Highly Accurate and Efficient Two Phase - Intrusion Detection System (TP-IDS) using Distributed Processing of HADOOP & Machine Learning Techniques

Abhijit Dnyaneshwar Jadhav (abhijit.jadhav29@gmail.com)
Koneru Lakshmaiah Education Foundation https://orcid.org/0000-0003-0151-6458
Vidyullatha Pellakuri
Koneru Lakshmaiah Education Foundation

Research

Keywords: accuracy, efficiency, intrusion detection, machine learning, security, TP-IDS

Posted Date: June 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-565099/v1

Highly accurate and efficient Two Phase - Intrusion Detection System (TP-IDS) using distributed processing of HADOOP & machine learning techniques

Abhijit D. Jadhav[1], Vidyullatha Pellakuri[2]

Department of Computer Science & Engineering[1]
Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.[1]
Assistant Professor, Department of Computer Engineering[1]
Dr. D. Y. Patil Institute of Technology, Pimpri, Pune-18[1]
abhijit.jadhav29@gmail.com[1]

Department of Computer Science & Engineering[2]
Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.[2]
pvidyullatha@kluniversity.in[2]

Abstract— **Network security and data security are the biggest concerns now a days. Every organization decides their future business process based on the past and day to day transactional data. This data may consist of consumers confidential data, which needs to be kept secure. Also, the network connections when established with the external communication devices or entities, a care should be taken to authenticate these and block the unwanted access. This consists of identification of the malicious connection nodes or identification of normal connection nodes. We expect, everytime whenever there is a connection request, it should be recognized as a type of normal node or malicious node connection request. For that, we use a continuous monitoring of the network input traffic to recognize the malicious connection request called as an intrusion and this type of monitoring system is called as Intrusion detection system(IDS). IDS helps us to protect our network and data from insecure and malicious network connections. Many such systems exists in the real time scenario, but they have critical issues of performance like accuracy and efficiency. These issues are addressed as a part of this research work of IDS using machine learning techniques. The TP-IDS is designed in two phases for increasing accuracy. In phase I of TP-IDS, Suppor Vector Machine (SVM) and k Nearest Neighbor (kNN) are used. In phase II of TP-IDS, Decision Tree (DT) and Naïve Bayes (NB) are used, where phase II is the validation phase of the system for increasing accuracy. Also, both the phases are**

having Hadoop distributed file system underlying data storage & processing architecture, which allows parallel processing to increase the speed of the system and hence achieve the efficiency in TP-IDS.

Keywords— accuracy, efficiency, intrusion detection, machine learning, security, TP-IDS

I. Introduction

Machine learning is the amazing asset of the technology, which enables us to provide solution almost for every problem in day to day life. It is being used by the software industries frequently for developing software solutions for business organizations. With the large data size, the machine learning can give excellent results. One of the critical concerns of the business organizations, is the security of the assets and unauthorized access. Every business organization wants to protect their assests like day to day transactional data, business functioning software systems, from the malicious attacks which can be with the intention of stealing data or damaging data and systems. So, it is required to examine every outside attempt entering the organizations network and block, if the malicious one. To examine such incoming network connection, the monitoring systems are established. Such, monitoring systems accepts the incoming network traffic and identifies it, whether a normal connection or malicious connection request. If, the malicious one, then it blocks the connection request, else allows the connection. Such, malicious connections are called as intruders and monitoring systems which detects the intruders are called as intrusion detection systems (IDS). [1] defines IDS as “An intrusion-detection system can be described at a very macroscopic level as a detector that processes information coming from the system that is to be protected”. This detector uses three kinds of information: longterm information related to the technique used to detect intrusions (a knowledge base of attacks, for example), configuration information about the current state of the system, and audit information describing the events that are happening on the system[1].

IDS can be categorized in two differen ways, like based on location and based on detection mechanism. The types of IDS based on location are Nework based IDS (NIDS) and Hos based IDS(HIDS). NIDS is the system which is placed at the network boundry or beween nework and the server of the nework, whereas HIDS is the system which is placed at the standalone system itself to monitor the traffic for internal and external attacks[2]. The types of IDS based on detection mechanism are Misuse detection systems and Anomaly detection systems. Misuse detection is the IDS which matches the known attack patterns with input traffic, whereas Anomaly detection is the IDS which defines the boundry for normalcy of the behavior and anything outside the boundry detected as the malicious or anomaly behavior. Many attempts have been made and come up with different IDS solutions of these types. Every such IDS system has its advantages and limitations based on the methods implemented as a part of the IDS architecture[3]. In the given categories of IDS, the anomaly detection has proven a valuable approach for providing the effective network security as well as network management[4]. Over the years many different models are developed for Anomaly detection, using different methods & techniques for achieving accuracy in the results. Most of the models are developed using the machine learning techniques like supervised learning techniques, unsupervised learning techniques, deep learning techniques etc. Machine learning is the process, which allows to develop automated decision systems based on the available sample of the data. In Machine learning, the larger the amount of the data, more the accuracy of the system is achieved, provided a good quality data is available. Anomaly based systems are also work same way and basically have three phases like parameterization, training phase and finally the detection phase[5]. Every machine learning model have very important training phase which ensures the learning of the model for automated decisions as a solution of the problem. By taking this advantage of the similarity between the anomaly detection systems and the machine learning models, the solution can be effectively built by developing a model with the combination of different machine learning techniques. The time requirement for general nework IDS is much higher than the service time in distributed processing environment[6]. The data of any size can be processed in distributed environment by using the HADOOP distributed file sysem. To increase the speed of the machine learning algorithms execution, this distributed processing is very effective data processing system and it is widely used. The intrusion detection is very critical operation from the time's perspective, it should identify the intrusion or malicious activities within acceptable time frame only, to avoid any loss of data confidentiality, data damage, damage to organizational data assets.

Many IDS systems are proposed and implemented, but every system has a limited success rate in view of the accuracy of the detection and efficiency or time requirement. Among the techniques used for IDS implementation, machine learning has gained the more attention as compared to other systems. Machine learning enables the more and more accuracy in detections with respect to use of the model and with respect to time. Hence, it is preferred by most of the researchers. Machine learning consists of different techniues like supervised machine learning, unsupervised machine learning and semi supervised machine learning. Every machine learning techniques has two phases- training and testing. In training phase, the machine learns for the problem input and output, wherein testing phase, machine is given only input data, based on the knowledge extracted during training, machine generates the output. More the data is passed in the training phase, more is the accuracy we obtain in testing phase. In Supervised machine learning techniques, the input data and its corresponding output is passed during training phase, machine has to find out the mapping function or transformation function between the input and output passed. Classifiation is the example of the Supervised machine learning technique. In Unsupervised machine learning techniques, the input data is passed only, machine has to extract the knowledge about the mapping function between input and output, also machine has to generate the output. Clustering is the example of the unsupervised machine learning technique.In semi supervised machine learning algorithm,we pass the input and output if available, else we pass only input, the machine has to extract the mapping function and output also, if not available. Supervised machine learning are very popular methods in developing machine learning models solutions. The popular supervised machine learning techniques of classification are SVM, Decision tree, naïve bayes etc. Unsupervised machine learning techniques are also very popular and effective, specially when the output data values are not available, these methods are proven effective methods. The popular unsupervised machine learning techniques of clustering are kNN clustering, k Means clustering, DBSCAN etc. Among the mentioned machine learning algorithms, supervised learning algorithms gives a good accuracy in case of known attacks but fails to detect the unknown attacks, which is the major area of concern. Wherein, unsupervised learning algorithms help us to detect the unkown attacks, but the result accuracy should be verified with other techniques. Its not possible to consider or trust the results generated by these methods alone. Classification techniques, during training phase gets the idea about input and its respective outputs, it only extracts the mapping function, because of which only the known input type is correctly classified by these classification techniques. Clustering techniques, during training phase gets only input values, it extracts the mapping function as well as the unlebelled output cluster of the input value.

Machine learning is oftenly useful when these techniques are used together to form a model as a solution to the problem. A combination of different algorithms under different machine learning techniques can be used to develop the effective solutions. Many different solutions for different problems are proposed and developed by researchers by using such combination of machine learning techniques only. Every techniques, in the mentioned supervised and unsupervised learning techniques, has its advantages and disadvantages. Inspite of this, these techniques, if combined properly, can give excellent results. In the classification techniques Support Vector Machine(SVM), Decision Tree (DT), Naïve Bayes(NB) are few of the popular techniques, which helps us to generate better performance models. Also, in clustering techniques kNN clustering, k Means, DBSCAN, Heirarchical clustering are also few of the popular clustering techniques, which helps us to deal with the problems having unkown and non categorized data. Intrusion detection is one of those problems where, the known attack patterns can be detected by using machine learning classification techniques. But, the major concern is when the attack is unkown attack pattern. Hence, here we need a model which can detect the unkown attack pattern also and that too with a higher accuracy. So, it is important to develop a model with the combination of such techniques, which can help us to get better accuracy results in both known and unknown attack input patterns. Also, it is important to get these results within a minimal time frame, and when we are using a combination of different techniques, it might slow down the execution time of the system. Hence, if we execute these combination of techniques with the underlying data processing system in distributed environment, then faster results can be generated. So, distributed and parallel data processing structures like HADOOP, which is a distributed file system for data storage and data processing of the system, can be used to develop the machine learning models with acceptable execution rates, specially in a time sensitive system like Intrusion Detection System.

II. LITERATURE SURVEY

The survey of different existing IDS implementations is carried out, before finalizing the solution. While carrying out survey, the advantages and disadvantages of these systems are observed. The survey study consists of few successfully implemented IDS solutions. The survey is as follows:

In [7], M. A. Rassam and et. al. have proposed a IDS solution based on smart and generic rule construction. The smart rules are the rules which are formed as a single rule in place of 2-3 rules, which can detect multiple attacks. Hence, these are also named as generic rules. In the said work, first step is data pre processing, followed by smart and generic rule construction, after which constructed rules learning is carried out. The advantage of the system is, because of construction of the smart rules, the smaller number of rules help to detect maximum number of attacks, hence less power consumption. But, this system is proposed considering, maximum attacks are from the internal systems of the network. So, it can not be considered for outside incoming attack detections, as the size of the network is large for outside incoming attack detections.

In [8], P. Amudha and et. al have presented the Intrusion detection system using hybrid swarm intelligence. The system is organized of two Intelligence algorithms, Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC). The data set used for the said work is KDDCUP 99. Initially the data preprocessing techniques are used for better results, feature selections are done by using SFSM(Single Feature Selection Method) and RFSM (Random Feature Selection Method). The PSO and ABC are applied as intelligence algorithms & hybrid model is developed, which gives the 99.5% accuracy for detection of known attacks. But, the problem is in detecting unknown attacks, which is not tested and presented by the researchers in this article. Also, the timeliness in detecting attacks is not considered while implementing this model. Hence, the model can not deal with unknown attack identification and time of detection is also the concern, that is required to be addressed.

In Intrusion detection system, the major areas of concern are the quality attributes which can not be compromised. The important measures are: Accuracy, Performance, Completeness, Fault tolerance, timeliness[9]. These properties should be developed and considered while developing the better intrusion detection system.

In [10], Zeeshan Ali Khan and et. al have carried out the survey about performance of different decision tree algorithms in data mining. From the survey, the importance of the decision tree algorthms is highlighted. The performance of ID3, C4.5, CART decision tree algorithms is explained in the article. These algorithms are useful in classification problems. These algorithms are used to construct the decision tree. In this process, the speed of the construction of the tree is the critical parameter. The performance of the algorithms in terms of speed is also presented in the said survey. Accordingly the C4.5 is the fastest algorithm among the given algorithms. This survey is useful for the researchers intending to deal with classification problem using decision tree.

In [11], Khairsat and et. al have presented the survey of intrusion detection systems. The paper has outlined the Intrusion Detection Systems with their advantages and disadvantages, data sets, and different challenges in IDS model development. The IDS systems for zero - day attack identification are reviewed in the survey. It is found, giving poor accuracy for new attacks detection. The survey has also examined the data sets and their effectiveness. The datasets used by different researchers for generating testing results for the IDS model does not consist of the new attacks. These datasets are developed in 1999, where are very old for testing new IDS systems developed in the recent years e.g. DARPA, KDD99 etc. Hence, the use of such old datasets leads to the inaccurate claims for effectiveness of the IDS systems and results can not be considered as genuinely accurate results.

In [12], T. Saranya and et. al have presented a survey about performance analysis of machine learning techniques in intrusion detection system. It briefs out various machine learning techniques and algorithms. It also explains the intrusion detection system in different application areas and their implementation using machine learning algorithms. The survey results states different IDS implementations using machine learning techniques like naïve bayes, decision tree, SVM, kNN, k Means, Deep learning, ensamble learning, ANN, DBN etc. It is important to note here, that machine learning techniques are useful techniques for develping effective intrusion detection system model.

In [13], P. Maniriho and et. al have presented the survey of machine learning techniques for anomaly based intrusion detection system. The generic model of the IDS using machine learning techniqes is explained, followed by the implementations of IDS using different machine learning techniques with the results. The results are generated using WEKA tool. The comparison is given among Random forest, decision stump, naïve bayes and SGD algorithms. The conclusion is that, Random forest generates best results for intrusion detection as compared to other techniques used in the review. It is important to note here that, the methods and data set used is suitable for known attack detection, hence unknown attack detection is important but not considered here. That's why, the results can not be generalized for new attack types, other than those which are not present in the data set. Also, this survey does not outline about the time performance of the algorithms in IDS, which is another important property to be considered in intrusion detection.

In [14], Bahlali and et. al have carried out the research work of anomaly based intrusion detection using machine learning techniques. The implementations and comparison of three machine learning techniques like logistic regression, decision tree and random forest, along with ANN in deep learning technique is presented in the work. The dataset USNW-NB15 is used, that suffers from issues such as imbalanced classes. Still, the accuracy claimed in the results by using these classifiers, is good. Among the used algorithms the ANN is found to be the best algorithm for accuracy of IDS model. The problem with the IDS model is performance in terms of timeliness, speed is not considered in the work. Also, the results can't be genuinely considered to compare the models, as dataset is old and does not reflect the new attacks.

In [15], F. Yihunie and et. al have presented the work of anomaly intrusion detection using machine learning techniques. The different five classification techniques are used in camparison with each other as a part of model. The techniques used are: SGD, logistics regression, random forest, SVM and sequential model. The data set used for training and testing of the models is NSL KDD dataset. The results have shown that, random forest has prvided the better accuracy as compared to other algorithms. It is important to note that, this accuracy is valid only for known attacks detection. This work does not consist of unkown attack detection, which is the important thing to be considered. Also, the work does not talk about the speed of the system, timeless in detection and also the fault tolerance nature of IDS. Hence, this IDS system needs improvement considering these properties, and model should be enhanced with better techniques.

In [16], Zamani Mahdi has provided the detail study of design of intrusion detection system using machine learning techniques. The machine learning techniques are divided into two parts: Artificial Intelligence and Computational Intelligence. These methods share many features in common. It is claimed in the study that, an effective intrusion detection system can be designed using machine learning approach. It can allow to design a system which will be accurate, falt tolerant, efficient etc.

In [17], Mazini M. and et. al have proposed the anomaly based intrusion detection using machine learning approach. The methods used for the design of this intrusion detection system is Artificial Bee Colony (ABC) and AdaBoost algorithms. The approach is sequential approach for the execution. The dataset used for testing of the system is NSL KDD dataset. The results obtained, are showing the better figures as compare to legendary methods of machine learning. But, in this approach, the feature selection is used, which ignore many features based on the inappropriate values of the features, and importance of the feature is not the criteria. Also, only classification of the known attack samples is given, hence accuracy can not be generalized for unknown malicious behaviours. The approach is implementated in the simulation environment, where assumptions may not be suitable in real time environment. The efficiency or time complexities are given in sequential execution, which are not suitable to achieve the timeliness in the current systems as the data and connection requests per second are huge in numbers.

In [18], Haq Nutan and et. al have presented the useful survey of machine learning techniues for intrusion detection systems. The study gives idea about, number of attempts of IDS design using machine learning approach. Few IDS are designed using single machine learning technique, few are designed sing hybrid methods and few are designed using ensamble approach. It also says, dataset used, plays important role in generating results, irrelevant and redundant features should be removed properly,best algorithm of feature selection should be used to avoid slowness of the system. And finally, it concludes that, hybrid approaches of implementation gives good results as compared to single approaches.

In [19], Amouri A have presented the intrusion detection system using machine learning techniques for mobile Internet of Things. The model is designed of a regression techniques and feature selection is used for removing redundant, irrelevant features from the dataset. The results are generated using simulation based environment. The accuracy observed in detection is claimed as good accuracy, but major concern is false positive rate (FPR), couldn’t be kept at minimum and it has shown variation to a concern level. This model is designed for IoT systems, hence can not be generalized for the different network types.

In [20], Dr.S.Malliga and et. al have presented the network intrusion detection system for IoT systems using machine learning and deep learning algorithms. Naïve Bayes is used as a machine learning technique, in comparison with ANN and RNN as deep learning techniques. It is concluded that, Accuracy achieved using deep learning techniques like ANN and RNN is more than accuracy achieved using machine learning approach of naïve bayes. The time requierement is also mentioned. But, here the important issue is, in deep learning, the amont of data required is huge in size and it has more time complexity of training as compared to machine learning techniqes. Also these ANN and RNN techniques are blind folded techniques, the base for decision making in the model, is not transparent. Also, the model presen, is for IoT and a sequential model.

In [21], Labonne and et. al have presented a survey study of intrusion detection system using machine learning techniques. It briefs out the work that has been carried out over the years by different researchers. The study reveals that, better solution is possible with machine learning in intrusion detection. It is also concluded that, NSL KDD dataset is better than KDD99 data set and is used by many researchers also.

In [22], Aslam and et. al have presented the hybrid approach for network intrusion detection system using machine learning approach and rules based system. The machine learning techniqes like decision tree, Seqential Minimal Optimization and simple logistic is used for the hybrid approach with rule based system. The accuracy observed is for known attacks from the dataset and hence can not be generalized for unknown attack identification. Also, model does not consider important properttis like speed, timeliness and fault tolerance of the IDS system.

In [23], A. Halimaa and et. al have presented the intrusion detection sysem model using machine learning approach. Two machine learning classifiers are used, SVM and Naïve bayes classifiers. It is concluded that, SVM provides good accuracy as compared to naïve bayes. Feature selection is used for selecting useful and relevant features for model development. These accuracy results are valid only for known attack detection and can not be considered in unknown attack detection systems. The timeliness is also the major area not addressed in this work.

In [24], YUJUN YANG and et. al have presented the work of support vector machine (SVM) and its performance. It is observed from the experiments that presented the work of support vector machine and its performance. It is observed from the experiments that SVM is faster machine learning technique. The analysis is done using three kinds of indicators sensitivity, specificity, time-consumption.

In SVM performance evaluation using different data sets, and different kernels linear,polynomial, sigmoid and RBF, RBF kernel gives good performance[25]. Also, the results observed are encouraging as compared to other techniques. The SVM provides better results for limited data set size, but in case of huge data size, time complexity is more[26]. Also, the SVM is not so popular for the imbalanced data and unkown labelled data[26]. Hence, the models can not be developed with stand alone SVM method.

In [27], Agarap and et. al have designed the intrusion detection system using neural network architecture with Gated Recurrent Unit (GRU) and Support Vector Machine

(SVM). In the presented model the SVM is used at a output layer if GRU-RNN, SVM is faster in terms of time complexity. The training and testing is done only with binary classification for SVM. The results should be generated with mlticlassification and that is one of the important things. Also, the classification using SVM gives better results for known attack traffic and should be verified with unknown attack traffic.

In [28], Jayshree Jha and et. al have designed the intrusion detection system using Support Vector Machine (SVM). The data set used for the model training and testing is NSL KDD dataset. With, feature selection, best features selected based on k Means algorithm criteria. With the reduction in the number of features, the time required for execution of the SVM is also reduced. But, the disadvantage is, SVM is a classification technique, which requires, prior knowledge of events and data. Hnece, in intrusion detection systems, SVM model is useful only in case of known attack detection and fails to detect unkown attacks attempts.

SVM gives good performance, when the effective feature selection algorithm is used. In [29], for cloud intrusion detection system (CIDS), the SVM is used to build a model with the Correlaion based feature selection (COFS), which helped to achieve better results.

In [30], S. Teng have presented the CAIDM model named as Collaborative and Adaptive Intrsion detection model using machine learning classification techniques, SVM and decision tree. The KDDCUP99 dataset is used for training and testing of the CAIDM model and it is concluded that, this model with the use of SVM and DT generates better results than using the single SVM for instrusion detection system. Also, the important point to note here, is that, both of these methods in CAIDM are classification methods, which are useful for prior knowledge attcaks and not for identification of new attack type input. Also, timeliness is not considered for the model execution.

In [31], Y. Chang and et. al have presented the network intrusion detection using the support vector machine and random forest. The random forest used for feature selection to increase the accuracy and time required for execution. By using, Random forest out of 41, 14 features are selected, which gave the good accuracy as compared to accuracy by using 41 attributes. These results are for known type of input and hence, accuracy can not be generalized for new attack type input.

In [32], T. Cover and et. al have presented the nearest neighbor pattern classification technique called as a k-NN classification. The data point is classified based on the distance between the other data points of other classes. The data point is classified to the class of the data point, whose distance is minimum as compared to other data points.

The work states that, the error is minimum for 1-NN and same can be implied for k-NN classification with minimized error. It is the classification technique, which is faster pattern classification technique.

In [33], Imandoust and et. al have proposed a study of kNN for prediction of economic events. The same method is used for regression by assigning the value of the property to the object property by calculating the average of the propert value for all its neighbors. Hence, kNN is used as efficient method for prediction of the economic events. The kNN is a faster algorithm and very useful, even in the absence of the prior knowledge of the events.

In [34], Ali N. and et. al have presented the detail study about performance of k nearest neighbor algorithm for heterogenous data sets. The kNN performance is evaluated using eucledian distance and manhatten distance formulas. It is found that, the eucledian distance formula does not provide better kNN results. Also, for heterogenous data sets, not much difference is observed in the performance of the kNN algorithm.

In [35], Benaddi Hafsa and et. al have proposed the robust intrusion detection system model using Principal Component Analysis (PCA)- Fuzzy Clustering- kNN. NSL KDD data set is used for preparing the model of IDS. It is concluded that, the important thing is to reduce the set of features in the data set to achieve the desired performance of the IDS. With the presence of kNN, it was possible for the model to classify the different attack types input effectively. But, the problem is the accuracy of the kNN should be verified properly. Also, the time complexity of the system is the area of concern in the given IDS model.

In [36], L. Li and et. al have presented the intrusion detection system based on binary classification and kNN. The model is divided into two steps. The first step is related to binary classification to detect the abnormal connections and in second step, kNN is used for detecting unusual and new input types. It is concluded that, when kNN is used, IDS model found accurate than the single binary classification in IDS. But, the accuracy isn’t verified after step 2, as it is required for kNN and also the timeliness is not tested in the given results, which can not be ignored in the IDS model.

In [37], Wenchao Li and et. al have presented the intrusion detection model for wireless sensor networks using the kNN classification. The attack type targeted for detection is flooding attack types. Model is proved to be effective in detecting the flooding attack types using kNN classification. But, the problem is, alone kNN is not effective when different attack types or new attack types are to be detected.

In [38], M. Kumar and et. al have presented the intrusion detection system using the decision tree. The results obtained are having better figures. The decision tree uses the pre existing knowledge for model building and gives better results for the known input types. For, unknown input types, it can not perform well and should be used with another methods to achieve better results.

In [39], Panda and et. al have presented the intrusion detection system using naïve bayes technique. The results obtained by the model are better than the neural network architectures. The model is built with two layers and distance between the information nodes is minimized. The results also shown that naïve bayes approach gives good results in less time and with low cost. But, the drawback of the system is, it generates more false positives as compared to other systems. Hence, it can be concluded that, naïve bayes should be used along with other technique for better results and reduction in false positives.

In [40], B. S. Sharmila and et. al have presented the intrusion detection system using PCA base naïve bayes algorithm. The results have shown that, better accuracy is achieved with PCA based naïve bayes algorihm as compared to traditional naïve bayes algorithm. This approach also helps to provide useful results even in presence of the missing values in the data sets. However, with increasing size of the data, the accuracy is decreased and speed of the system also slow downs. Hence, naïve bayes can be used with other techniques to achieve better results.

III. DEFINITIONS AND TECHNIQUES

From the survey, it is found that, using one of the methods in intrusion detection system is not sufficient for achieving better accuracy in detecting attacks. So, we are designing the intrusion detection system model using the combination of different machine learning technques and distributed processing architecture as follows:

1. Support Vector Machine (SVM)
2. K Nearest Neighbor (kNN)
3. Decision Tree (DT)
4. Naïve Bayes (NB)
5. HDFS

1. **Support Vector Machine**: Support vector machine which is abbreviated as SVM, is the Supervised Machine learning technique, used for classification of the data. In Support vector machines, the reproducible hyperplane is produced, which maximizes the margins between the classes[41]. These margins include the boundary points of the classes, which are called as support vectors. Support vectors and hyperplane help to

classify the data points into separate classes. The diagram showing support vectors and hyperplane separation is as shown in fig. 1 as follows.

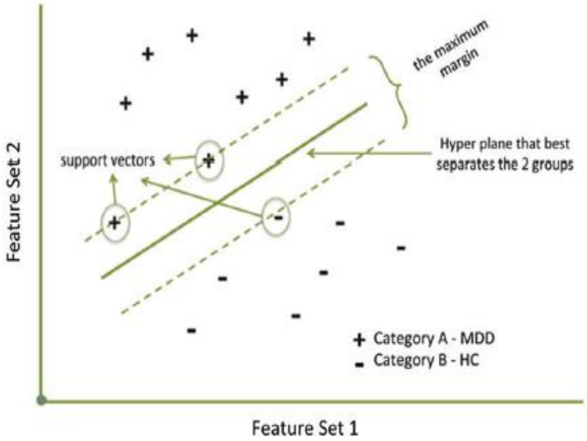


Fig. 1: Support Vector Machine

In intrusion detection system, the data can be classified into two different classes by hyperplpane. First class is the set of points with normal behavior i.e. normal connection request from genuine nodes, which can be categorized as normal input traffic. Second class is the set of points with unwanted behavior i.e. malicious connection request from malicious nodes, which can be categorized as intruder nodes. This is the effective machine learning technique to achieve accuracy as well as timeliness in the classification.

2. **K Nearest Neighbor**: k nearest neighbor is a classification technique used to classify the data points based on the distance between the data points. The distance formulaes that can be used are Eucledian distance (mostly used) [42], manhatten distance formulaes. Many times, it is important to carry out the data preprocessing before training the kNN model[42]. In next step, the distance between the data point to be classified and the training data points is calculated[42]. After, finding distance to all data points, the distance array is sorted and first k distance values are selected. Finally, the data point will be classified to the class, whose maximum data points are present in first k distance values array, as these are the closest point with minimum distance from the data point to be classified.

In intrusion detection system, every incoming input traffic behavior is matched with the available data points in normal behavior and malicious behavior, and that wll be categorized to the class, where the similarity is more and more points are closer to the new data point to be categorized. kNN algorithm is suitable in intrusion detection because of following important advantages:

a. Accuracy
b. Time Complexity/Calculation Time is less
c. Easy interpretation of the output

d. Predictive Power of kNN

3. **Decision Tree:** Decision Tree is the classification technique, which unlike SVM, helps for multiclassification. Decision trees are also called as classification trees. The classification trees are used to classify the data points into classes which are belonging to the response variable[43]. Decision trees are generated with the tree structured nodes, where the internal nodes are the conditional attribute nodes, based on which the path to the class is selected and the leaf nodes the class nodes which are the classification nodes or categorical nodes for the data points. The example of host based IDS decision tree is as shown in fig. 2 [44].

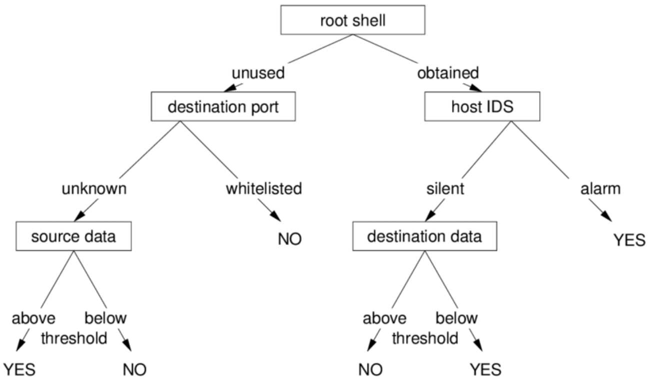


Fig. 2: Sample host based IDS decision tree

Herein, the decision tree is very helpful for the intrusion detection system, as the attack type can be identified with the categorical or leaf nodes of the decision tree. Whenever, unknown attack input is encountered, the new leaf node will be created with the conditional attributes path and the decision tree is continuously updated for new attacks also.

4. **Naïve Bayes:** Naïve Bayes is a classifier which provide better accuracy with less computational and storage requirements for classification of the data[45]. Naïve Bayes classification is useful when the overall probability is the product of the independent probabilities of the variables. In naïve bayes, it is assumed that, the features are independent of each other and they contribute to the classification independently. Naïve bayes is effective to use when a large size data sets to deal with and also, it is easy to implement.

Fig. 3 shows the probability expression for the naïve bayes classification. It states that, there are two probabilities which are required to calculate the final probability of the object being classified to a specific class. The prior probability, likelihood and posterior probabilities are as shown in fig. 3.

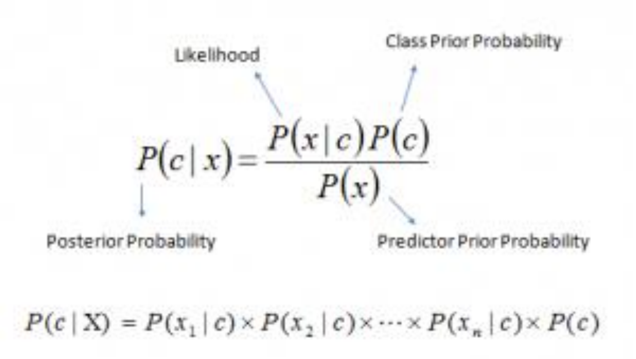

Likelihood
Class Prior Probability
$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X)=P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$

Fig. 3: Naïve Bayes Probability Concept

The posterior probability is the probability of getting the object classified to that class. Here, in the technique, the posterior probability is calculated for every class for the given object and will be classified to the class with the maximum posterior probability.

5. **HDFS:** HDFS is a part of Hadoop. It is called as Hadoop distributed file system, which is very popular because of two important things- fault tolerance and distribted & parallel data processing architecture[46]. Hadoop consists of two important things namenode and datanodes. Each name node acts as a master and keeps the metadata of the data nodes and manages the several number of data nodes typically thousand in nmbers. Such number of namenodes can be created in Hadoop. With this architecture, data can be divided into multiple blocks, and stored at multiple data nodes. This distributed data storage enables the parallel processing of the data, without any restriction on size of data processing. Hadoop is the massively parallel data processing architecture. The simple HDFS architecture is as shown in fig. 4[46]:

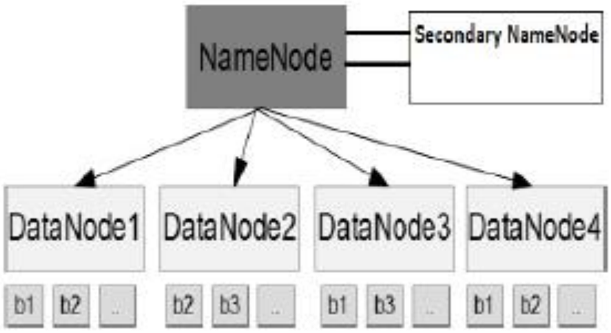


Fig. 4: HDFS Architecture

In intrusion detection system, the fault tolerance and parallel processing to reduce the time complexity of the execution is very important feature, that can be easily achieved with the Hadoop or HDFS system, by using HDFS underlying data storage and processing architecture for all machine learning techniques.

IV. OBJECTIVES

From, the literature study, few important things are uncovered, of which the solution in real time can be provided with different machine learning techniques. To develop the solution, the following objectives are defined and achieved:

1. To design the appropriate intrusion detection system architecture with the help of machine learning techniques.
2. To develop the two phase intrusion detection system for increasing accuracy of the intrusion detection.
3. To implement the intrusion detection model, which can detect the malicious activities in timeliness manner.
4. To generate the fault tolerant intrusion detection system by using machine learning techniques and HDFS.

The objective are defined by considering the current security requirements of the organizational networks. The effective and efficient intrusion detection system implementation is the overall objective of this research work.

V. SOLUTION METHODOLOGY

The model for anomaly intrusion detection system is presented in this context. The model is designed using machine learning techniques. From the survey study, it is observed that, the intrusion detection system model designed using either of the machine learning techniques, does not provide desired features alltogether. With single technique as a part of the IDS model, disadvantages of the technique limits the performance potential of the system. Hence, it is required to bring multiple machine learning techniques together, which are complimentry to each other & design the intrusion detection system model which can provide all the required features as well as performance without limiting its performance potential.

So, as per the research gap, the four different machine learning techniques such as Support Vector Machine(SVM), k Nearest Neighbor (kNN), Decision Tree (DT), Naïve Bayes (NB) found useful and important to be a part of the accurate, efficient intrusion detection system along with massively parallel distribted data storage & processing system such as HDFS. The architecture of the system is as shown in fig. 5. The architecture consists of two phases, in which first phase is for identifying the type of input, whether it is a normal input or anomaly input. The methods used for Phase I are SVM and kNN. SVM is the best method for known attack detection and it has high accuracy rates. kNN is a method which acts like a clustering technique, even for unknown or new input, it can provide faster and approximately accurate anomaly detection. Hence, here though SVM couldn't identify the unknown input sample correctly, kNN will provide idea about the type of input based on the similarity criteria, with more number of minimum distance points from a specific category or class to classify the input type. If, either or both of the SVM & kNN detects the incoming input traffic as anomaly, then the access will be blocked and the input traffic information will be passed to Phase II of the TP-IDS. Else, if both of these techniques identify the input traffic as normal connection request, then connection request will be accepted & access will be given. In this case, Phase II will not be executed.

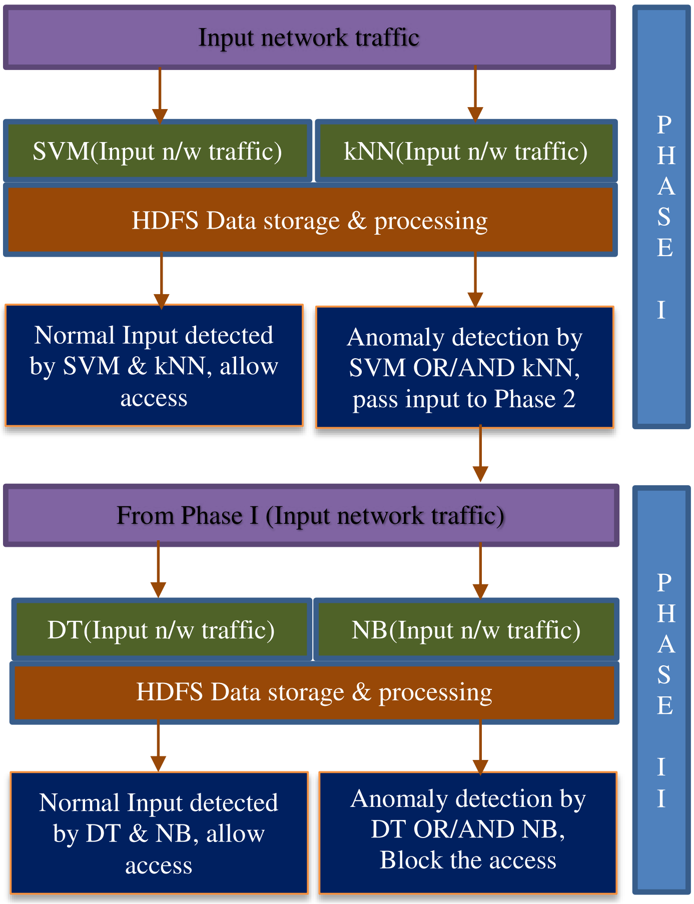

Input network traffic
SVM(Input n/w traffic)
kNN(Input n/w traffic)
HDFS Data storage & processing
Normal Input detected by SVM & kNN, allow access
Anomaly detection by SVM OR/AND kNN, pass input to Phase 2
PHASE I
From Phase I (Input network traffic)
DT(Input n/w traffic)
NB(Input n/w traffic)
HDFS Data storage & processing
Normal Input detected by DT & NB, allow access
Anomaly detection by DT OR/AND NB, Block the access
PHASE II

Fig. 5: TP-IDS Architecture

In Phase II of the TP-IDS, the machine learning techniques used are decision tree and naïve bayes. Here, the phase II is used as a validation Phase of TP-IDS. Here, again if either or both of the Decision tree and naïve bayes detects the input traffic as anomaly, then the final output will be anomaly with type of attack description, and network access will be blocked. Else, if both of these methods of phase II identify the input traffic as normal input, then network connection request will be accepted and access will be given.

Here, the Phase I and Phase II techniques are executed in underlying Hadoop distributed file system, which is distributed hence parallel & faster, fault tolerant. Also, with Phase II validation, we increase the accuracy & decrease the False Positive Rate (FPR) as well as False Negative Rate (FNR). By using, HDFS the two techniques used in each phase will be executed in parallel with parallel input processing in each technique. This enables to save the processing time and detection time of the TP-IDS model. This architecture helps to achieve the desired performance of the TP-IDS.

VI. DATASET AND DATA PRE PROCESSING

The important thing to achieve the accuracy with any machine learning model, is the structure of the data set & quality of the data in the data set. Quality dataset is the data set, where noise is not present in the data values. Apart from this, it is also important that, only relevant attributes or features should be selected to train the model. When the irrelevant and redundant attributes are removed, the remaining set of the relevant features helps to increase the accuracy of the model.

In this intrusion detection model TP-IDS, the data set used is NSL KDD data set. The data sets prior to NSL KDD data set like KDD99, are the data sets which are unbalanced data set and affects the quality of the output during testing phase of the machine learning model. Hence, NSL KDD is the most popular data set now a days & it is used by many of the researchers who worked in the area of intrusion detection systems. NSL KDD data set consists of total 43 attributes, where one of the feature is attack type. It is the 42nd feature of the data set. It is the output variable for TP-IDS.

One of the important thing in the model training is data pre processing. In the data pre processing task, the feature selection method is used. There are many different feature selection techniques like Single feature selection technique(SFST), Random feature selection technique (RFST), Coorelation based feature selection technique (CFST). Among these techniques, Correlation based feature selection technique is found to be effective and useful in this application scenario of TP-IDS. Hence, the Correlation based feature selection is used for removing the redundant and irrelevant features from the dataset. In correlation based feature selection, the heauristic function evaluation is used to find the correlation[47]. The equation is as follows:

$$M_S = \frac{l\overline{t_{cf}}}{\sqrt{l + l(l-1)\overline{t_{ff}}}},$$

The features are considered irrelevant, when the correlation value is less than the threshold. When the value is more then it is relevant feature and can be considered in the model training.

With CFST, the 29 features are selected as relevant features with the attack type feature of NSL KDD & same are used for training & testing of the TP-IDS model. The same is shown in fig. 6.

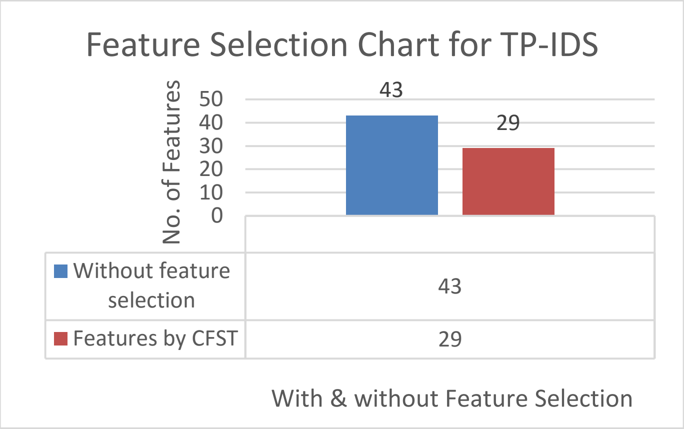


Fig. 6: Feature Selection Chart

VII. RESULTS AND DISCUSSION

The TP-IDS model is implementated using R programming and Hadoop underlying distributed file system. The R programming is one of the popular statistical programming langage specifically designed for data analytics and machine learning model building. Hadoop distributed file system has helped to increase the speed of the data processing by huge factor and could generate the excellent results.

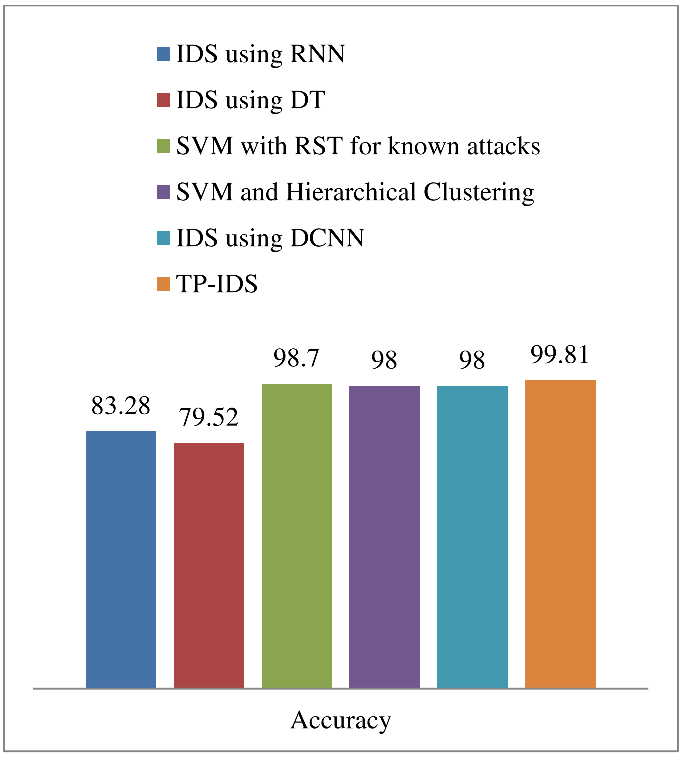


Fig. 7: Accuracy results comparison

In fig. 7 the accuracies of different IDS models are given in comparison with the TP-IDS model. It is observed that, with the hybrid model concept, when different complimentary machine learning techniques are used, the highly accurate model is generated. The results are generated in different test conditions like known and new input values also.

It is also observed that, the output categorizations are generated in quick time expected, because of the use of Hadoop distributed file system. It enables the parallel and faster data processing in TP-IDS model. The model is trained by using all samples given in the NSL KDD training data set i.e. total 1,25,973 records of input are passed for training the TP-IDS model. One of the reason for getting such high accuracy is also the size of the data input and the features passed to the model. The quality data has enabled to generate the highly accurate TP-IDS model.

VIII. CONCLUSION AND FUTURE WORK

So, hereby it is studied that, machine learning can give better results in building a models for intrusion detection systems. When different machine learning techniques are combined together to overcome the disadvantages of each other, one of the better intrusion detection systems like TP-IDS can be designed. By using the two phase model, with second phase as validation phase, the false positive rate and false negative rate is reduced by much amount and accuracy is increased by better scale.

Also, with the use of distributed processing data architecture like HADOOP-HDFS, the processing speed of the system is increased by massive amount, hence it helped to achieve the timeliness in the system TP-IDS. One of the important requirement such as fault tolerance is also achieved with the help of such distributed architecture. In future, the methods can be replaced with different methods from different machine learning techniques such as unsupervised or semi supervised category & new model with improved performances can be achieved.

ACKNOWLEDGEMENT

We are very much thank ful to our respected research committee members and coordinators, and respected research dean of the Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India. We are also thankful to all our colleagues fo our Institute and others who directly or indirectly helped us through out our research work.

AUTHORS' CONTRIBUTIONS

All authors have made substantial contributions to the article. Abhijit has done the literature, proposed model and writing this articles, where Dr. Vidyullatha has provided useful guidance throughout the research.

FUNDING

This research is funded by authors only.

Availability of data and materials

All data generated or analyzed during this study are included in this published article. The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable. This research does not involve any human participants, human data, or human tissue.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

Abhijit D. Jadhav[1], Research Scholar at Department of Computer Science & Engineering[1], Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India[1] and Assistant Professor, Department of Computer Engineering[1], Dr. D. Y. Patil Institute of Technology, Pimpri, Pune-18.[1] Dr. Vidyullatha Pellakuri[2] Associate Professor at Department of Computer Science & Engineering[2], Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.[2]

References

[1] Debar, H., Dacier, M. & Wespi, A. A revised taxonomy for intrusion-detection systems. Ann. Télécommun. 55, 361–378 (2000). https://doi.org/10.1007/BF02994844

[2] Venkata Ramani Varanasi, Shaik Razia, "Intrusion Detection using Machine Learning and Deep Learning", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-4, November 2019

[3] Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, Kuang-Yuan Tung, Intrusion detection system: A comprehensive review, Journal of Network and Computer Applications, Volume 36, Issue 1, 2013, Pages 16-24, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2012.09.004.

[4] Juan M. Estevez-Tapiador, Pedro Garcia-Teodoro, Jesus E. Diaz-Verdejo, Anomaly detection methods in wired networks: a survey and taxonomy, Computer Communications, Volume 27, Issue 16, 2004, Pages 1569-1584, ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2004.07.002.

[5] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, E. Vázquez, Anomaly-based network intrusion detection: Techniques, systems and challenges, Computers

& Security, Volume 28, Issues 1–2, 2009, Pages 18-28, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2008.08.003.

[6] J. B. D. Cabrera, J. Gosar, W. Lee and R. K. Mehra, "On the statistical distribution of processing times in network intrusion detection," 2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601), Nassau, Bahamas, 2004, pp. 75-80 Vol.1, doi: 10.1109/CDC.2004.1428609.

[7] M. A. Rassam, M. A. Maarof and A. Zainal, "A novel intrusion detection framework for Wireless Sensor Networks," 2011 7th International Conference on Information Assurance and Security (IAS), Melacca, Malaysia, 2011, pp. 350-353, doi: 10.1109/ISIAS.2011.6122778.

[8] P. Amudha, S. Karthik, S. Sivakumari, "A Hybrid Swarm Intelligence Algorithm for Intrusion Detection Using Significant Features", The Scientific World Journal, vol. 2015, Article
ID 574589, 15 pages, 2015. https://doi.org/10.1155/2015/574589

[9] Hervé Debar, Marc Dacier, Andreas Wespi, Towards a taxonomy of intrusion-detection systems, Computer Networks, Volume 31, Issue 8, 1999, Pages 805-822, ISSN 1389-1286, https://doi.org/10.1016/S1389-1286(98)00017-6.

[10] Zeeshan Ali Khan, Peter Herrmann, "Recent Advancements in Intrusion Detection Systems for the Internet of Things", Security and Communication Networks, vol. 2019, Article
ID 4301409, 19 pages, 2019. https://doi.org/10.1155/2019/4301409

[11] Khraisat, A., Gondal, I., Vamplew, P. et al. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecur 2, 20 (2019). https://doi.org/10.1186/s42400-019-0038-7.

[12] T. Saranya, S. Sridevi, C. Deisy, Tran Duc Chung, M.K.A.Ahamed Khan, Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review, Procedia Computer Science, Volume 171, 2020, Pages 1251-1260, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.04.133.

[13] P. Maniriho and T. Ahmad, "Analyzing the Performance of Machine Learning Algorithms in Anomaly Network Intrusion Detection Systems," 2018 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2018, pp. 1-6, doi: 10.1109/ICSTC.2018.8528645.

[14] Bahlali, Ahmed Ramzi. (2019). Anomaly-Based Network Intrusion Detection System: A Machine Learning Approach. 10.13140/RG.2.2.29553.84325.

[15] F. Yihunie, E. Abdelfattah and A. Regmi, "Applying Machine Learning to Anomaly-Based Intrusion Detection Systems," 2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, NY, USA, 2019, pp. 1-5, doi: 10.1109/LISAT.2019.8817340.

[16] Zamani, Mahdi. (2013). Machine Learning Techniques for Intrusion Detection.

[17] Mazini, M., Shirazi, B., Mahdavi, I., Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms, Journal of King Saud University – Computer and Information Sciences (2018), doi: https://doi.org/10.1016/j.jksuci.2018.03.011.

[18] Haq, Nutan & Avishek, Md & Shah, Faisal & Onik, Abdur & Rafni, Musharrat & Md, Dewan. (2015). Application of Machine Learning Approaches in Intrusion Detection System: A Survey. International Journal of Advanced Research in Artificial Intelligence. 4. 10.14569/IJARAI.2015.040302.

[19] Amouri A, Alaparthy VT, Morgera SD. A Machine Learning Based Intrusion Detection System for Mobile Internet of Things. Sensors (Basel). 2020;20(2):461. Published 2020 Jan 14. doi:10.3390/s20020461.

[20] Dr.S.Malliga, S.Darsniya, P.S.Nandhini. A NETWORK INTRUSION DETECTION SYSTEM FOR IoT USING MACHINE LEARNING AND DEEP LEARNING APPROACHES. International Journal of Advanced Science and Technology. Vol. 29, No. 3s, (2020), pp. 1017-1023.

[21] Labonne, Maxime. (2020). Anomaly-based network intrusion detection using machine learning.

[22] Aslam, Urooj & Batool, Ezzat & Ahsan, Syed & Sultan, Abdullah. (2017). Hybrid Network Intrusion Detection System Using Machine Learning Classification and Rule Based Learning System. International Journal of Grid and Distributed Computing. 10. 51-62. 10.14257/ijgdc.2017.10.2.05.

[23] A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 916-920, doi: 10.1109/ICOEI.2019.8862784.

[24] Yang, Yujun & Li, Jianping & Yang, Yimei. (2015). The research of the fast SVM classifier method. 121-124. 10.1109/ICCWAMTIP.2015.7493959.

[25] Srivastava, Durgesh & Bhambhu, Lekha. (2010). Data classification using support vector machine. Journal of Theoretical and Applied Information Technology. 12. 1-7.

[26] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, Asdrubal Lopez, A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing, Volume 408, 2020, Pages 189-215, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2019.10.118.

[27] Agarap, Abien Fred. (2017). A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data. 10.1145/3195106.3195117.

[28] Jayshree Jha, Leena Ragha, Ph.D. Intrusion Detection System using Support Vector Machine. International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868.

[29] W. Wang, X. Du and N. Wang, "Building a Cloud IDS Using an Efficient Feature Selection Method and SVM," in IEEE Access, vol. 7, pp. 1345-1354, 2019, doi: 10.1109/ACCESS.2018.2883142.

[30] S. Teng, N. Wu, H. Zhu, L. Teng and W. Zhang, "SVM-DT-based adaptive and collaborative intrusion detection," in IEEE/CAA Journal of Automatica Sinica, vol. 5, no. 1, pp. 108-118, Jan. 2018, doi: 10.1109/JAS.2017.7510730.

[31] Y. Chang, W. Li and Z. Yang, "Network Intrusion Detection Based on Random Forest and Support Vector Machine," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, China, 2017, pp. 635-638, doi: 10.1109/CSE-EUC.2017.118.

[32] T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967, doi: 10.1109/TIT.1967.1053964.

[33] Imandoust, S.B. & Bolandraftar, Mohammad. (2013). Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. Int J Eng Res Appl. 3. 605-610.

[34] Ali, N., Neagu, D. & Trundle, P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Appl. Sci. 1, 1559 (2019). https://doi.org/10.1007/s42452-019-1356-9.

[35] Benaddi, Hafsa & Ibrahimi, Khalil & Benslimane, Abderrahim. (2018). Improving the Intrusion Detection System for NSL-KDD Dataset based on PCA-Fuzzy Clustering-KNN. 1-6. 10.1109/WINCOM.2018.8629718.

[36] L. Li, Y. Yu, S. Bai, Y. Hou and X. Chen, "An Effective Two-Step Intrusion Detection Approach Based on Binary Classification and k -NN," in IEEE Access, vol. 6, pp. 12060-12073, 2018, doi: 10.1109/ACCESS.2017.2787719.

[37] Wenchao Li, Ping Yi, Yue Wu, Li Pan, Jianhua Li, "A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network", Journal of Electrical and Computer Engineering, vol. 2014, Article ID 240217, 8 pages, 2014. https://doi.org/10.1155/2014/240217

[38] M. Kumar, M. Hanumanthappa and T. V. S. Kumar, "Intrusion Detection System using decision tree algorithm," 2012 IEEE 14th International Conference on Communication Technology, Chengdu, China, 2012, pp. 629-634, doi: 10.1109/ICCT.2012.6511281.

[39] Panda, Mrutyunjaya & Patra, Manas. (2007). Network intrusion detection using naive bayes. 7.

[40] B. S. Sharmila and R. Nagapadma, "Intrusion Detection System using Naive Bayes algorithm," 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Bangalore, India, 2019, pp. 1-4, doi: 10.1109/WIECON-ECE48653.2019.9019921.

[41] Derek A. Pisner, David M. Schnyer, Chapter 6 - Support vector machine, Editor(s): Andrea Mechelli, Sandra Vieira, Machine Learning, Academic Press, 2020, Pages 101-121, ISBN 9780128157398, https://doi.org/10.1016/B978-0-12-815739-8.00006-7.

[42] João Elias Vidueira Ferreira, Clauber Henrique Souza da Costa, Ricardo Morais de Miranda, Antonio Florencio de Figueiredo, The use of the k nearest neighbor method to classify the representative elements, Educación Química, Volume 26, Issue 3, 2015, Pages 195-201, ISSN 0187-893X, https://doi.org/10.1016/j.eq.2015.05.004.

[43] Vijay Kotu, Bala Deshpande, Chapter 4 - Classification, Editor(s): Vijay Kotu, Bala Deshpande, Predictive Analytics and Data Mining, Morgan Kaufmann, 2015, Pages 63-163, ISBN 9780128014608, https://doi.org/10.1016/B978-0-12-801460-8.00004-5.

[44] Staudemeyer, Ralf. (2012). The importance of time: Modelling network intrusions with long short-term memory recurrent neural networks. 224.

[45] E.R. Davies, Chapter 13 - Basic classification concepts, Editor(s): E.R. Davies, Computer Vision (Fifth Edition), Academic Press, 2018, Pages 365-398, ISBN 9780128092842, https://doi.org/10.1016/B978-0-12-809284-2.00013-7.

[46] Mohd Rehan Ghazi, and Durgaprasad Gangodkar. "Hadoop, MapReduce and HDFS: A Developers Perspective" Procedia Computer Science, vol. 48, 2015. doi:10.1016/j.procs.2015.04.108.

[47] Agnieszka Wosiak, Danuta Zakrzewska, "Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis", Complexity, vol. 2018, Article
ID 2520706, 11 pages, 2018. https://doi.org/10.1155/2018/2520706

Figures

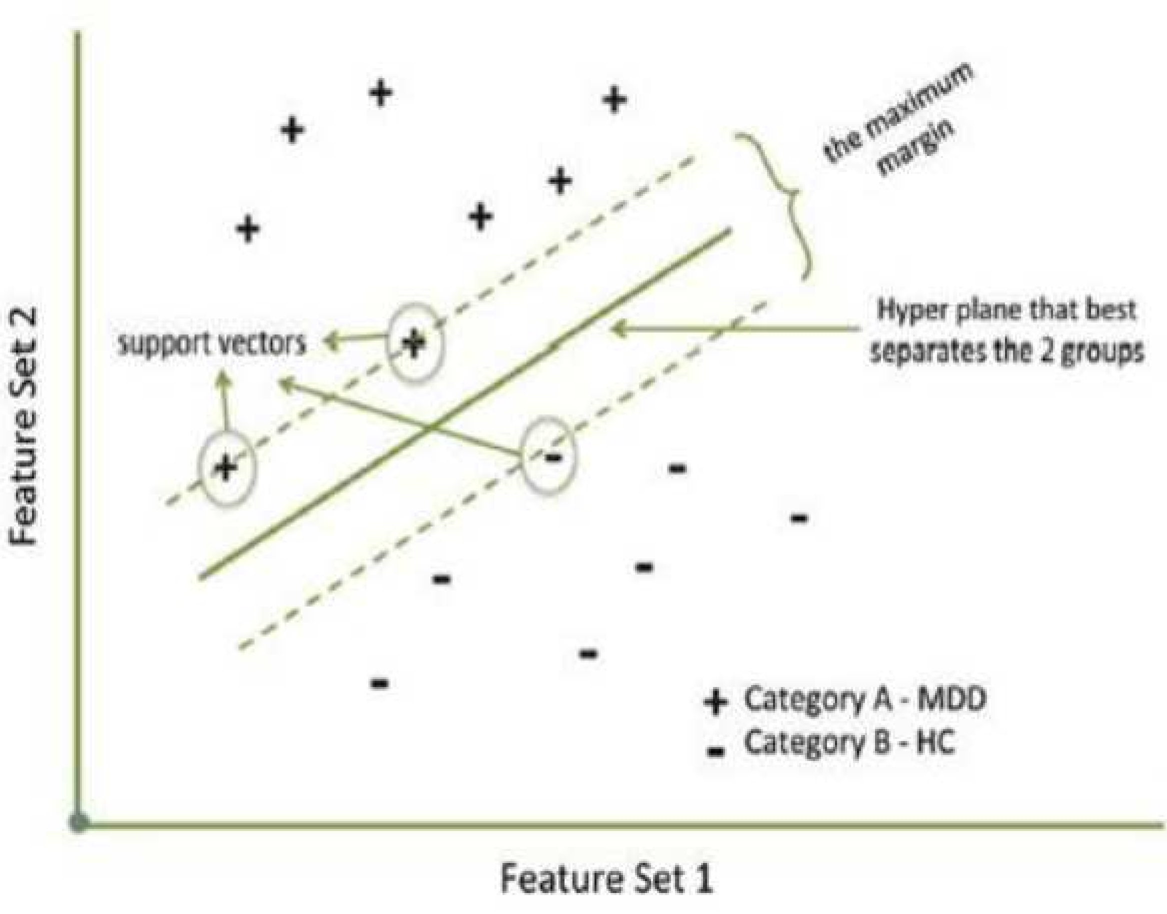


Figure 1

Support Vector Machine

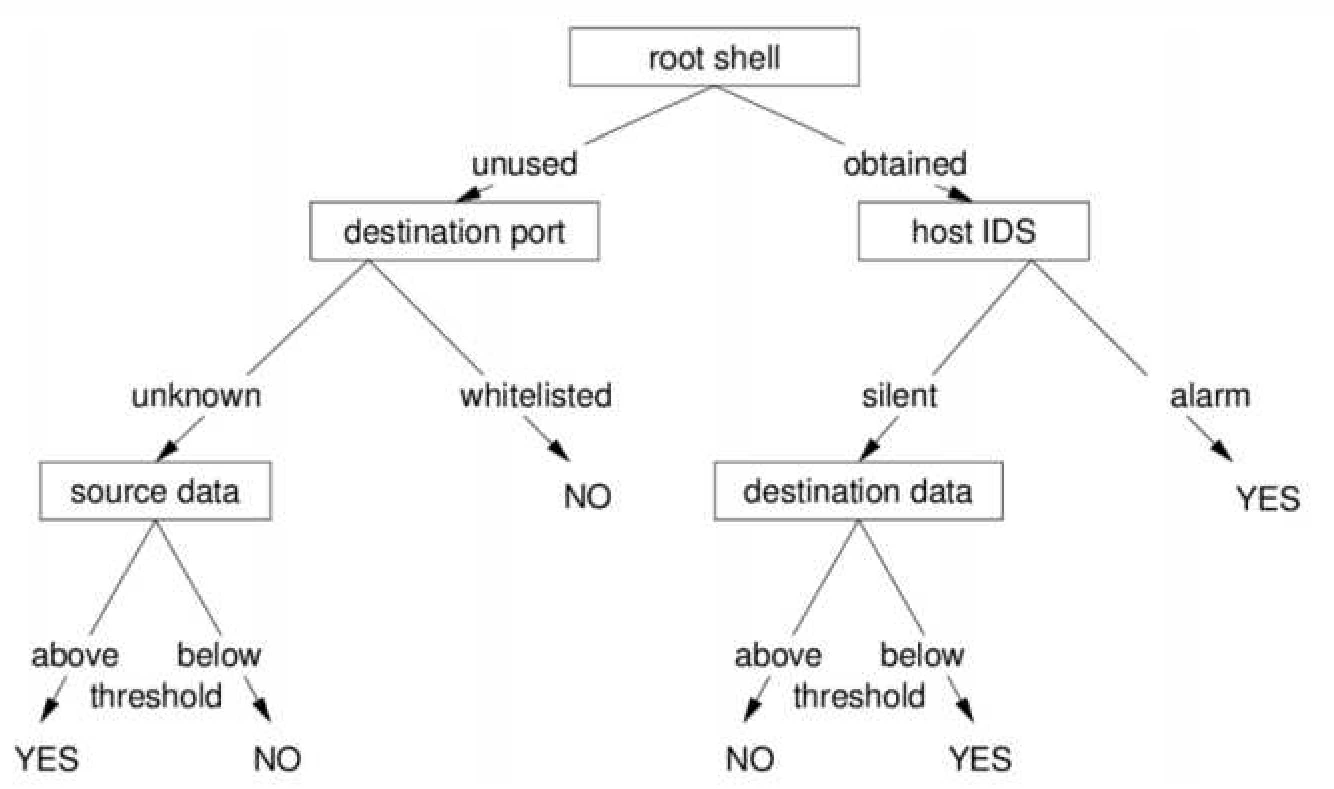
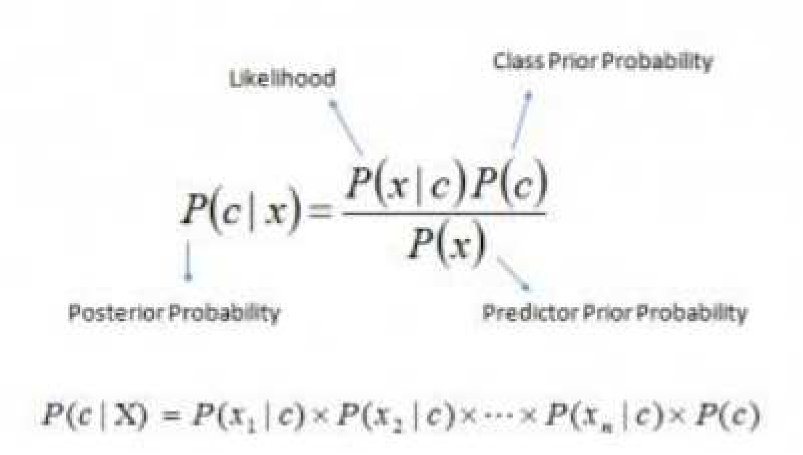


Figure 2

Sample host based IDS decision tree



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 3

Naïve Bayes Probability Concept

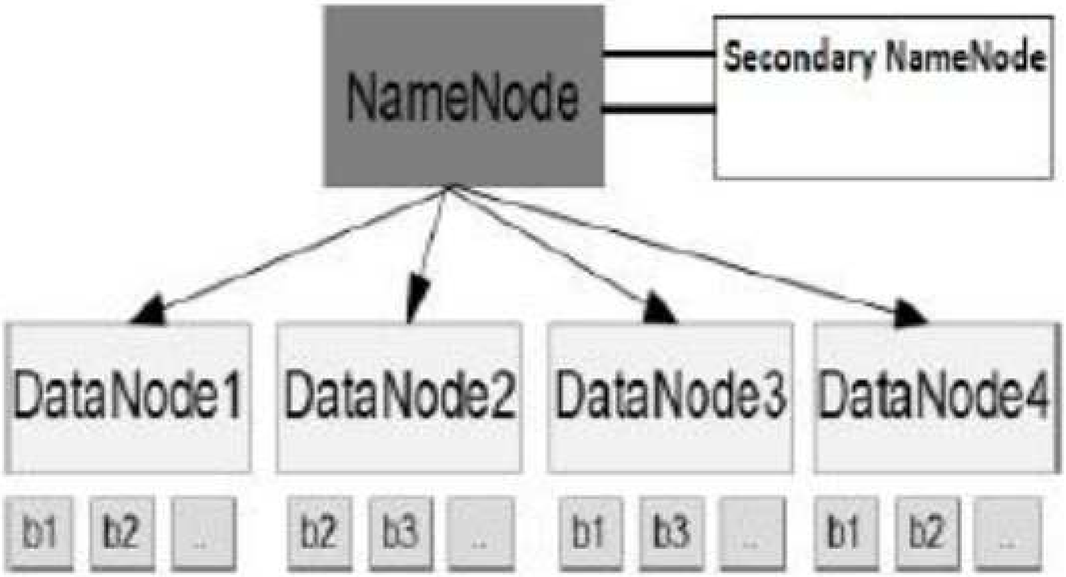


Figure 4

HDFS Architecture

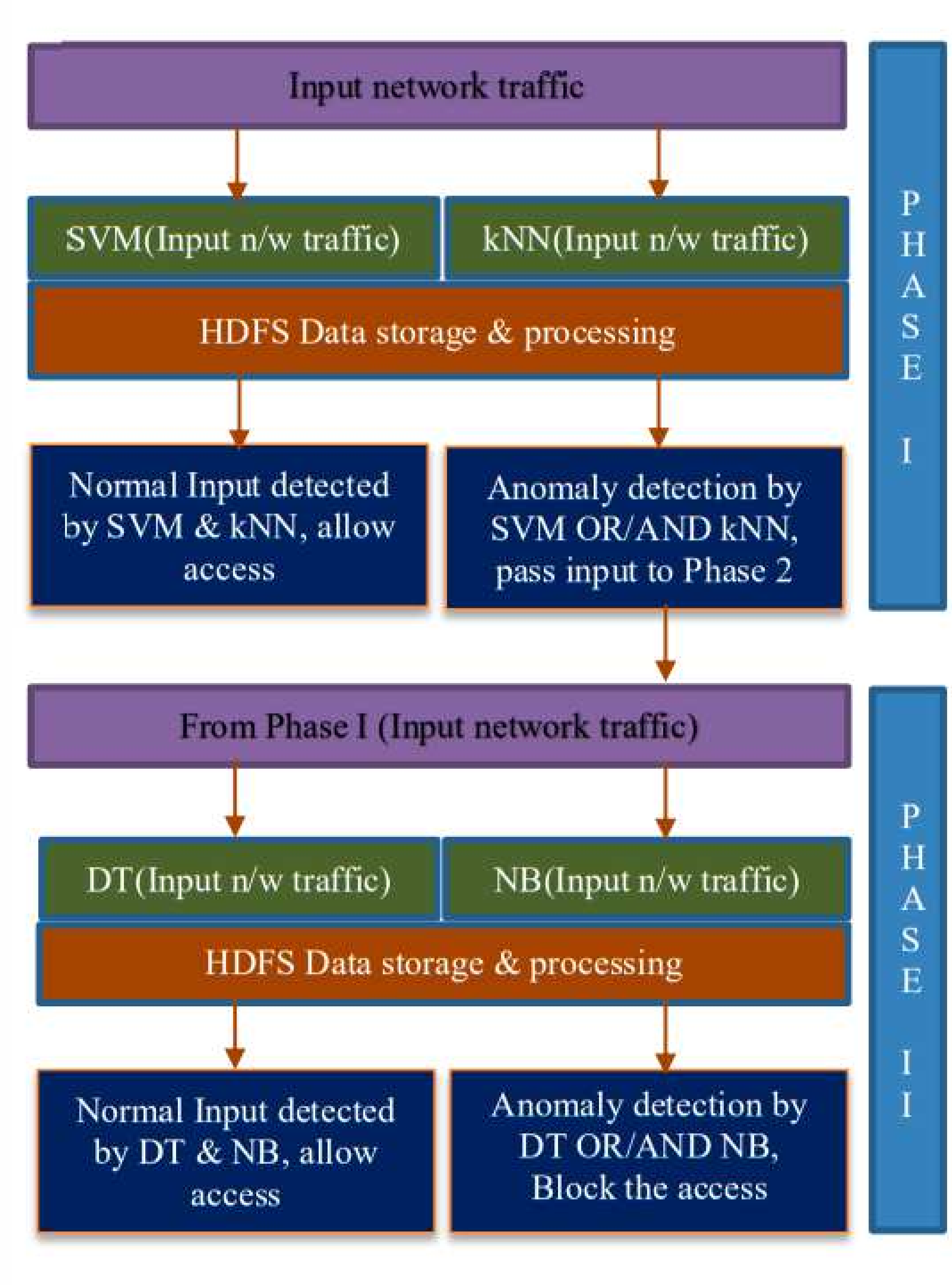

Input network traffic
SVM(Input n/w traffic)
kNN(Input n/w traffic)
HDFS Data storage & processing
Normal Input detected by SVM & kNN, allow access
Anomaly detection by SVM OR/AND kNN, pass input to Phase 2
PHASE I
From Phase I (Input network traffic)
DT(Input n/w traffic)
NB(Input n/w traffic)
HDFS Data storage & processing
Normal Input detected by DT & NB, allow access
Anomaly detection by DT OR/AND NB, Block the access
PHASE II

Figure 5

TP-IDS Architecture

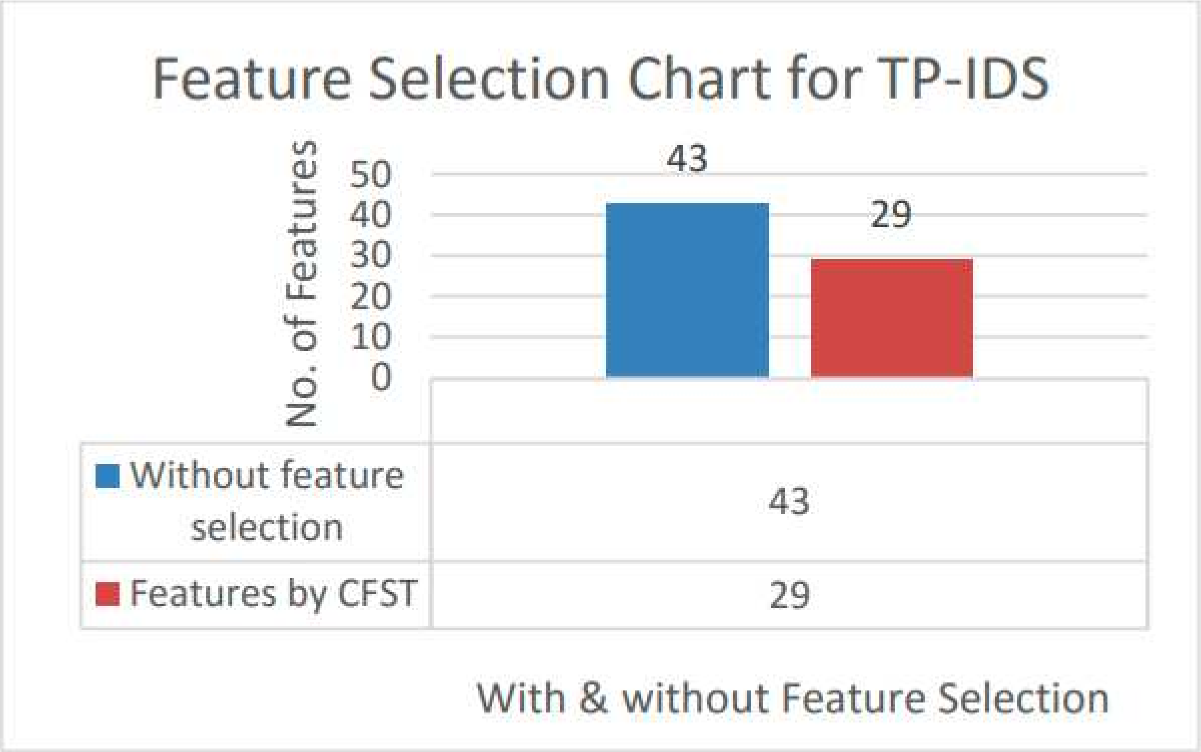


Figure 6

Feature Selection Chart

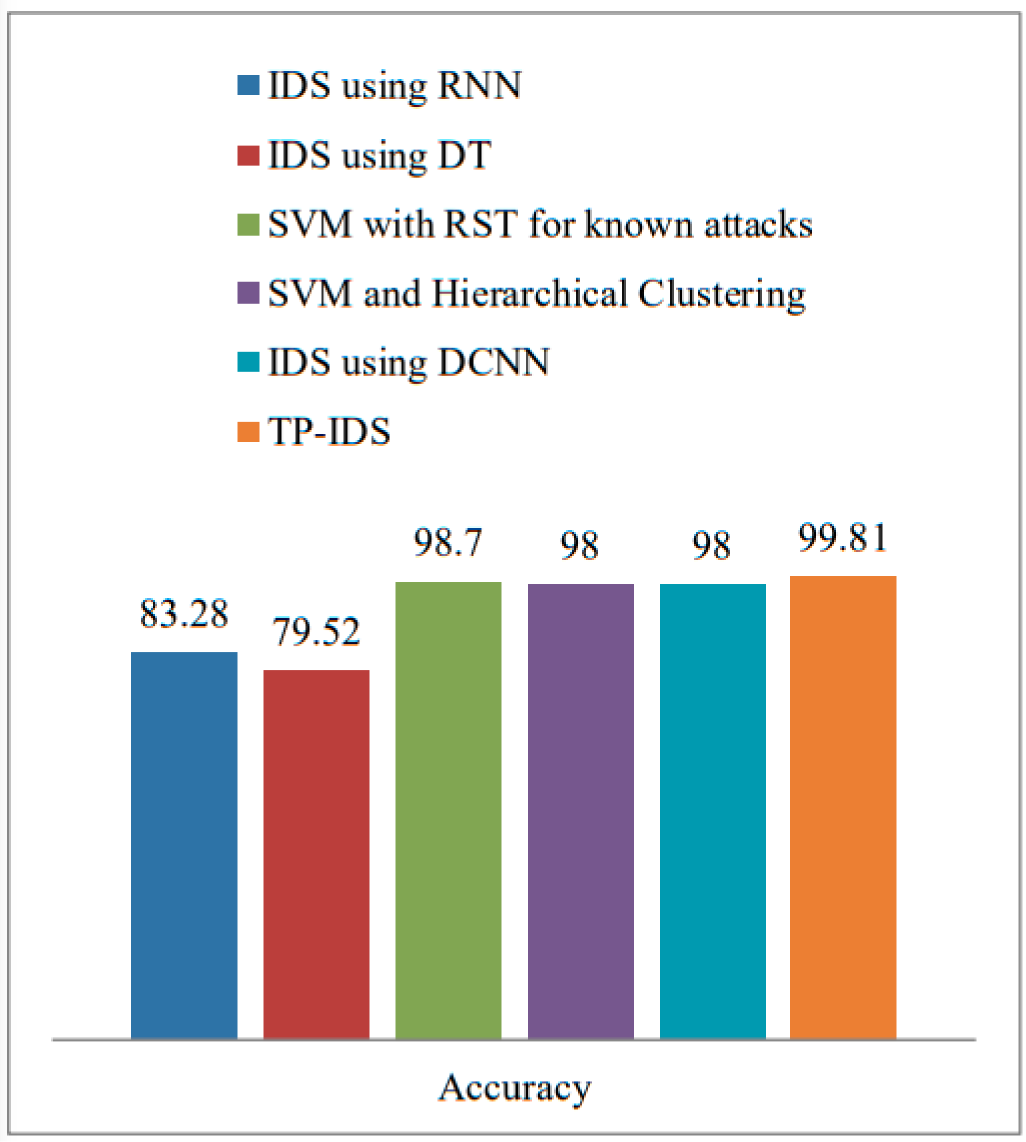


Figure 7

Accuracy results comparison