iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** XI **Month of publication:** November 2026

DOI: https://doi.org/10.22214/ijraset.2026.85708

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

A Comparative Study of Decision Tree and Naïve Bayes Algorithms for Network Intrusion Detection

Richa Kirad[1], Diwakar Tripathi[2]
Department of Computer Science, S.S. Maniar College of Computer and Management, Nagpur, India

Abstract: ***A network intrusion is any unauthorized access to a computer network. For detecting a network intrusion, the defenders should have a clear understanding of how attacks work. In a network environment, intrusions possess a major security issue which can be an unauthorized activity on a computer network which is generally difficult to detect. Through this project we aim to monitor computer network for malicious activity using Intrusion Detection System(IDS). In this work, we aim to use Supervised Machine Learning Algorithms such as Naive Bayes Classifier and Decision Tree Classifiers, and compare their output and efficiency. The machine learning algorithms are used on a labelled dataset, which classifies the connections as good or bad. As a result, the accuracy of the classification result has to be maximized by maintaining low false-negative and low false-positive rates.***

Keywords: ***DoS, ICI, Decision Tree, Naïve Bayes, IDS***

I. INTRODUCTION

One of the important technology in business sector is Intrusion detection. Along with cybersecurity, it is considered as one of the most active and important area in research. Nowadays, a single computer or laptop is part of multi networked and distributed systems which are prone for intrusions, intrusions that are increasing in severity and number. For this, several protective measure have been put in place like firewalls to check the activities of intruders. However, this technique cannot guarantee the full protection of the system. Therefore, in order to monitor the networks for intrusions or attacks more efficiently, a more dynamic mechanism like IDS or Intrusion Detection System is used and considered as a second line of defence. This system will allow reporting signs of intrusions in order to take the necessary action.

A. Intrusion Detection System

A computer network operates not only by a small network such as a closed LAN such as a part of business systems, but also in connection with an external network in some cases. However, connecting to an external network means that it becomes side by side with unauthorized access. For this reason, firewalls, proxy servers, etc. are arranged in the connection path to the outside and centralized, and the communication is managed and controlled on that, measures to prevent illegal access and intrusion attempts are taken. However, it is not always possible to prevent all intrusions (such as spoofing) even if taking protective measures of them, fraudulent access that has invaded by penetrating the network of protection and penetrating, data corruption and destruction of the system, moreover it is impossible to remove concerns of receiving data theft. Therefore, using IDS we can detect signs of unauthorized access and notify the administrator. Administrators can be alert prior to actual initial damage, take few defensive measures such as line disconnection if possible, and prevent system destruction anyhow beforehand [4].

Basically there are two types “anomaly type” and "signature type" as a method of preventing and detecting intrusion. Firewalls generally handle information up to the transport layer, but the IDS checks all the data contained in the packet [6]. However, among firewall products in recent years, there are not many products that have functions close to IDS and intrusion prevention system (IPS). These systems constantly check the amount and type of communication, check the unevenness of the communication type which is impossible in ordinary work, the increase of the abrupt traffic volume, and further check the connection with other than the specific communication partner. (Anomaly type) In the signature type, we judge from the characteristics of packets considered to be attacked. In general, if there is an attempt of unauthorized access to exploit target vulnerabilities by various attack methods and trying to obtain access rights and administrative authority to access internal information, it can be detected as a preliminary phase of attack [2].

B. Algorithms Generally Used for Intrusion Detection System

1) *Support Vector Machines:* Support Vector Machine is a binary classification algorithm. Given a set of 2 types in N dimensional place, SVM generates a total of (N1) dimensional hyperplane to separate all those points into groups.

©IJRASET: All Rights are Reserved

2) *Naïve Bayes Classification:* Naïve Bayes Classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. The featured image is the equation-

$P(A|B) = P(B|A) \cdot P(A) / P(B)$

with P(A|B) is posterior probability, P(B|A) is likelihood, P(A) is class prior probability, and P(B) is predictor prior probability.

3) *Decision Tree:* In the field of machine learning, the decision tree is a predictive model, and from observation results a conclusion is derived. The Internal nodes are defined as variables, and branches to child nodes indicate possible values of that variable. A leaf represents a predicted value of a variable with respect to a variable value which is represented by route to destination.
4) *K-means Clustering:* The k-means method or k- means clustering is an algorithm of non-hierarchical clustering. It is named like this because it uses the average value of clusters and equally divides it into k number of clusters given. It is also called k-means method. In this algorithm, it is reduced to the minimum value dependent on the initial value, not the minimum value.
5) *Artificial Neural Networks:* Neural networks have been used for both intrusion prevention and intrusion detection. ANN is basically used to detect abnormal intrusions. It can also be used for learning typical characteristics of system admin users, and legal use of unauthorized IDS.

The neural network receives packets from the network stream and analyses information on instances of misuse.

C. Datasets

1) *NSL KDD:* NSL-KDD is a data set proposed to solve some of the inherent problems of the KDD '99 dataset. However, this new version of the KDD dataset has some issues. Complete representation of existing real networks cannot be done due to the lack of NIDS public data sets. Also It can be applied as a valid benchmark dataset that can help researchers differentiate various intrusion detection methods.
2) *DARPA:* Under the sponsorship of DARPA ITO and the Air Force Research Institute, the Cyber System Technology Group from the Lincoln Research Institute which developed an intrusion detection system for computer networks. Also evaluation first made by the Aeronautics Research Institute a formal, reproducible, statistically significant evaluation of the intrusion detection system. These evaluation efforts were carried out in 1998 and 1999.
3) *KDD CUP'99:* In 1998 the KDD Cup '99 dataset was created by processing the tcpdump part of the DARPA Intrusion Detection System (IDS) Evaluation Data Set. Lapoln Lab signed a contract with DARPA [Lippmann et al] and created it.
4) *ISCX:* ISCX 2012 data set depicts dynamically generated data reflecting network traffic Invasion. There are various multi-level attack scenarios when packets are captured. By running randomly synthetical user profiles, normal traffic is detected [25].
5) *ADFA:* The Australian Defence Force Academy Linux Data Set (ADFA - LD) and the Australian Defence Force Academy Windows Data Set (ADFA - WD) have created a new generation dataset system that includes a labelled system call trace for the safety of various applications preventing it from various modern attacks [25].
6) *KYOTO:* Kyoto 2006 + was built on actual traffic data for few years. It consists of 14 statistical features which is derived from the KDD'99 data set that can be used for further analysis and evaluation of IDS [25].

II. LITERATURE SURVEY

Table I. Literature Survey

Authors	Algorithms	Datasets	Attacks	Accuracy
Sandhya P., Ajith Abraham, Johnson T.	DT-SVM Model	KDD Cup 1999	Normal Probe, U2R, R2L, Dos Attack	98%
Mrutyunjaya Panda and Manas P.	Naïve Bayes	KDD Cup 1999	Normal Probe, U2R, R2L, Dos Attack	99%
Iwan S., Adam P. B. and Garry W.	Naïve Bayes, Nearest Neighbor and Rule Induction		Normal Probe, U2R, R2L, Dos Attack	
Srinivas M., Andrew S. and Ajith A.	Support Vector Machine, Multivariate Adaptive		Normal Probe, U2R, R2L, Dos Attack	97%

	Regression Splines, Linear Genetic Programs			
Md Reazul Kabir, Abdur O., Tanvir S.	Wrapper Type Bayesian Network		Normal Probe, U2R, R2L, Dos Attack	
Warusia Y., Nur Izura U., Zaiton M., and Md. Nasir S.	K-Means Clustering, Naïve Bayes Classification	DARPA'99, ISCX 2012	Normal Probe, U2R, R2L, Dos Attack	97%
Gvin W., Dr. Taghi M. K.	Nearest N., Naïve Bayes, DT	SANTA and IRSC		
Kanagalakshmi R., V. N. Raj	Hidden Naïve Bayes Multiclass Classifier	KDD Cup 1999	Dos Attacks	
Zainab A., Antoun Y., Sara M.	Decision Tree	Data provided by a faculty of Science's Web Server of a University	Dos Attacks, U2R, R2L	89%
Jasmin K., Samed J., Abdulhamit S.	DT and Random DT	NSL KDD	Normal Probe, U2R, R2L, Dos Attack	88%
Dewan Md. Farid, Nouria F., M. Z. Rahman	Hybrid Learning Algorithm using Naïve Bayes Classifier	KDD'99	Normal Probe, U2R, R2L, Dos Attack	99%
Uzair B., Manzoor C.	J48 Decision Tree and Naïve Bayes Classifier		Normal Probe, U2R, R2L, Dos Attack	DT -100% NB – 90%
Amirhossein Gharib, Iman S.		DARPA, ADFA, KYOTO, ISCX, KDD'99, LBL, UMASS	Normal Probe, U2R, R2L, Dos Attack	
Robert M., Shay P.	Decision Tree, Bayesian Networks, ANN	NSL KDD	Dos Attack	DT+NB– 99% ANN-96%
Rutvij P., Jayati P.	C5 Classifier			
Paul N. Bennet	Naïve Bayes			
Ron Kohavi	Naïve Bayes			
Kiran D., Deepak K., Arif K.			TCP SYN Flood, Back/Land, Buffer over Flow, Abnormal Packets	
Naila B. A., Mohammed G.	K-means clustering	KDD'99		75%-98%
Amit k., Harish Chandra M., Rahul M.	Combination of Naïve Bayes Classifier and K-means Clustering	KDD'99, NSL KDD	Dos, Probe, U2R, R2L	
Li Tian, Wang Jiawen	K-means clustering	KDD Cup'99	Dos, Probe, U2R, R2L	98.60%
Mohammadreza E., Sara M., Fatimah S.	C4.5 and SVM	KDD'99	Dos, Probe, U2R, R2L	
Petrussenko D.	LERAD	DARPA		
Md Reazul K., Abdur O., Tanvir S.	Genetic Algo., Bayesian Network, Wrapper Approach	NSL-KDD	Dos, Probe, U2R, R2L	99.90%
Nahla A., Salem B., Zied E.	Naïve Bayes and Decision Tree	KDD'99		DT>NB

III. MACHINE LEARNING

Machine learning is to learn from data repetitively and to find patterns hidden there. By applying the learning result to the new data, we can predict the future according to the pattern. Algorithms implemented by manual programming can be automatically constructed from a large amount of data, so they are applied in various fields [4]. There are various machine learning algorithms used to predict the result, some of them that the we are going to use in our work are:

A. Decision Tree Classifier

The Decision Tree Classifiers [11] organized a series of test questions and conditions in a tree structure. Following are the different parameters related with Decision Tree. For the detection of various attacks most of the researchers have used decision tree algorithm to improve its accuracy.

Entropy: The loss while transferring data from parent node to child node.

Entropy = - (N+/N log(N+/N)- (N-/N log(N-/N))

N+ is No. Of positive tuples in child node

N- is No. Of negative tuples in child node

Information Gain: It is the amount of information which is transferred from parent node to child node. It is the opposite of Entropy [9].

We split the dataset on the parameter where we get more Information Gain or less Entropy.

Given below is an example of Decision Tree Algorithm.

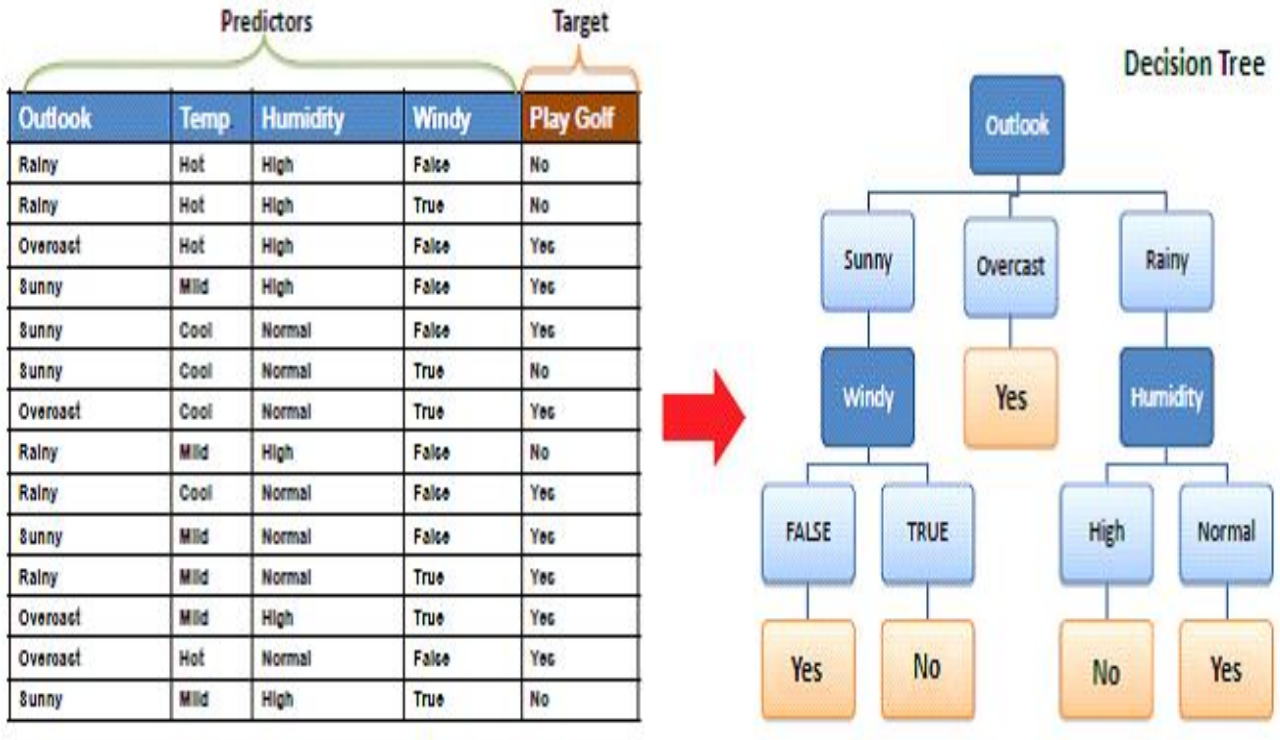


Figure 1: Example of Classification Tree

Most of the authors have split the dataset on various parameters, like number_of_connections or port_number_connections, which are common on every packet and contend by almost all packets. After that the dataset is again split on different parameter such as normal error or service, which are present on mainly packets, but not on all. When the packets go in the analysis phase, its parameters are checked against the tree where it transverse down the tree according to the content of the packets. Finally, it reaches the point where it is either allowed or rejected.

B. Naive Bayes Classifier

Naive Bayes Classifier theorem [3] is probabilistic theorem which predicts the occurrence of happening of an event based on the prior evidence. It calculates the probability of happening of an event A if event B has already happened. For this it assumes that all the events are mutually exclusive. Following are the terminologies used in the theorem:

Posterior(P[A/B]): Probability of happening of event A when event B has already happened

Prior(P[A]): Probability of happening of event A

Likelihood(P[B/A]): Probability of happening of event B when event A has already happened

Evidence(P[B]): Probability of happening of event B

Hence, $P[A/B] = P[A]. P[B/A]/P[B]$

Researchers have used this classifier to calculate various attacks such as DOS, Probe, U2R.

IV. REVIEW WORK

Mrutyunjaya P., Manas Panda, et. al. has proposed an Intrusion Detection System with a probabilistic approach of Naive Bayes Theorem. The author has used KDD'99 dataset. The authors have assumed that given data belongs to a particular connection, and they have calculated the probability of belonging to a certain class for each connection. Naive Bayes Theorem is a simple theorem and gives an accurate result as compared to the other Simple Theorems. The disadvantage of this approach is that authors has used a very old dataset. New datasets are available with new and concrete information's. These datasets can also detect new type of attacks. Secondly, the Naive Bayes does not give a concrete output. Being a probabilistic approach, the attacker can intrude in the system with some tweaks in its packets [3].

Kanagalakshmi. R, Naveenantony Raj, et. al. has proposed a system using Naive Bayes Multiclass Classifier Model. They have extended the Naive Bayes theorem which relaxes the conditional independence assumption imposed on Naive Bayes Model. They have also included information about various algorithms which can be used in making intrusion detection system. But they have not given any information about their implementation [29].

Latifur Khan, Mamoun Awad, Bhavani Thuraisingham, et. al., have proposed an intrusion detection system using Support Vector Machines and Hierarchical Clustering. They have claimed that the SVM algorithm is more efficient for IDS than any other Algorithm. They have also used a Dynamically Growing Self-Organizing Tree (DGSOT) which overcomes the drawbacks of traditional hierarchical clustering algorithms [31].

Mohammadreza E., Sara M., Fatimah S., et. al., compared C4.5 and SVM to show the performance of both algorithm and FAR values too. Among these two C4.5 works better compared to other [32].

Petrussenko D. has implemented an IDS with LERAD to detect attacks in network packets and TCP flow to capture a new form of attack patterns. He found LERAD performs well using DARPA dataset with high detection rate [20].

Md Reazul K., Abdur O., Tanvir S., et. al., have proposed an IDS using Wrapper Approach. The authors have used NSL KDD dataset and have first filtered the packets using a Genetic algorithm and then have applied Bayes Theorem to classify the packets. They have compared various Selection algorithm and claims that their Algorithm performs the best. IDS used less features during the training and validation phase for reducing time complexity [15].

Iwan S., Adam PB., Gary W., have proposed an efficient IDS using Dimensionally Reduction of the unwanted variables. The author claims that some of the variables may be redundant or make less contribution to the detection process. The authors have also compared and contrasted various algorithms using his wrapper approach and have graphically represented it [6].

Nahla A., Salem B., Zied E., et. al., have compared Naive Bayes and Decision Tree Algorithm for Intrusion Detection System. The authors have used a very old Dataset which is already outdated, KDD'99. Although the accuracy of both the algorithms is good, the author finds that the accuracy of Decision Tree is slightly greater than Naive Bayes Theorem [21].

Suseela T., et. al., have proposed a multilevel hierarchical K-Map for intrusion detection system. The major advantage of this approach was the computational effectiveness. The reduced network size was also an advantage. On the down side, K-Map's approach does not engage expensive point-to-point computation in grouping the data into K- map clusters [5].

Iftikhar A., Azween., et. al., have proposed a paper to overcome presentation issues an optimized interference detection algorithm by means of soft computing techniques. The SVM algorithm was applied for classification. They have also applied Genetic Algorithm as a favorable or optimal choice of principal component.

Zainab A., Antoun Yaccoub, Sara M. et. al. has proposed a Network Intrusion System with Data Mining Technique. In the paper, the author has used Decision Tree algorithm for the classification. Author has used the collection of data provided by a faculty of

science's web server of a particular University. This approach classifies the packets only when the algorithm reaches the node. The disadvantage of this approach is Author has used the dataset provide by a certain University which may or may not contains enough data of attacks or different information for a certain attack [30].

Jasmin K., Samed J., Abdulhamit S., et. al. has proposed an Intrusion Detection System which involves a combination of two algorithms. The authors have combines Decision Tree algorithm and Random Tree algorithm for more accurate results. In the paper, the author claims that to combine two classifiers, the sum rule outperforms other classifications scheme. The authors have achieved slightly higher accuracy when combining the classifiers than a standalone classifier. The authors have not achieved an agreeable accuracy. In IDS, accuracy of 88% would mean that out of 100 bad connections, it has detected 88 correctly. But a single computer gets a high number of connections each day, around in 1000. Thus, this would mean that a large amount of connections could enter the system which could harm the computer [11].

Sravan J., Ravi R., et. al., have proposed a paper in which they have done an exhausted literature survey of all the algorithms available that can be used for IDS. The authors have also compared the output of those author's work and have contrasted their Detection Rate [2].

Sandhya Peddabachigari, Ajith Abraham, Crina Grosan, Johnson Thomas has proposed a Network Intrusion Detection System using Hybrid Intelligence System. In this they have used decision trees (DT) and support vector machines(SVM) which are combined as a hierarchical hybrid intelligent system model (DT–SVM) and an ensemble approach combining the base classifiers. They have used the KDD Cup 1999 Intrusion detection contest data. This data was prepared by the 1998 DARPA Intrusion Detection Evaluation program by MIT Lincoln Labs. The data set is first passed through the DT and node information is generated. Node information is determined according to the rules generated by the DT. The key idea here is to investigate whether the node information provided by the DT will improve the performance of the SVM. With this new approach they have concluded that the DT gives better or equal accuracy for Normal Probe, U2R and R2L classes. The hybrid DT–SVM approach improves or delivers equal performance for all the classes when compared to a direct SVM approach [5].

Gavin Wolf, Dr. Taghi M. Khoshgoftaar from Florida Atlantic University Boca Raton, Florida USA have proposed a Network Intrusion Detection System using Machine Learning Techniques. The researchers have used four different types of classification models for their analysis which are 5Nearest Neighbor (5-NN), two forms of C4.5 Decision Trees (C4.5D and C4.5N), and Naïve Bayes (NB). Instead of using old Datasets like KDD 99 / DARPA, KYOTO and ISCX they have used SANTA and IRSC. As this dataset seem to be meaningful improvements over the three most popular datasets, they have not yet been made publicly available. This researcher mainly focused on the dataset they have used and not on the algorithms [4].

Kanagalakshmi, V. Naveenantony Raj have proposed Network Intrusion Detection System using Hidden Naive Bayes Multiclass Classifier Model. This paper mostly focuses on Hidden Naive Bayes model, the observation results show that the HNB model exhibits a superior overall performance in terms of accuracy, error rate, and misclassification cost compared with the traditional naive Bayes model, leading extended naive Bayes models and the Knowledge Discovery and Data Mining (KDD) Cup 1999. HNB model performed better than other leading state-of-the art models, such as Support Vector Machine, in predictive accuracy. The results also indicate that HNB Model significantly improves the accuracy of detecting denial-of-services (DoS) attacks. They have concluded that hidden naive Bayes is a promising model for datasets with dependent attributes, like that the KDD’99 intrusion detection dataset [29].

Zainab A., Antoun Yaccoub, Sara M. et. al. has proposed a Network Intrusion System with Data Mining Technique. In the paper, the author has used Decision Tree algorithm for the classification. Author has used the collection of data provided by a faculty of science's web server of a particular University. This approach classifies the packets only when the algorithm reaches the node. The disadvantage of this approach is Author has used the dataset provided by a certain University which may or may not contain enough data of attacks or different information for a certain attack [30].

Mrutyunjaya P., Manas P., et. al. has proposed an Intrusion Detection System with a probabilistic approach of Naive Bayes Theorem. The author has used KDD'99 dataset. The authors have assumed that given data belongs to a particular connection, and they have calculated the probability of belonging to a certain class for each connection. Naive Bayes Theorem is a simple theorem and gives an accurate result as compared to the other Simple Theorems. The disadvantage of this approach is that authors has used a very old dataset. New datasets are available with new and concrete information’s. These datasets can also detect new type of attacks. Secondly, the Naive Bayes does not give a concrete output. Being a probabilistic approach, the attacker can intrude in the system with some tweaks in its packets [18].

Research paper published by Dewan Md. Farid, Nouria Farbia and Mohammad Zahiur Rahman introduced a new hybrid learning algorithm for adaptive network intrusion detection system using Naive Bayes classifier which analyses the large volume of network data and considers the complex properties of various attack behaviours to improve the performance of prevention and detection. The attacks of KDD'99 dataset detected with 99% accuracy rates. One main disadvantage is false positive rate is bit lower than expected which needs to be improved [10].

Research paper published by Uzair Bashir and Manzoor Chachoo focuses on designing an IDS using machine learning algorithms. The algorithms chosen by them are J48 Decision tree and Naive Bayes classifier. The results obtained by the experiments revealed that J48 performs well for designing and implementation of IDS than Naive Bayes algorithm. The detection rate of Naive Bayes classifier reaches up to 90% mark whereas detection rate of J48 Classifier reaches up to 100% mark. The scenario for False positive rate is similar [26].

In the research paper published by Amirhossein gharib, Iman Sarafaldin they have analyzed various existing datasets such as ADFA, DARPA, ISCX, KYOTO, LBNL, KDD'99, UMASS, DEFCON, CDX for the training and validation of IDSs. Various traffic generation strategies such as configurability, low cost, Monitoring, Attack, scalability have been compared and reviewed, and presented as a new framework which can be used to evaluate datasets with characteristics such as: Network Traffic, Labelling, Interaction, Packet Capturing, Protocols, Attack Diversity, Metadata, etc [25].

In a Research Paper published by Robert Moskovidic, Shay Pluderman four classification algorithms were applied on various feature subsets for detecting malicious code in (HIDS) host based intrusion detection system. Also an Experimental design of 4 different hypotheses was made to check accuracy of machine learning algorithms on various attacks. While Decision Trees and Bayesian Networks achieved 99% accuracy, ANN achieved 96%, and Naive Bayes achieved 92%. The categorization of worm activity is is more challenging than classifying to a generic worm behaviour. Thus, the results can be improved while evaluating this method [23].

In the Research Paper Published by Rutvij Pandya and Jayati Pandya survey of various classification algorithms such as Decision trees, Rule based Classification, K-Nearest neighbour, SVM are done through C5 Classifier to find the Speed, Scalability, Interpretability, Accuracy and Transparency of Proposed System. In the proposed system C5 classifier classifies the data in less compared to other classifiers. Further scope of classifier is achieved by implementation of new features like Cross Validation and Model Complexity.

Paul N. Bennet has explained the conditions and theoretical reasons that give rise to the nearly binary posterior estimates of Naive Bayes while assessing the calibration of Naive Bayes posterior estimates. Using this explanation, Paul has justified that modelling requires the posterior given by the log odds scores of Naive Bayes. Paul designed his own Naive Bayes Unigram model to predict log odd scores of the model. Finally, he considered and evaluated an initially promising candidate sigmoid family for this approach [14].

In the Research Paper Published by Ron kohavi, He explained Naive-Bayes induction algorithms were previously shown to be surprisingly accurate on many classification tasks even when the conditional independence assumption on which they are based is violated. Generally, in larger databases, the accuracy of Naive-Bayes does not scale up as compared to decision trees. So a new algorithm, NBTree was proposed by him, which induces a hybrid of decision-tree classification tree and Naive Bayes. To evaluate a NBTree a large set of data files were used to improve the accuracy [17] .

Kiran D., Deepak K., Arif K. Proposed a hybrid IDS system to detect various types of attacks like Buffer over Flow, TCP SYN Flood and Abnormal packets in Network IDS and also for Unauthorized accessing and login/ and logout error. Attack comparisons are also done on the type of attacks and Number of packets as input. The proposed model offers several advantages over traditional IDS. It supports High Scalability, High Availability and High Security. The IDS basically captures and Monitors IP Packets for potentially dangerous user activities on the network. There are basically two types of host-based intrusion detection software: host wrappers and agent-based software. Host wrappers are tools that can be configured to monitor at all the network packets, login attempts or configuration attempts on the monitored machine. The agent-based software has similar abilities, but it can be used to detect changes in system files and also manipulation in user privileges. [28]

This paper proposed by Naila B., Aissa and Mohammed Guerroumi is anomaly based detection scheme that uses an unsupervised clustering approach combined with a genetic algorithm for Anomaly based detection. This paper proposes an anomaly based detection scheme that uses an unsupervised clustering approach combined with a genetic process. The main purpose of CG-AD is to obtain an optimal homogenous partitioning of normal and anomaly instances. Due to the computation of a confidence interval in the fitness function, a cluster of rejected instances is created. The authors have evaluated the performance of scheme using different subsets of KDD'99 dataset [13].

In the research paper Published by Amit Kumar, Harish Chandra Maurya and Rahul Mishra they have proposed a hybrid based IDS. Basically 2 classification techniques are used for their proposed architecture which can be used in Hybrid format. The first one is Bayes Theorem or Naive Bayes Classifier for Intrusion Detection System and K-means Clustering. They have discussed about detecting and preventing various attacks such as DOS attacks, U2R and R2L [1].

Because of the k-means algorithm's shortcomings about dependence and complexity, the paper published by Tong H. and Wang Jiawen, Lu W. put forward an improved clustering algorithm through studying on the traditional means clustering algorithm. The new algorithm focuses on the strong points from the k-medoids and improved relations trilateral triangle theorem. The observation experiments proved that the new algorithm could improve accuracy of data classification and detection efficiency significantly. The results show that this algorithm achieved the desired objectives with a very efficiency and high detection rate. Cluster algorithm in the network intrusion detection system mainly focuses for abnormal detection technology. The system is established by clustering method. The model obtains very high fault detection rate while maintains the lower false drop rate [24].

In the research published by Mrutunjaya Panda and Manas patra a cost matrix is developed which calculates average classification of each connection. In this paper, a framework of IDS is proposed based on Naive Bayes algorithm. The framework builds the patterns of the network services over data sets. With the built patterns, the framework detects attacks in the datasets using the Naive Bayes algorithm. Compared to the Neural network based approach, this approach takes less time and supports low cost factor and higher detection rate [27].

V. PROPOSED WORK

We propose to build a Host based Intrusion Detection system using some Machine Learning Algorithm Mainly Naive Bayes Decision Tree Algorithms and compare the accuracy and time required of both. For this, we are using a more efficient and recent dataset.

In case of Decision Tree for our project, we split our dataset on various parameters, like number_of_connections or port_number_connections, which are common on every packet and contend by almost all packets. After that the dataset is again split on different parameter such as normal_error or service, which are present on mainly packets, but not on all. As we transverse down the tree, we observe that the packets get more specific. This is continued till we get an efficient tree where the
connections are clearly classified based on their packets fields. When the packets go in the analysis phase, its parameters are checked against the tree where it transverse down the tree according to the content of the packets. Finally, it reaches the point where it is either allowed or rejected.

In case of Naive Bayes for our project we find the probability of each and every attacks from the dataset based on different parameter like, P[number_of_connection, DOS] or P[service,U2R] etc. We then apply the Bayesian theorem to the parameters of the packets which are analysed in the analysis phase. As we have multiple parameters, the probability is calculated on all the fields, like
P[A/B] = (P[A]). (P[B 1 /A]). (P[B 2 /A]). (P[B 3 /A]). (P[B 4 /A]) ..... (P[B n /A]/P[B])

We select a cut-off probability. Above it, we declare the packet to be malicious. Generally, the cut-off probability is 0.65. If the prediction algorithm finds the packet to be malicious, then it will raise an alarm where the IDS system will know about the malicious activity and a log will be entered containing the information of the attacker for future references. Finally, such connections will be terminated.

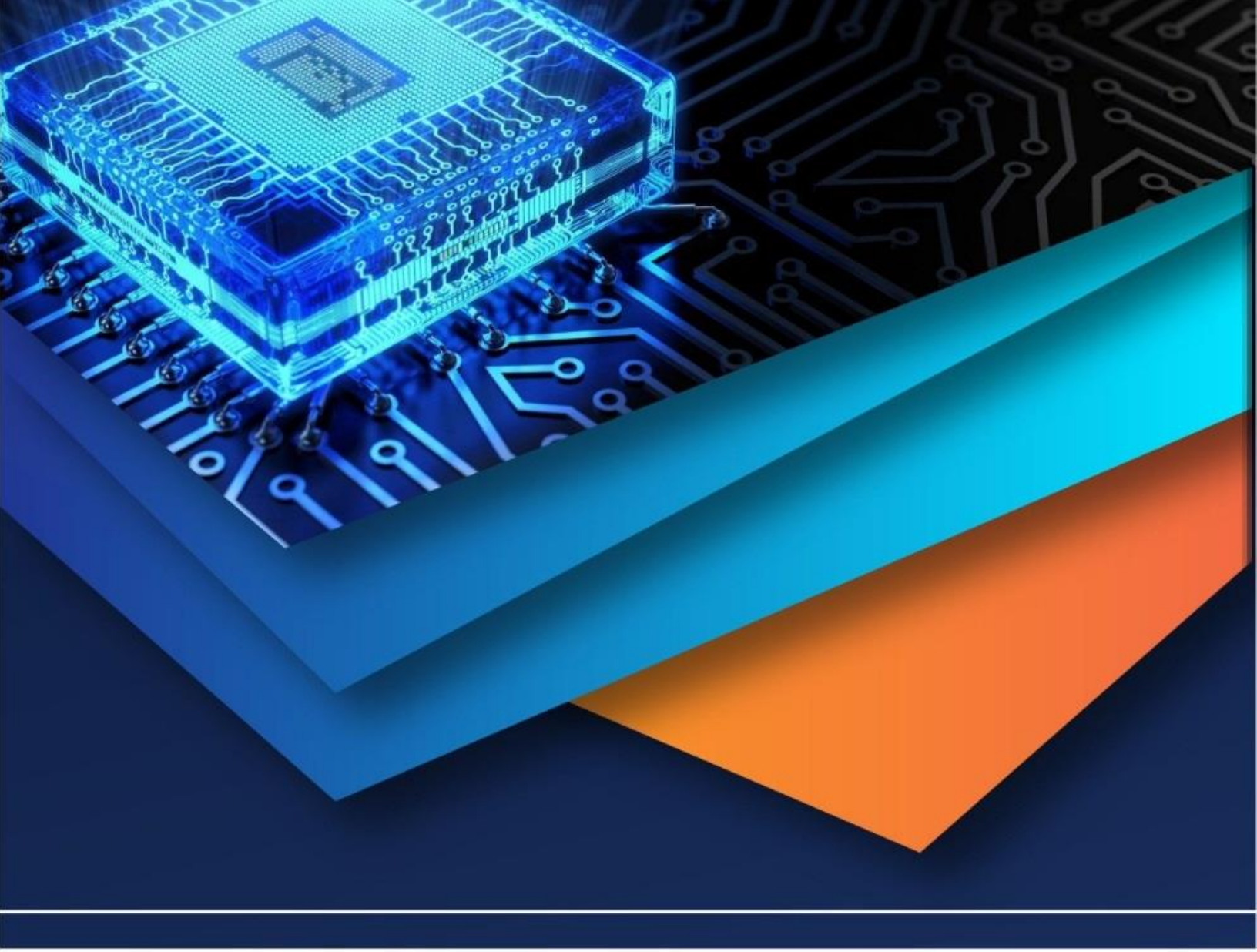
VI. CONCLUSION

From the review work done from all research papers of IDS, we have come to a conclusion that most of the authors have used NSL KDD dataset or KDD'99 Dataset, although some of them have used ADFA and ISCX dataset to detect new types of attack. From the reference papers we see that authors have mainly used Decision Tree, Naive Bayes or SVM to calculate accuracy under various attacks. Some of them have also used Combined classifier which uses combination of such machine learning algorithms.

REFERENCES

[1] Harish C., Palash Chaturvedi, Amit K. Saxena, "A Systematic Literature Survey in IDS", International Journal on Recent and Innovation Trends in Computing and Communication, Vol 5, Issue:6, 2017
[2] Sravan K. Jonnalagadda, Ravi P. Reddy, "A Literature Survey and Comprehensive Study of Intrusion Detection", International Journal of Computer Applications, Volume 81, 2013
[3] Mrutyunjaya Panda, Manas R. Patra, "Network Intrusion detection using Naive Bayes", International Journal of Computer Science and Network Security, Vol 7, 2007

[4] Gavin wolf,Dr.Taghi,M.Khoshgoftaar, "Using Machine Learning for Network Intrusion Detection: The Need for Representative Data", Florida Atlantic University Boca Raton, Florida USA, May 2016
[5] Suseela T., Q.A. Zhu, "Hierarchical Kohonenen net for anomaly detection in network security J.Huff",IEEE Transactions on Sysytem,Man and Cybernatics,Part(B),Cybernatics (Volume: 35, Issue:2, April 2005)
[6] Ajith Abraham, Sandhya Peddabachigari, C. Grosan, Johnson Thomas, "Modelling intrusion detection system using hybrid intelligent systems", Journal of Network and Computer Applications, 2007
[7] Iwan Syarif, Adam Prugel-Bennett, Gary Wills, "Data Mining Approaches for Network Intrusion Detection: from Dimensionality Reduction to Misuse and Anomaly Detection", Journal of Information Technology, Vol 3, 2012
[8] Therese R. Metcalf, Leonard J. LaPadula, "Intrusion Detection System Requirements: A Capabilities Description in terms of Network Monitoring and Assessment Module of CSAP21", MITRE, MP 00B0000046, 2000
[9] Mukkamala, Andrew Sung, Ajith Abraham, "Designing Intrusion Detection Systems: Architectures, Challenges and Perspectives", Journal of Network and Computer Application, Vol 28, Issue 2, 2005
[10] Dewan Md. Farid, Nouria Farbia and Mohammad Zahiur Rahman," Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection" International Journal of Network Security & Its Applications 2.2 (2010) 12-25
[11] Jasmin Kevric, Samed Jukic, Abdulhamit Subasi, "An effective combining classifier approach using tree algorithms for network Srinivask intrusion detection", Neural Computing and Application, Vol 28, Issue 1 2017
[12] Mohammadreza E., Sara M., Fatimah S. "Intrusion detection using data mining techniques", Information Retrieval (HYPERLINK http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5462929)
[13] Naila B. Aissa, Mohammed Guerroumi, "A genetic clustering technique for Anomaly-based Intrusion Detection Systems", Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference
[14] Paul N. Bennett, "Assessing the Calibration of Naive Bayes' Posterior Estimates", Carnegie Mellon University, 2000
[15] Md Kabir, Abdur Onik, Tanvir Samad", A Network Intrusion Detection Framework based on Bayesian Network using Wrapper Approach", International Journal of Computer Applications, Vol 166, Issue 4, 2017
[16] Tamas Abraham, "IDDM: Intrusion Detection using Data Mining Techniques", DSTODCD, Defense Science and Technology Organization Document Control, DSTO-GD-028, 2001
[17] Kohavi R (1996) Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In: Proceedings of the second international conference on knowledge discovery and data mining.
[18] Panda M, Abraham A, Patra MR (2010) Discriminative multinomial Naive Bayes for network intrusion detection. In: Sixth international conference on information assurance and security
[19] Panda M, Abraham A, Patra MR (2012) A hybrid intelligent approach for network intrusion detection. Procedia Eng 30:1–9.
[20] Petrussenko D.,Philp K. Chan, "Incrementally Learning Rules for Anomaly Detection", Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)
[21] Nahla Amor, Salem Benferhat, Zied Elouedi, "Naive Bayes in Intrusion Detection System", ACM, 2004
[22] Kumar S., Spafford E. H. (1994) "An Application of Pattern Matching in Intrusion Detection," Technical Report CSD-TR-94-013. Purdue University.
[23] Robert Moskocovic,Shay Pludeman, et.al "Host Based Intrusion Detection using Machine Learning", Intelligence and Security Informatics,2007 IEEE
[24] Lu, W., Tong, H. (2009). Detecting Network Anomalies Using CUSUM and EM Clustering. In: Proceedings of the 4th International Symposium on Advances in Computation and Intelligence, p. 297–308.
[25] Amirhossein Gharib, Imar Sharafaldin , Arash Lashkari, Ali Ghorbani " An Evaluation Framework for Intrusion Detection System", Institute of Electrical and Electronics Engineers,2016
[26] Uzair Bashir, Manzoor Chachoo, "Intrusion detection and prevention system: Challenges & opportunities" Computing for Sustainable Global Development (INDIACom), 2014 International Conference
[27] Panda, M. and Patra, M.R. 2009. Ensemble of classifiers for detecting network intrusion. In: Proceedings of the International Conference on Advances in Computing, Communication and Control (Mumbai, India, 2009) 510–515.
[28] Kiran D., Deepak K., Arif K. International Journal of Computer Applications (0975 – 8887) Volume 65– No.23, March 2013.
[29] Kanagalakshmi R., V. Naveenantony Raj, International Journal of Science, Technology & Management Volume No.03, Issue No. 12, December 2014.
[30] Zainab Assaghir, Antoun Yaacoub, Sara Makki International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 1
[31] Latifur Khan, Mamoud Awad, Bhavani Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering", The VLDB Journal (2007) 16: 507-521
[32] Mohammadreza E., Sara M., Fatimah S., et. al., compared C4.5 and SVM to show the performance of both algorithm and FAR values too. Among these two C4.5 works better compared to other.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)