Two Phase - Intrusion Detection System (TP-IDS) model using Machine Learning Techniques

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

Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.[1]
Assistant Professor, Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune-18
abhijit.jadhav29@gmail.com
Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.
pvidyullatha@kluniversity.in[2]

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract— Network security is one of the important issues that should be handled carefully and security to our data and important information is necessary when communicating with outside world over the internet. We have so many examples of the attacks that have been happened in the past and threatened a huge loss to the world. Over the years, a lot of work is done in this area and many security systems are developed, many security algorithms are implemented which have strong security fundamentals. Also, these systems have proven strong at the time of highly insecurity incidences or attacks. But, the problem is still there in identifying the malicious node entry in the network correctly without any false positives or false negatives, also within quick time before getting any type of access to the network by such malicious nodes which are also called as intruders. We need a system which should be operating faster even in heavy incoming network traffic scenarios, and generating correct identification of the intruders as well as giving easy, faster network access to non intruders or normal connecting nodes by correctly identifying them as normal nodes. In this research work, we are carrying out the implementation of the Intrusion detection System (IDS) model which performs faster as well as classifies the communicating nodes correctly as intruder or normal user. This IDS model is being implemented in two phases, hence a Two Phase Intrusion Detection System, abbreviated and named as TP-IDS model. We are using the machine learning techniques to develop this model, which makes this model a strongly secure IDS model and helping to better identify the known as well as unkown attack. We are using first phase for the identification and second phase for the validation of the first phase results of our model, which provides high accuracy. In first phase of the IDS we are using Support Vector Machine (SVM), k Nearest Neighbor (kNN) and in second phase we are using Decision Tree (DT), Naïve Bayes (NB) for first phase validation. Also, the issue of efficiency is handled by underlying data processing infrastructure, which is HADOOP that increases efficiency or speed of the TP-IDS.

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

I. INTRODUCTION

Intrusion Detection System (IDS) is the monitoring system, which continuously monitors to the incoming network traffic to detect the unauthorized & malicious node network access, If any. The purpose of the attack can be, to attack the host system or to enter into the network as a malicious node entry and harm the network devices or important assets of the organization. Based on this, IDS can be categorized into two types as host based intrusion detection system (HIDS), network based intrusion detection system (NIDS)[1]. In host based IDS, the IDS continuously monitors the system's security for attacks and it is installed in the host itself. We do not require any special equipements or hardware for the system setup[1]. In network based IDS, the IDS monitors the network traffic for security attacks and detects the malicious requests to enter the network. In Network based IDS, IDS can act as a gateway for the network or can be setup between network and the server[1]. We herewith are focusing on the network based IDS. It is important from the security point of view, to detect the malicious nodes which are called as intruders, before they enter into the network.

The IDS are of twor types, based on detection capability. The IDS which uses the pattern matching and identifies the attacks which are previously encountered & detected by the system or which are known attacks, such IDS type is called as misuse detection systems. Misuse detection systems are also called as signature based IDS or knowledge based IDS[2]. The IDS which identifies unknown malicious behaviors which are different from the normal standard behaviors, are called as Anomaly detection systems, which is the another type of IDS[2]. Because of the nature of the Anomaly detection system to identify the unknown behaviors, it has been very popular and the topic of interest for the researchers. Anomaly detection systems have an advantage over misuse detection systems, that, Anomaly detection systems can detect unknown attacks also, where as Misuse detection systems fail to detect the unkmown attacks. Hence, we are considering the implementation fo Anonmaly detection syste. So, herewith in this research work, we are focusing on the implementation of network based anomaly detection system to detect both known and unkmown attacks. Much of the research work is done in this area, but the results are still not satisfactory, the major problems we are facing in the existing anomaly detection systems, are, false positives and false negatives, also the time required for the detection of unknown attacks is more than the ideal time, that is important, before the intruder harm the network or systems. In this work, we are working to develop a model of IDS which will overcome these issues of false negatives, false positives and the time required for detection of the intruder. False positives are the network traffic input of normal behavior, which is detected as a malicious behavior[3]. False negatives are malicious nerwork traffic input, which is detected as a normal behavior pattern[3]. In both the cases, it is important that, network access should be given to the normal user or genuine user and blocked for the malicious user or intruder. To implement this, we should develop the IDS system which can easily identify such behaviours without any false positives and without any false negatives.

Many researchers have taken efforts to develop the IDS systems with different techniques. But, the machine learning is the mechanism which can help us to achieve the IDS model with better accuracy and efficiency, also the accuracy of the model will be increased with the use in the operating environment. Machine learning with big data helps in improving the computation time and accuracy of the system[4]. Machine learning helps the system to get trained for the different patterns for categorization. In our system, the model accepts the different kind of network traffic inputs and categorizes it as an attack or normal traffic. Also, while doing this, the model learns to identify the input patterns from time to time and hence, it helps the system to tune its accuracy. Machine learning helps the system to improve accuracy automatically without writing any explcit code for it[5]. Also, it is done in less time also to maintain the timeliness of the system, if proper system structure is organized for the development and execution of the system. In IDS model, if we want to achieve the accuracy, then we can not rely on any single technique[6], a combination of techniques forming a model is necessary for getting the desired accuracy in the system. We, have so many methods available in machine learning, that can be used for the IDS model building. But, as per survey, we have identified the research gap and the well performing methods are grouped together to form a model, so that the disadvantages of one method will be compensated by the presence of another method in the model and vice versa. As per our survey, we found four different useful machine learning techniques, which can help us to improve the accuracy of the model to the highest extent. The techniques are Support Vector Machine (SVM), k Nearest Neighbor (kNN), Decision Tree (DT) and Naïve Bayes (NB) classifier. Over the time, these methods are proved to be very useful for generating better results in the models. This is for accuracy purpose only. Another question arises, when we use such four different methods as a part of the model, will it affect the performance speed of the system?, yes it will. But, to avoid this problem and in fact increase the speed of the system to supreme level, we will use the fastest underlying architecture. The name of this architecture is Hadoop Distributed File System (HDFS). HDFS is the massively parallel and distributed data processing architecture, which unbeilievably increases the system speed to fastest level. With this type of combinations , we are trying to achieve the higher accuracy with very less time execution requirements. Hence, we will implement the accurate and efficient IDS system at the end of this article.

II. LITERATURE SURVEY

We have carried out the survey to find out the existing implementations in IDS using machine learning techniques. During this study, we tried to find out, the IDS models proposed by the researchers with the use of machine learning techniques. We have noted down the assumptions in the research work carried out, the dataset used, the accuracy and performance results of the IDS systems.

In [7], Bahlali has evaluated the performance of anomaly-based NIDS by using multiple machine learning algorithms like linear regression (LR), decision tree (DT), Random Forest (RF) and Artificial Neural Network on the UNSW-NB15 dataset. It is observed in the results shows that ANN gives better overall performance with correct classification and less false negatives. The major problem here, is that, USNW-NB15 dataset is used which has several issues such as imbalanced classes and the recorded malicious traffic. From this, it can be concluded that, only the seen attacks will be identified and system fails to identify the unseen aatacks. Also, the accuracy in classifying the network traffic is less which is near to 77, this is for multi classification. Hence, the methods used with the dataset USNW-NB15, does not give the desired accuracy and efficiency results.

In [8], F. Yihunie et. al have analysed the performance of five different machine learning techniques in intrusion detection system. The techniques are Random Forest (RF), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Sequential Model and Support Vector Machine (SVM). The team have used NSL KDD dataset for the training and testing of these techniques. It is concluded that, the Random forest technique performs better as compared to other four techniques. The important part in the research work carried out by the authors is the identification of known attacks only. All these methods are used in comparative manner and hence, the accuracy is not validated. Also, no comment is made on the tperformance of the system for unknown attack identification. As we have studied, we can not have IDS model with either of the machine learning technique, as neither of the techniques gives 100% accuracy in intrusion detection. So, authors have concluded that, combination of machine learning techniques is required for the accurate IDS model.

In [9], Zamani has studied different machine learning techniques for the application in intrusion detection system. The author has considered that, the machine learning techniques can be classified in two categories like Artificial Intelligence and Computational Intelligenece. Although they have similarities, but also have different unique properties are offered by these techniques which gives advantage to deveop the fault tolerant, adaptive and high computational speed IDS system. Author has concluded that, combination of different such techniques gives the better results in intrusion detection systems. The machine learning techniques helps us to achieve the desired accuracy and performance in the intrusion detection system.

In [10], Nguyen Thanh Van et. al have stated the research work carried out in IDS development using deep learning techniques. Authors have evaluated the performance of two deep learning techniques such as stacked Autoencoder (SAE), Stacked Restricted Boltzmann Machines (RBM). It is concluded that the Stacked Autoencoder gives good performance as compared to stacked RBM, as RBM requires more time for computations. The computations in RBM are complex, hence it is slower as compared to the SAE. Slow computation rate or more time requirement for training of the IDS using deep learning is the inherited disadvantage from deep learning techniques. Hence, though deep learning techniques helps us to identify the unknown attacks up to some extent, the inability of slow computations or poor performance speed discourages the use of deep learning in intrusion detection systems, as timeliness is the important requirement in security issues of the systems.

In [11], Mehrnaz Mazini et. al have used two important techniques namely artificial bee colony (ABC) and AdaBoost algorithms. The ABC algorithm is used for the feature selection to select the subset of the unbalanced class of features. The unbalanced class of features and records unnecessarily slow down the working of the data mining techniques and algorithms, hence ABC is used in the research work implementation of IDS model, which helps to improve the speed as well as accuracy of the IDS model. The authors have concluded that, the accuracy of the IDS model using ABC and AdaBoost algorithms is improved as compared to legendary techniques, but the work is carried out with the known attack samples and hence can not be considered as the ideal method for the IDS model. Another disadvantage of the system is the performance speed of the system is not as per the ideal timing requirements, hence the improvements in the model are necessary.
In [12], Erxue Min et. al have proposed the IDS model using the word embedding and text-CNN methods. Also for final classification of the input network data, random forest is used. The word embedding technique is used to fetch the semantic relations from the payloads, which reduces the feature dimensions. Also, text-CNN is used to extract the features from the payloads and finally random forest classifies the input. In the work that is carried out by the authors, the payloads used are manual and hence, it is difficult to get the the accurate results in the actual working environment. The accuracy is the only part of concern in the work and that too not stated with the convincing results. Hence, it can be concluded that, the techniques are not very helpful to be used together to form the IDS model as accuracy and performance speed issues are not very well addressed by this research work.
In [13], Nutan Farah Haq et. al hav carried out the survey of machine learning techniques for intrusion detection systems. Authors have carried out the study of 49 different papers and have studied the performance issues and key factors in the area of IDS using machine learning techniques. One of the important conclusion of the authors is that, the feature selection is the key factor for improving the performance of the system during training. If, the better feature selection algorithm is used, then the performance of the IDS system can be enhanced easily, is proved through the survey work. Also, one of the important conclusion of the work is that, the group of machine learning techniques should be used to form a model, instead of a single machine learning method in IDS model building.
In [14], Amouri A et. al have implemented the intrusion detection system using the two layer model of machine learning. The results are showing the accuracy is not as per desired requirements of the security domain. The false positive rate (FPR) is appeared to be increasing the existing used IDS model. The system is basically given for the Mobile AdHoc Networks (MANETs) and Wireless Sensor Networks (WSN). The IDS model is based on heuristic based approach in the first stage identification and linear regression in the second stage identification of the network traffic. The important issues of IDS which accuracy and efficienacy are remained unaddressed in the work.
In [15], Dr.S.Malliga et. al have used the Deep learning approach and techniques for the IDS model implementation. The techniques used are Artificial Neural Network (ANN) and Recurrent Neural Network (RNN). The authors have stated about the accuracy of the model with these methods in comparison with machine learning techniques. The results are not shown with the training and testing parameter values, samples used, features selected etc. Only accuracy values are shown which are not convincing statistical values without any support information. Also, deep learning techniques take more time for the training and hence, very compromising to use in the timeliness system like intrusion detection. Also, authors have not stated the comparison with the actual machine learning techniques like SVM, Decision tree etc. The statistics given is in very much generalized form, which is not givng clear idea about comparative value information.
In [16], JABEZ J et. al have proposed the IDS model structure by using the machine learning technique based on outlier detection mechanism named as Neighborhood Outlier Factor (NOF). The authors have used the old KDD dataset which has unbalanced structure and hence the results obtained in the research work, can not be assumed to be true in the actual operating environment of the IDS model. The approach helps to identify the attacks, but the results should be verified with some other mechanism, that is not included in this work. We can not trust the results without verification phase. IDS with only single machine learning technique, never gives accurate results, it might neglect the high impact of false positives (FPR) and false negatives (FNR). No statistics is given by the authors about FPR and FNR, which is definitely the major part of concern in IDS results accuracy.
In [17], Ambreen Sabha have performed the survey of intrusion detection system using machine learning techniques. Authors have drawn few important conclusions from the survey. The first conclusion is that, every technique in machine learning has merits and demerits which gets inherited in the IDS and affects its performance, hence a proper selection of set of machine learning techniques to form a IDS model is required. Secondly, author has commented on the quality of the datasets used to validate the work, KDD CUP99 dataset has many issues and hence NSL KDD is the better option identified by the authors, as it is choiced by many researchers also. The dataset choice and feature selection also plays very important role in achieving the accuracy and efficiency of the IDS model.
In [18], Urooj Aslam et. al have proposed the intrusion detection model using rule based learning and machine learning classification technique. The authors have used SNORT for rule based learning, which helps IDS to identify the network traffic input as attack or normal input and simple logistic, J48 and Sequential Minimal Optimization (SMO) for machine learning classification, which finally classifies the input. The approach used is not the validation based and either of the result, if positive then will be treated as final result as attack and access will be blocked. With this approach, though author is stating about decreasing the false positive rate, but it is clear that, with OR of these techniques, false positives can not be reduced at all. So, model fails in giving the validated accuracy, also, with so many algorithms in sequential execution is a time consuming execution of the system and hence, slows down the IDS model.

In [19], Abien Fred M. Agarap has used the Gated Recurrent Unit (GRU) & Support vector machine (SVM) as a part of intrusion detection system model. The research work is carried out in comparison with GRU Softmax IDS implementation, where the softmax is replaced with SVM for speed improvements. But, still the SVM used is a linear SVM and used for binary classification of the input. The GRU is a technique of recurrent neural nework, which takes time to train the model. This model is not convincing for its use in real time scenario, as it is developed in comparison with only GRU softmax and not considered for performance improvements with other existing IDS models or solutions. Hence, the results stated in the article are only for comparison with GRU softmax and does not show the usability of the work in the real environment or scenario. Also, the dataset used is the private dataset of the Kyoto University, which again the area of concern, whether the features and records used in the dataset are balanced or not. No, feature selection technique is proposed for improving the accuracy of the system.
In [20], Mohammad Kazim Hooshmand et. al have proposed the two phase intrusion detection system model. In first phase authors have used random forest for identification of the input as an attack or normal. In second phase authors are using the neural network to categorize the attack traffic into specific category. Authors have concuded that accuracy is achieved through two phase approach, but still they faced challenge in finding out the appropriate attack category and improvements are required for the same.
From, all that survey we have carried out, concludes that, no single machine learning technique is useful to develop the intrusion detection systems. Hence, a group machine learning techniques is important for developing the IDS model, which will be giving accurate results by reducing the false positive rate and false negative rate values. So, appropriate combination of machine learning techniques is required to detect the known as well as unkown attacks correctly. Based on the survey that we have carried out, we identified the research gap and this new ideal intrusion detection system model is proposed here.

III. DEFINITIONS AND TECHNIQUES

In this research work, we are using different machine learning techniques in two phase architecture implementation of the intrusion detection system. Also, we are using HDFS as the underlying infrastructure of the system. The techniques we are using are:

1. Support Vector Machine (SVM)
2. K nearest neighbor (kNN)
3. Decision Tree (DT)
4. Naïve Bayes
5. Hadoop Distributed File System (HDFS).

Here, we will try to highlight few important strengths of these techniques.

1. **Support Vector Machine (SVM)*:*** Support Vector Machine can be abbreviated as SVM, is the classification technique in machine learning. SVM classifies the data points by hyperplane between the separated classes. The support vectors are the data points at the boundry of the classes and hyperplane maximizes the distance between these support vectors to better separate the data points of the distinct classes. The advantage of the support vector machine is that, it is faster regardless of high dimensionality of the available data[21]. It is highly accurate machine learning technique with the requirement of pre existing knowledge of the events. In case of unknown events, SVM does not perform well and this is the limitation of the SVM[21]. SVMcan be used for binary as well as multiclassification. But, the accuracy of the SVM binary classification is very high and acceptable in the security systems.

2. **k Nearest Neighbor clustering(kNN):** k nearest neighbor is a supervised learning technique, that classifies the data points based on minimum distance from k data points. In a technique, we calculate the distance of the data point to be classified from all data points. Out of these, first closest k data points are selected, and the data point will be classified to a class, where majority of the data points of the class are in selected closest k data points[22]. kNN, though a supervised technique, we will be using it as a clustering technique. The important advantage of the kNN is, it does not need any training and one of the faster technique in machine learning for classification of unkown events. The limitation of the kNN is, its accuracy should be validated by validation mechanism.

3. **Decision Tree (DT):** Decision Tree is a classification machine learning technique. In decision tree, the leaf nodes are the classes to which the data points are classified. The internal nodes are the conditions, which are the intermediate nodes along the path from root element to the final class of the data points. Th e technique is very popular and provides useful way to deal with the adverse security identifications[23]. The decision tree has important advantages like, the classification of the data points to their specific class, identifying unkown

events, performance speed and proven very useful in intrusion detection systems[24]. The disadvantage of the decision tree is, even in case of small change in tree structure, the complexity becomes more for classification and it may affect the execution time of an algorithm.

4. **Naïve Bayes classifier:** Naïve Bayes Classifier is the probability classification technique. It works on probability independence theory. To calculate the probability of the occurred end result, it considers the probability of occurring evidence variables independently to occur the end result[25]. It makes strong attribute independence assumption, means the probability of one attribute is independent of other attributes in the group[26]. Naïve Bayes gives good accuracy for the classification of events with general attribute informations. It is a technique which calculates prior and posterior probability of the event happenings.

5. **Hadoop Distributed File System (HDFS):** Hadoop Distributed File System is a massively parallel data storage and processing infrastructure. It is highly fault tolerant and it can be installed in a commodity hardware, a low cost systems also[27]. Hadoop has namenodes and datanodes for parallel and distributed processing. Name node is a node which is a master node and manages data nodes. It distributes data among data nodes, which are the actual data storage and processing nodes. Name node also stores the meta data, of the data which is distributed among data nodes. It uses block structure for dividing and storing data in data nodes. Because of the massive parallel processing, it is the most efficient processing infrastructure.

IV. OBJECTIVES

The important objectives of the research work proposed are as follows:

1. To develop the Two Phase IDS (TP-IDS) which is anomaly based network intrusion detection system with timeliness in detection.
2. To generate accurate intrusion detections for all types of network attacks including unknown input events.
3. To dvelop faster intrusion detection system using machine learning techniques and HDFS storage and processing infrastructure.

We have set these objectives, by studying the literature and finding the research gap that strongly states about accuracy and efficiency requirements in the proposed IDS model development.

V. ROPOSED METHODOLOGY

The Fig. 1 shows the proposed architecture of the Two Phase Intrusion Detection System named as TP-IDS. It is basically targeting the anomaly based network intrusion detection. As shown in fig. 1, we are dividing our work in two phases. First phase consists of Suport Vector Machine and k nearest neighbor techniques which will be independently executed using HDFS infrastructure to achieve performance speed. If, both or one of these techniques recommends the input network traffic as an attack, then the input network traffic will be passed to phase two for validation of the first phase result. If, neither of the first phase technique recommends the input network traffic as an attack, then the input network traffic will be considered as a normal user connection and access to network will be allowed. In phase II, the decision tree and naïve bayes algorithms will be executed in parallel and independently using HDFS infrastructure. If both or either of the technique, recommends input traffic as an attack, then validation of the Phase I result will be successful and the connection request will be interpreted as an attack and access to the network will be blocked. If both of the techniques recommends the input traffic as a normal traffic then, validation result of the first phase result will be false and input connection request will be considered as normal and access to network is allowed. The algorithm is as follows:

TP-IDS Algorithm (Input network traffic)

a. Call to **Phase I():**
b. Call to **ParallelSVM()** and **ParallelkNN()**
c. **If (ParallelSVM() recommends input network traffic as an attack OR ParallelkNN() recommends input network traffic as an attack)** then
   Call to Phase II()
   Else
   Input network traffic is normal user connection and access to network allowed.
d. **Phase II():**
e. **If (ParallelDT() recommends input network traffic as an attack OR ParallelNB() recommends input network traffic as an attack)** then
   Block network access to connection request.
   Else

Input network traffic is normal user connection and access to network allowed.

f. End.

The TP-IDS algorithm is the proposed algorithm of the research work carried out. Phase II() is called only when ParallelSVM() and/or ParallelkNN() recommends the input network traffic as an attack, as Phase II() is the Phase I() result validation phase, else Phase II() will not be called. Many existing systems failed because of the false positives, as there only methods in the architecture recommends normal connection as an attack and this is only because, this result is not validated after first identification. In our work, though Phase I () methods recommends the input as an attack, still it will be considered as an attack only when Phase II() methods will also recommend it, an attack, else, it will be considered as a normal connection. This is how the first attack detection is validated with two more methods and that makes this architecture more accurate and effective than existing architectures

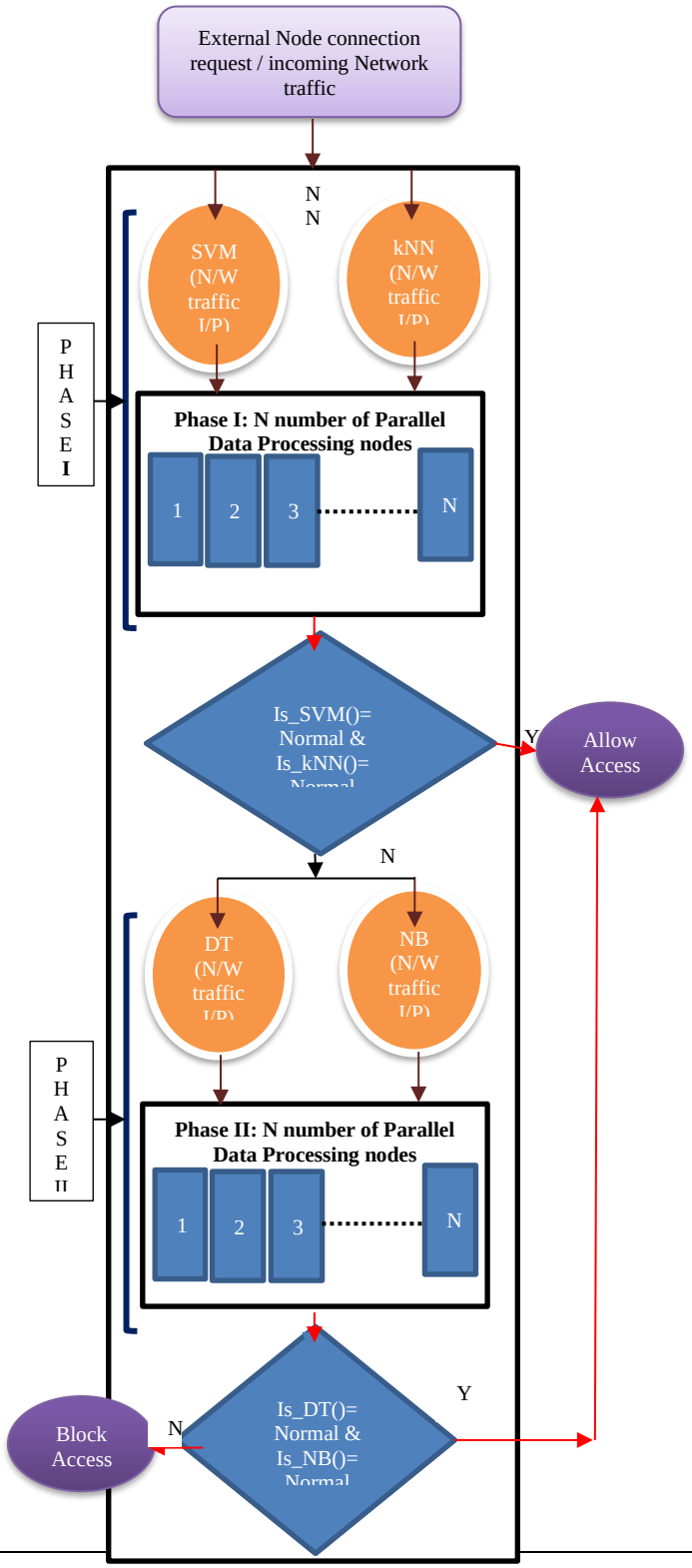

Fig. 1: TP-IDS Architecture

VI. DATASET AND DATA PRE

a. NSL KDD Dataset:

We are using the NSL KDD data set. It has 43 attributes, 1,25,973 training data records and 22,544 testing data records of different types of network attacks. The attack wise records are as shown in fig. 2 graph.

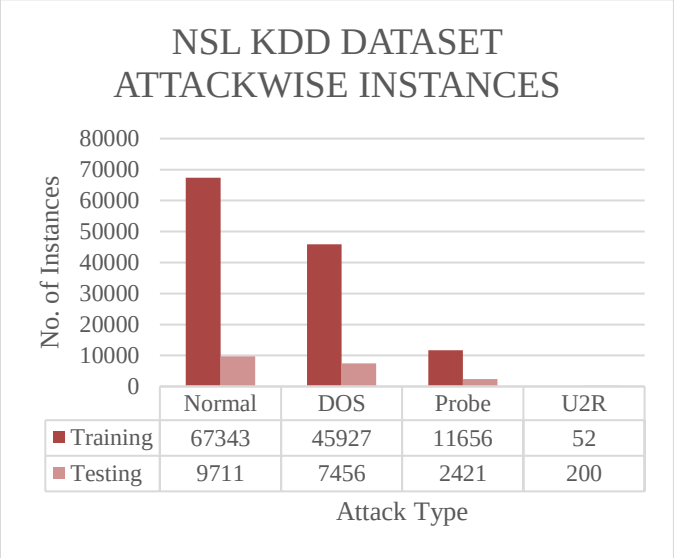

Fig. 2: Attack types and number of instances in NSL KDD training-testing datasets

NSL KDD dataset has few important characteristics, because of which it is mostly used by many researchers. It does not have any redundant records of the data, which definitely avoids any possibility of biased results. It is one of the better datasets to develop the effective security system.

b. *Data Pre processing:*

In data pre processing work, we are focussing more on feature selection process. Feature selection is very important to improve the accuracy of the model. It is important to identify and use only those features which have any relation with the output variable. We have many algorithms and techniques of feature selection like correlation based feature selection (CFS), Information Gain, Gain ratio. From these methods, we are using CFS i.e. correlation based feature selection technique. With this technique, we have selected 29 attributes from the available 43 attributes of the NSL KDD dataset. Also, few attributes values are converted to normalize form to increase the accuracy and generate the valid results for kNN algorithm.

VII. CONCLUSION AND FUTURE WORK

Here, we have completed the effective intrusion detection model TP-IDS implementation, which can provide us high accuracy than any of the existing IDS models and also the faster performance speed that can be achieved through underlying distributed as well as parallel data storage and data processing infrastructure HADOOP. The model is developed in two phases to take care of accuracy validation and reduce false negative & false positives in the results. In future work, we will be focusing on the structure of the machine learning techniques used, whether we can change these methods to get more accurate and faster IDS model to reach the best IDS solution ever.

ACKNOWLEDGEMENT

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

REFERENCES

[1] 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

[2] Ansam Khraisat* , Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman, “Survey of intrusion detection systems: techniques, datasets and challenges”, Cybersecur 2, 20 (2019). https://doi.org/10.1186/s42400-019-0038-7

[3] P. Garcı´a-Teodoro, J. Dı´az-Verdejo, G. Macia´-Ferna´ndez, E. Va´zquez, “Anomaly-based network intrusion detection: Techniques, systems and challenges”, computers & security 28 (2009) 18–28, doi:10.1016/j.cose.2008.08.003

[4] Suad Mohammed Othman, Fadl Mutaher Ba-Alwi , Nabeel T. Alsohybe and Amal Y. Al-Hashida, “Intrusion detection model using machine learning algorithm on Big Data environment”, J Big Data 5, 34 (2018). https://doi.org/10.1186/s40537-018-0145-4

[5] 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”, Third International Conference on Computing and Network Communications (CoCoNet’19), Procedia Computer Science 171 (2020) 1251–1260

[6] Pascal Maniriho , Tohari 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, 978-1-5386-5813-0/18/$31.00 ©2018 IEEE

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

[8] 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.

[9] Zamani, Mahdi. (2013). “Machine Learning Techniques for Intrusion Detection”.

[10] Nguyen Thanh Van, Tran Ngoc Thinh and Le Thanh Sach, "An anomaly-based network intrusion detection system using Deep learning," 2017 International Conference on System Science and Engineering (ICSSE), Ho Chi Minh City, 2017, pp. 210-214, doi: 10.1109/ICSSE.2017.8030867.

[11] Mehrnaz Mazini, Babak Shirazi, Iraj Mahdavi, “Anomaly network-based intrusion detection system using a reliable hybrid ar tificial bee colony and AdaBoost algorithms”, Journal of King Saud University - Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2018.03.011

[12] Erxue Min , Jun Long , Qiang Liu , Jianjing Cui and Wei Chen, “TR-IDS: Anomaly-Based Intrusion Detection through Text-Convolutional Neural Network and Random Forest”, Hindawi, Security and Communication Networks, Volume 2018, Article ID 4943509, 9 pages, https://doi.org/10.1155/2018/4943509

[13] Nutan Farah Haq, Musharrat Rafni, Abdur Rahman Onik, Faisal Muhammad Shah, Md. Avishek Khan Hridoy, Dewan Md. Farid, “Application of Machine Learning Approaches in Intrusion Detection System: A Survey”, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.3, 2015

[14] 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

[15] 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

[16] JABEZ J , Dr.B.MUTHUKUMAR, "Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach", International Conference on Intelligent Computing, Communication & Convergence, Procedia Computer Science 48 (2015) 338 – 346

[17] Ambreen Sabha, "Anomaly-based Intrusion Detection using Machine Learning Algorithms - A Review Paper", International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 08 | Aug 2020

[18] Urooj Aslam, Ezzat Batool, S. Nadeem Ahsan and Abdullah Sultan, "Hybrid Network Intrusion Detection System Using Machine Learning Classification and Rule Based Learning System", International Journal of Grid and Distributed Computing, Vol. 10, No. 2 (2017), pp.51-62, http://dx.doi.org/10.14257/ijgdc.2017.10.2.05

[19] Abien Fred M. Agarap, "A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data", ICMLC 2018, February 26–28, 2018, Macau, China

[20] Mohammad Kazim Hooshmand, Doreswamy, "Machine Learning Based Network Anomaly Detection", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-4, November 2019

[21] Jayshree Jha and Leena Ragha. Article: Intrusion Detection System using Support Vector Machine. IJAIS Proceedings on International Conference and workshop on Advanced Computing 2013 ICWAC(3):25-30, June 2013

[22] 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.

[23] Simon, G. J., Xiong, H., Eilertson, E., & Kumar, V. (2006). Scan detection-a data mining approach. SIAM International Conference on Data Mining, 6, 118-129.

[24] 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, 2012, pp. 629-634, doi: 10.1109/ICCT.2012.6511281.

[25] Panda, Mrutyunjaya & Patra, Manas. (2007). Network intrusion detection using naive bayes. IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007.

[26] Saurabh Mukherjee, Neelam Sharma,Intrusion Detection using Naive Bayes Classifier with Feature Reduction,Procedia Technology,Volume 4, 2012, Pages 119-128, ISSN 2212-0173, http//doi.org/10.1016/j.protcy. 2012.05.017.

[27] Miss Gurpreet Kaur Jangla , Mrs Deepa.A.Amne, Development of an Intrusion Detection System based on Big Data for Detecting Unknown Attacks, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 12, December 2015, ISSN (Online) 2278-1021, ISSN (Print) 2319 5940.