtain confidential details without altering it. Release of message contents and analysis of traffic are examples of passive attacks. In active attack, the attackers get illegal access and further modify, delete, encrypt and decrypt the data. Active attacks include message modification, repudiation, service denial, replay, and masquerade. It is very hard to detect passive attacks since they have no effect on the data or device. IDS perform a significant part in detecting various forms of attacks. Any attack, whether passive or active or any one of the attacks which fall in the following categories can be considered.

- Denial of Service (DOS): Here, the network is filled with unusable traffic by intruders such that the resources are kept busy and users are prevented from using the network. Land, Back, and Mail Blood Smurf attacks are examples of DOS attack.

- Probe attack: It makes use of a software/program to monitor or collect information about the network activity. Satan, Ipsweep, Mscan, Saint, and Nmap are examples of these attacks.

- Remote to Local (R2L): Here, an intruder can transmit packets via certain devices but does not have access to the device's authorized account. In this situation, the intruder often exploits any weakness to get access to the device as a user. Named, Phf, Sendmail, and Guest are the examples of this type of attack.

- User to Root (U2R): An attacker has gained access to the user and is attempting to get superuser benefits. Perl, Ps, Eject, and Ffbconfig are examples of this class.

Attack Category	Probing Attacks	DoS Attacks	U2R Attacks	R2L Attacks
Known Attacks	ipsweep, satan, nmap, portsweep	Teardrop, pod, land, back, Neptune, smurf	Perl, loadmodule, rootkit, buffer_overflow	ftp_write, phf, guess_passwd, warezmaster, warezclient, imap, spy, multihop
Novel Attacks	saint, mscan	mailbomb, udpstorm, apache2, processtable	Xterm, ps, sqlattack, httptunnel	Named, snmpguess, worm, snmpgetattack, xsnoop, xlock, sendmail

Table 1: Known and novel attack types

2. *DATASET*

Dataset is the key component to train machine learning to detect anomaly threats. However, the analysis from this study shows that many researchers are still relying on an outdated dataset, KDDCup99 and NSL-KDD (a variant of KDD00 dataset), which have been criticized by many as outdated and not relevant in current network infrastructure. This dataset was produced in 1999, which is almost 20 years old. Rapid development and changes in Information Technology such as cloud computing, social media and Internet of Things are changing the landscape of network infrastructure. These changes have the driving force in changing threat attack itself. Therefore, many research results that demonstrate high accuracy is being viewed as overstated, because the dataset being used does not represent the current threat or infrastructure.

The KDDCup99 dataset is a popular dataset and has been used for the Third International Knowledge Discovery and Data Mining Tools Competition. Each connection instance is described by 41 attributes (38 continuous or discrete numerical attributes and 3 symbolic attributes). Each instance is labelled as either normal or a specific type of attack. These attacks fall under one of the four categories: Probe, DoS, U2R, and R2L .

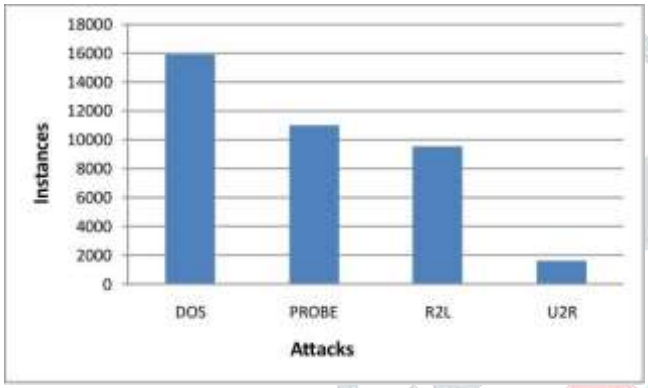


Figure 2:Instances in Training Dataset

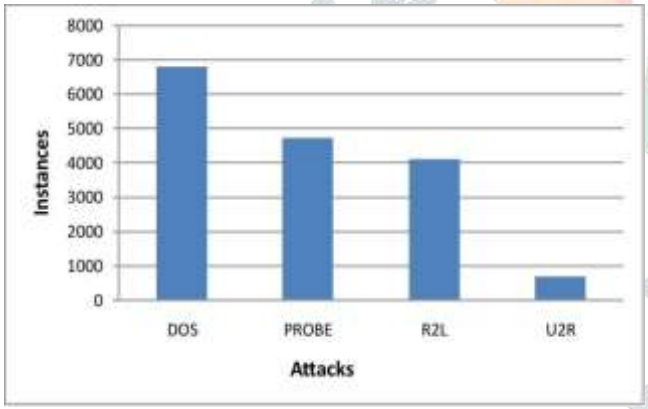


Figure 3. Instances in Testing Dataset

Table II

DESCRIPTION OF THE NSL KDD DATA SET

S. NO.	***Feature***	**Definition**
1.	Duration	Connection's length (in sec)
2.	Src-bytes	The amount of data(bytes) transmitted from source to destination
3.	Dst-bytes	The amount of data(bytes) transmitted from destination to source
4.	Land	Land=1 if the connection belongs to the same host,

		otherwise 0
5.	Wrong-fragment	Total no. of wrong fragments
6.	Urgent	Total no. of urgent messages
7.	Hot	Total no. of hot symbols
8.	Num-failed-logins	The total no. of unsuccessful login attempts
9.	Logged-in	Logged-in=1 if user is logged in, otherwise logged-in=0
10.	Num-compromised	The no. of conditions that have been compromised
11.	Root-shell	Root-shell=1 if root shell is generated, otherwise 0
12.	Su-attempted	Su-attempted=1 If su root attempted, otherwise 0
13.	Num-root	Total no. of connected roots
14.	Num-file-creations	The total no. of file created
15.	Num-shells	Total no. of shell prompt
16.	Num-access-files	Total no. of operations performed on access files
17.	Num-outbound-cmds	Total no. of outgoing commands
18.	Is-host-login	Is-host-login=1 If host is login, otherwise 0
19.	Is-guest-login	Is-guest-login=1 If guest is login, otherwise 0
20.	Count No.	Total no. of connections made to the same host in the last two sec
21.	Srv-count.	Total no. of connections made to the same service in the last two sec
22.	Serror-rate	Proportion of connections with a syn error
23.	Srv-serror-rate	Proportion of connections with a syn error
24.	Rerror-rate	Proportion of connections with a rej error
25.	Srv-rerror-rate	Proportion of connections with a rej error
26.	Same-srv-rate	Proportion of connections to the same service
27.	Diff-srv-rate	Proportion of connections to the different service
28.	Srv-diff-host-rate	Proportion of connections to the different hosts
29.	Dst-host-count	The total no. of connections to the same destination host
30.	Dst-host-srv-count	The total no. of connections to the same destination host and service
31.	Dst-host-same-srv-rate	Proportion of connections that have the same destination host and service
32.	Dst-host-diff-srv-rate	Proportion of connections on the current host that use a different service
33.	Dst-host-same-src-port-rate	Proportion of current host connections with the same source port
34.	Dst-host-srv-diff-host-rate	Proportion of connections of same service and different hosts
35.	Dst-host-serror-rate	Proportion of current host's connections with serror
36.	Dst-host-srv-serror-rate	Proportion of serror connections on the current host of a service
37.	Dst-host-rerror-rate	Proportion of current host connections with an rst error
38.	Dst-host-srv-rerror-rate	Proportion of current host of service connections with rst error
39.	Protocol-type	Protocol type, tcp, udp, etc.
40.	Service	Type of network

41.	Flag	Status of flag
42.	xAttack	Attack type

NSL KDD dataset description is given in table2. DOS has maximum instances in the training as well as in the testing data set whereas U2R has the least number of instances in both training and testing sets.

The NSL-KDD dataset was developed in 2009, but it is actually an improved version of the KDDCup99 dataset. NSL- KDD tries to improve KDDCup99 dataset by removing redundant records, including the imbalanced number of instances and the variety of attack classes . However, it still inherited the fundamental limitation of the dataset.

KDDCup99 has many drawbacks. Firstly, this dataset was developed in 1999 using a Solaris-based operating system to collect a wide range of data due to its easy deployment. However, there are significant differences in today's operating systems which barely resemble Solaris. In this age of Ubuntu, Windows and MAC, Solaris has almost no market share.

Secondly, the traffic collector used in KDD datasets, TCPdump, is very likely to become overloaded and droppackets from a heavy traffic load. More importantly, there is some confusion about the attack distributions of these datasets. According to an attack analysis, Probe is not an attack unless the number of iterations exceeds a specific threshold, while label inconsistency has been reported .

Thirdly, the emergence of new technologies such as cloud computing, social media and the Internet of Things has changed the network infrastructure drastically. These changes will also result in new types of threat.

The other two popular datasets are ISCX 2012 and UNSW- NB15. ISCX 2012 is a dataset created by Information Security Centre of Excellence (ISCX) at University of New Brunswick in 2012. This dataset consists of seven days of data withlabelling of normal (one) or attack (two). The dataset has no classification of the types of attack, thus it will only provide binary classification. However, this dataset is no longer available. This is because the centre has created a new dataset, called CICIDS2017 . The centre has also changed its name to Canadian Institute for Cybersecurity (CIC). Unfortunately, no article was found using this new dataset at the time of this study.

Another popular dataset is UNSW-NB15, this dataset was created by Australia Centre for Cyber Security (ACCS) using IXIA PerfectStorm to generate nine types of attack. These nine types of attack are namely fuzzers, analysis, backdoors, DoS, exploits, generic, reconnaissance, shellcode, and worms. The dataset has a total of 47 features with two labels. First is named as ‘Label’, where zero indicates normal and one indicates an attack. Second label is named as ‘attack_cat’, which provides the type of attack .

3.Introduction of Implemented Machine Learning Algorithms

Machine Learning algorithms are the advancement of conventional algorithms . Such algorithms allow systems to automatically learn themselves from data and make them smarter. Because of their learning and classification skills,these algorithms are currently employed in practically every industry to handle a wide range of issues. These algorithmsare primarily categorized into supervised or unsupervised. In the section, we will go over several relevant machine learning approaches for detecting and classifying network attacks using IDS.

3.1 Supervised Learning: In supervised learning , the data are split into two sets, one is the training set and the other is the testing set.Training set data are used to train the model and the testing set are used for input in that model as shown in figure 5. Some of the supervised learning algorithms are: Naive Bayes, Logistic Regression and Support Vector Machine. The supervised model has divided into two categories. The first one is the classification in which the output variable is categorical data. The second one is the regression in which the output class is a real value. The advantage of supervised learning is that it helps to solve the various types of real-world problems.

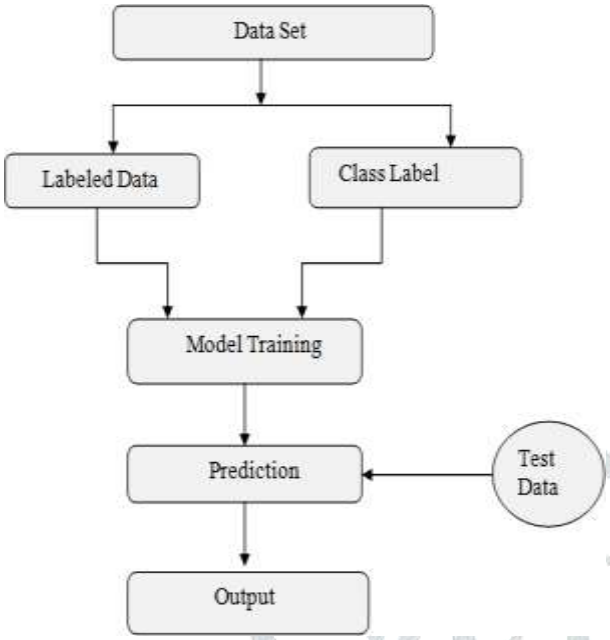


Figure 2. Architecture of Supervised Learning

- Naive Bayes

A group of supervised learning algorithms built on the Bayes theorem is referred to as the Naive Bayes [25]. It aids in the resolution of classification-related problems. It is typically used in text classification tasks where the training data sets are of high dimension. Being a probabilistic classifier, predictions are done based on the probability of the object. The Bayes theorem is discussed in equation 2.

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

where A and B are the events. P(A) and P(B) are the independent probability. P(A/B) is the probability of A given that B is true. P(B/A) is the probability of B given that A is true. It performs well in multi-class classification as compared to binary classification. One of the disadvantages is that all the relationships between features are not learned as the algorithm makes an assumption that all features are unrelated or independent.

- Support Vector Machine

Support Vector Machine (SVM) is a form of supervised learning technique. It is used for both regression and classification purposes however, most of the time it is used in classification problems. It is mostly used for two group classification problems due to its excellent accuracy and capacity to analyze high dimensional data. Support Vector Machine is a fast and dependable classification algorithm that performs very well with a limited amount of data contains data with similar features as shown in figure 3. The Support Vector Machine is basically of two types: Linear SVM and Nonlinear SVM. Linear SVM is mainly used for data that are linearly separable i.e a

straight line can divide the data set into two classes. When there is a nonlinearly separable data in that case we use nonlinear SVM.

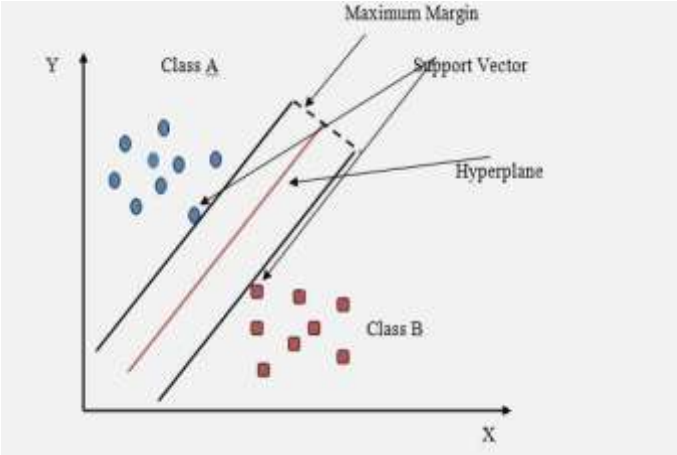


Figure 3. Architecture of Support Vector Machine

The SVM algorithm's objective is to find a hyperplane which can separate the data set into a certain number of groups that and Nonlinear SVM. Linear SVM is mainly used for data that are linearly separable i.e a straight line can divide the data set into two classes. When there is a nonlinearly separable datain that case we use nonlinear SVM.

- Logistic Regression

It is used in solving classification problems. Logistic Regression evaluates the relationship between the dependent and independent variables.

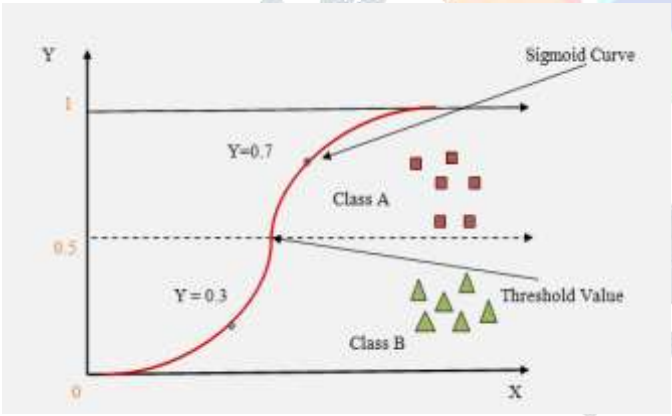
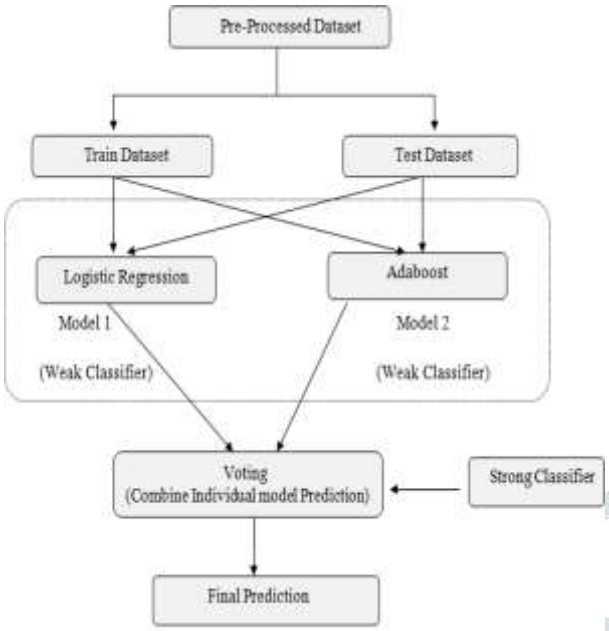


Figure 4. Logistic Regression Architecture

LR works for both binary and multi-class classification but LR performs better on binary class classification. The probability of an event occurring is anticipated by fitting the data to the logistic function. The logistic function selects values between 0 and 1. If the value is greater or equal to 0.5, it is labeled as 1, else it is labeled as 0 as shown in figure 4. The advantages of logistic regression are that it performs better when the data are linearly separable and it is less prone to the over-fitting problems. The major disadvantages of the algorithm are that nonlinear problems can't be solved using logistic regression and if there is a high dimensional dataset then there is a chance of over-fitting.

- Proposed Model

The Proposed model is an ensemble technique in which the Adaboost is combined with Logistic Regression. Adaboost is a machine learning technique developed to improve classification efficiency . The basic workingidea of boosting algorithm is as follows: data are initially



.Figure 5. Architecture of Proposed Model

divided into groups using draft rules. Every time the algorithm is executed, additional rules are added to this preliminary set of rules. In this manner, misclassification is reduced. In this approach, all the weak classifiers combine to create a strong classifier capable of detecting different types of attacks. The main advantage of the Adaboost approach is that net classification error is evaluated in each learning step. The architecture of the proposed model is explained in figure 5.

3.2 Unsupervised Learning*:* Unsupervised learning is a type of machine learning technique in which models are not supervised by training data sets , Here, without any prior training of data, the machine's objective is to categorize the unsorted data according to similarities and patterns as demonstrated in figure 9. Some of the unsupervised learning algorithms are DBSCAN, K-Means, and Genetic K-Meansclustering.

- DBSCAN

DBSCAN stands for Density-Based Spatial Clusteringof Applications with Noise. DBSCAN is a member of the unsupervised machine learning algorithm. DBSCAN [29] is a density-based clustering and it forms the cluster based on the density. It can find clusters of various shapes and sizes from huge quantity of data that include noise and outliers. The architecture of DBSCAN is illustrated in figure 10

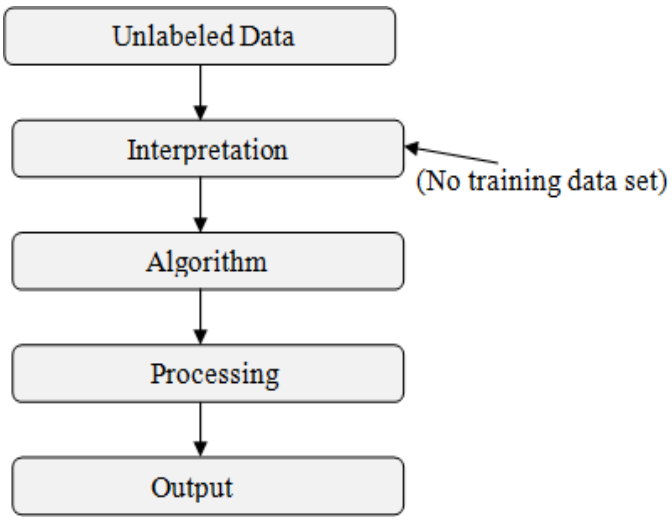


Figure 6. Architecture of Proposed Model

divided into groups using draft rules. Every time the algorithm is executed, additional rules are added to this preliminary set of rules. In this manner, misclassification is reduced. In this approach, all the weak classifiers combine to create a strong classifier capable of detecting different types of attacks. The main advantage of the Adaboost approach is that net classification error is evaluated in each learning step. The architecture of the proposed model is explained in figure 6.

Unsupervised Learning: Unsupervised learning is atype of machine learning technique in which models are not supervised by training data sets [28], Here, without any prior training of data, the machine's objective is to categorize the unsorted data according to similarities and patterns as demonstrated in figure 9. Some of the unsupervised learning algorithms are DBSCAN, K-Means, and Genetic K-Meansclustering.

- DBSCAN

DBSCAN stands for Density-Based Spatial Clusteringof Applications with Noise. DBSCAN is a member of the unsupervised machine learning algorithm. DBSCAN

[29] is a density-based clustering and it forms the cluster based on the density. It can find clusters of various shapes and sizes from huge quantity of data that include noise and outliers.

- K-MEANS CLUSTERING

K-Means clustering [16] is one of the simplest and most popular unsupervised machine learning algorithm. The K-Means algorithm identifies the K number of centroids. The centroid concept is used to cluster the data points. After every iteration, the centroid value is evaluated using the averaging concept. The objective of the algorithm is to minimize the sum of distances of data points from their respective clusters. The method takes unlabeled data as input, separate it into k number of clusters and performs the same procedure till the optimal cluster is discovered. The main advantage of K-Meansis that if the data sets are distinct, then it gives the best results. The main disadvantage of the algorithm is that it needs prior specification for the number of clusters and sometimes choosing the centroid randomly cannot give fruitful results

- IGKM

Genetic K-Means (IGKM) is a method in which the number of clusters is not known in advance. Genetic Algorithm (GA) is used to determine the optimal value K. The fitness function (evaluating function) minimizes the amount of clusters while maximizes the separation and effectiveness as much as possible.

4 Performance Metrics

The following performance measures are used to measure and compare the effectiveness of various IDS based on ma- chine learning .

- True Positive (TP) - Here, an attack is identified and confirmed to be an attack. This sort of circumstance is classified as a true positive.

- False Positive (FP) – Here, an attack is detected but it is not actually an attack. A false positive is therefore only a false warning.

- True Negative (TN) - Data that are appropriately classified as normal and is normal. This sort of circumstance is classified as a true negative.

- False Negative (FN) - Attack data that has been erroneously classified as normal. This is the most vulnerable stage since there is no information of the attack that has been already occurred.

- The sum of the TP and TN observations to the total num-ber of observed values is known as accuracy. Accuracy typically determines the total number of classifications that are valid. The formula of accuracy is explained in equation 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision is the ratio of true positive observation to the summation of true and false positive observations as shown in equation 4.

$$Precision = \frac{TP}{TP + FP}$$

Recall calculates the number of valid classifications pe- nalized by number of missing entries. The formula of recall is discussed in equation 5.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score is a combination of precision and recall as shown in equation 6. A good F1 score means that there are lesser false positives and negatives. Its value lieswithin 0 and 1. An F1 score of 1 is depicted as perfect while an F1 score of 0 is a failure.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.1 Experiment Results

Weka is open-source software that offers tools for pre- processing of data, execution of various machine learning algorithms, and visualization tools, allowing us in building machine learning algorithms and help to apply in real-lifescenario. It is written in Java and runs on almost any platform [30].

Table III
EXPERIMENT RESULTS OF UNSUPERVISED LEARNING MODELS

Unsupervised Learning					
Algorithm	Attack	Accuracy	Precision	Recall	F-Score
DBSCAN	DOS	94.18%	96.7%	88.6%	92.5%
	Probe	79.51%	38.8%	79.4%	52.2%
	R2L	80.91%	70%	99.9%	13.1%
	U2R	79.4%	26.2%	87.5%	40.3%
K-Means	DOS	94.54%	97.1%	89.1%	92.9%
	Probe	79.05%	38.9%	85.5%	53.5%
	R2L	79.83%	54%	99.9%	10.3%
	U2R	78.47%	26.2%	94.6%	41.1%
IGKM	DOS	82.7%	97.2%	95%	96.1%
	Probe	54.77%	95.7%	99.9%	97.8%
	R2L	31.58%	73%	99.9%	13.7%
	U2R	29.60%	64.3%	99.9%	78.3%

The hardware specification is as follows: Intel i5 10 generation, 1.19 GHz machine with 8GB of Random Access Memory (RAM) and 512GB of Read-only Memory (ROM). In this experiment, after the pre-processing phase, The public NSL KDD data set are split into 70-30 ratio, 70% for training the model and 30% for testing the model. The target class of the dataset is attack and the IDS identifies four different types of attacks i.e, DOS, Probe, R2L and U2R. The well-known algorithms of supervised and unsupervised learning are applied on the pre-processed dataset. In supervised learning, we have used Support Vector Machine, Naive Bayes and Logistic Regression. The proposed model also falls in the category of supervised learning. In unsupervised learning, we have employed DBSCAN, K-Means and Genetic K-Means clustering. The performances of different unsupervised learning algorithms are discussed in table 3. while the performance of different supervised learning algorithms are shown in table 4. Out of all these supervised and unsupervised learning algorithms, the proposed/ensemble model performs better than any other algorithms either it is supervised or unsupervised. The ensemble model obtained an accuracy equal to 99.91%, 99.60%, 99.90% and 98.15% on DOS, Probe, R2L and U2R Respectively.

Table IV

EXPERIMENT RESULTS OF SUPERVISED LEARNING MODELS

Supervised Learning					
Algorithm	Attack	Accuracy	Precision	Recall	F-Score
SVM	DOS	91.18%	87%	91.8%	89.3%
	Probe	81.33%	42.7%	95.2%	58.9%
	R2L	85.57%	63%	64.4%	11.4%
	U2R	84.98%	34.2%	96.4%	50.5%
NB	DOS	93.9%	89.8%	95.8%	92.7%
	Probe	93.40%	69.2%	95.8%	80.4%
	R2L	97.8%	38.1%	76.3%	50.8%
	U2R	86.1%	35.8%	94.6%	52%
LR	DOS	98.4%	98.7%	97.4%	98%
	Probe	98.19%	93.8%	93.4%	93.6%
	R2L	98.95%	75%	40.7%	52.7%
	U2R	97.8%	90.2%	82.1%	86%
Proposed Model	DOS	99.91%	99.80%	99.99%	99.99%
	Probe	99.6%	99.4%	98.2%	98.8%
	R2L	99.90%	98.2%	94.9%	96.6%
	U2R	98.15%	87.7%	89.3%	88.50%

4.2 Performance Analysis

In this section, we have compared the results obtained by our proposed model with the results obtained by previously proposed models. Table 5 presents the results that we obtained and compares them with the results of B. Selvakumar [19]. It is found that the Proposed model performs better than the Decision tree (C4.5) and Bayesian Network (BN).

5. CONCLUSION AND FUTURE WORK

The paper initially provided a background on the intrusion detection system and its importance in the cyber security space. The NSL-KDD dataset were analyzed and pre-processed using the chi-square test, which reduces the number of features from the dataset and avoids the over- fitting problem. Supervised and unsupervised machine learning algorithms are applied to the pre-processed dataset. When the performance of all the algorithms are compared then it is found that the ensemble model outperforms all other models. In the future, we will conduct an extensive study of ML algorithms to provide a better solution for the IDS by taking a real-time dataset.

REFERENCES

[1] Suman Thapa and Akalanka Mailewa. The role of intrusion detec- tion/prevention systems in modern computer networks: A review. In *Conference: Midwest Instruction and Computing Symposium (MICS)*, volume 53, pages 1–14, 2020.

[2] Tejvir Kaur, Vimmi Malhotra, and Dheerendra Singh. Comparison of network security tools-firewall, intrusion detection system and honeypot. *Int. J. Enhanced Res. Sci. Technol. Eng*, 200204, 2014.

[3] Ashwini Pathak and Sakshi Pathak. Study on decision tree and knn algorithm for intrusion detection system.

[4] Asmaa Shaker Ashoor and Sharad Gore. Importance of intrusion detec- tion system (ids). *International Journal of Scientific and Engineering Research*, 2(1):1–4, 2011.

[5] L Haripriya and MA Jabbar. Role of machine learning in intrusion detec- tion system. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 925–929. IEEE, 2018.

[6] Marcel Jung, Octavian Niculita, and Zakwan Skaf. Comparison of different classification algorithms for fault detection and fault isolation in complex systems. *Procedia Manufacturing*, 19:111–118, 2018.

[7] MA Jabbar, Rajanikanth Aluvalu, et al. Rfaode: A novel ensemble intrusion detection system. *Procedia computer science*, 115:226–234, 2017.

[8] C Kalimuthan and J Arokia Renjit. Review on intrusion detection using feature selection with machine learning techniques. *Materials Today: Proceedings*, 33:3794–3802, 2020.

[9] Xianwei Gao, Chun Shan, Changzhen Hu, Zequn Niu, and Zhen Liu. An adaptive ensemble machine learning model for intrusion detection. *IEEE Access*, 7:82512–82521, 2019.

[10] Manjula C Belavagi and Balachandra Muniyal. Performance evalua- tion of supervised machine learning algorithms for intrusion detection. *Procedia Computer Science*, 89(2016):117–123, 2016.

[11] Ali A. Ghorbani, Wei Lu and M. Tavallaee, Network intrusion detection and prevention: Concepts and Techniques, *Advances inInformation security*, Springer, 2010.

[12] A. O. Adetunmbi, S.O. Falaki, O. S. Adewale, and B. K. Alese, Network Intrusion Detection based on rough set and k-nearest neighbour, *Intl. Journal of computing and ICT research*, 2(1) (2008), 60-66.

[13] C. Elkan, Results of the KDD'99 classifier learning. SIGKDD Explorations, 1(2) (2000), 63-64.

[14] C. Krugel and T. Toth, Using decision tree to improve signature based intrusion detection, in: *Proceedings of RAID*, 2003, G. Vigna, E. Jonsson, and C. Kruegel, eds, Lecture Notes in Computer Science, Vol. 2820, 173-191.

[15] D. E. Denning and P. G. Neumann, Audit trail analysis and usage data collection and processing, Technical report project 5910, SRIInternational,1985.

[16] D. E. Denning, An intrusion detection model, *IEEE Trans. On Software Engineering*., SE-13(2) (1987),118-131. IEEE Computer SocietyPress, USA.

[17] D. Barbara, J. Couto, S. Jajodia, L. Popyack, and N. Wu, ADAM: Detecting intrusions by data mining, in : *Proceedings of 2nd Annual IEEE workshop on Infor. Assu. Secur.*, Jun 2001, New York , 11-16.

[18] G. Wang, J. Hao, J. Ma and L. Huang, A new approach to intrusion detection using artificial neural networks and fuzzy clustering, *Expert system with applications*, 37 (2010), 6225-6232, Elsevier.

[19] H. Debar and B. Dorizzi, A neural network component for an intrusion detection system, in: *Proceedings of the IEEE Computer Society. Symposium on research in security and privacy*, Oakland, CA, May 1992, 240-250.

[20] H. A. Nguyen and D. Choi, Application of data mining to network intrusion detection : classifier selection model, in: *Proceedings of Challenges for Next Generation Network Operations and Service Management (APNMOS 2008)*, Y. Ma, D. Choi, and S. Ata eds, Lecture notes in computer science, Vol. 5297 (2008), 399-408, Springer.