IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 5, October, 2021, pp. 359-366
Journal homepage: http://iieta.org/journals/ria

Accuracy Based Fault Tolerant Two Phase - Intrusion Detection System (TP-IDS) Using Machine Learning and HDFS

Abhijit Dnyaneshwar Jadhav[1,2*], Vidyullatha Pellakuri[1]

[1] Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India
[2] Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune 411018, India

Corresponding Author Email: abhijit.jadhav29@gmail.com

https://doi.org/10.18280/ria.350501

Received: 10 August 2021
Accepted: 20 September 2021

Keywords:
TP-IDS, HDFS, machine learning, accuracy, timeliness, fault tolerance, innovation

ABSTRACT

Security of the organizational assets like data, from external intruders is an important issue now a days. Already, a lot of research has been done and many researchers have come up with the useful solutions also. This research is about developing the model for monitoring network systems to identify the entry requests of malicious nodes and such models are called as Intrusion Detection Systems (IDS). But, many of these IDS have kept the scope open for improvements, in terms of reducing false positives, reducing false negatives, increasing the speed of intrusion detections, fault tolerance systems. To achieve these enhancements in IDS, different solutions are proposed and implemented. When the attempts are made for increasing accuracy, timeliness of the system is not achieved, also with timeliness, fault tolerance and accuracy is compromised. In this research, we have implemented the time effective and accuracy based system with one important feature of fault tolerance, by using the distributed file processing architecture and machine learning techniques. The system is designed and implemented in two phases, and hence named as Two Phase- Intrusion Detection System (TP-IDS). Each phase of the system proposed consists of machine learning techniques which are executed in the underlying Hadoop Distributed File System i.e. HDFS data storage and processing architecture. HDFS is distributed and parallel architecture, which helps to achieve the timeliness and also fault tolerance due to distributed nature of HDFS. Machine learning techniques are helping us to learn from the experience and increase the accuracy of the model. The results of the research have shown significant improvements in all aspects like accuracy, timeliness and importantly fault tolerance.

1. INTRODUCTION

1.1 Machine learning: Let's classify accurately

Machine learning is the branch of artificial intelligence which helps the community to develop solutions which are based on training and testing of the models. The important thing is the amount of data required for training of the model, which affects the accuracy of the model. Higher is the amount of data, more is the accuracy rate and lower is the amount of data, less the chances of producing highly accurate model. Hence, data plays very important role in machine learning models. Now, a days, the use of the internet world is increased unbelievably to a huge amount and hence the huge amount of data is generated every day. This data can be, if used effectively, then with the help of machine learning techniques, the solutions in the business environments for decision making and automation of the business process can be achieved to make it more sophisticated, and manpower independent. Similar kind of approach of machine learning can be used for targeting the solution model for intrusion detection systems. Machine learning is very effective approach for developing solution models in the network security field and problems. Many researchers have already provided with proven solutions about use of machine learning approach for security field and intrusion detection system as well.

1.2 Intrusion detection system problem and enhancement

Intrusion Detection System (IDS) is one of the application areas where improvements can be achieved in the dimensions of accuracy, timeliness and fault tolerance using machine learning techniques along with parallel processing Hadoop Distributed File System (HDFS). The challenges in the existing IDS models are as shown in Figure 1. The intrusion is the entity tries to enter the organization network with unauthorized access as a malicious node. Intrusion Detection System is the monitoring system of network traffic to identify the intrusions and alert on intrusion detection, if any Intrusion Detections [1] can be employed at host systems as host based intrusion detection systems or at networks connecting computers as a network based intrusion detection systems. There are two types of intrusion detection systems Misuse detection and Anomaly detection system. Misuse detection system uses the pattern matching, identifies the known patterns for intrusion detection [2]. Hence, it gives better accuracy in case of known attacks and fails for unknown attack identification [3]. Anomaly detection system defines the baseline for the normal behavior and anything outside the baseline is detected as an attack. Anomaly detection system is useful for both networks based and host based intrusion identifications and hence should be improved with the architectures to achieve better results in accuracy [4].

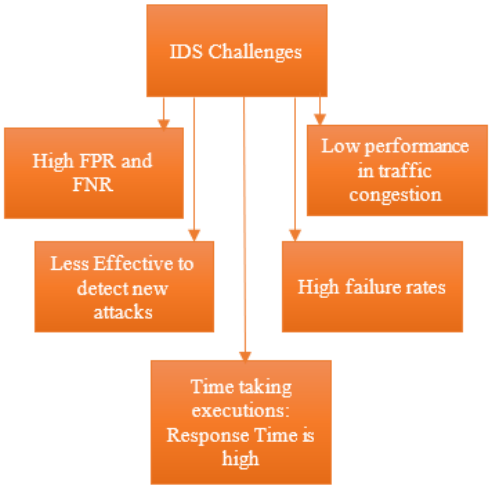

Figure 1. Challenges in IDS: Problem Motivation

Machine learning techniques can be effectively applied to design the model of anomaly based intrusion detection system to achieve the accuracy goals [5]. There are different machine learning categories like classification, clustering, regression. Classification is the machine learning technique used for classification of the input data into labelled or known classes. Clustering is the machine learning technique used for grouping of data items based on similarity or dissimilarity measure into different classes. Regression is the machine learning technique used for prediction of the continuous and discrete value data variables. It can be easily understood that, we can use classification and clustering in combination for developing the intrusion detection model to achieve the highly accurate results. To improve the timeliness and fault tolerance is the important part in achieving the improvements in the existing intrusion detection system models. It is important to identify the attack or intrusion before it damages the network and organizational assets like data. This feature of the intrusion detection system is called as timeliness. Also, though any of the hardware nodes of the system fails, still system should work with other backup or available hardware nodes and should be up for 24X7, this feature is called as fault tolerance.

1.3 TP-IDS with HDFS to achieve timeliness and fault tolerance

To achieve the timeliness and fault tolerance, the distributed data processing system is very useful. Here, we are using the Hadoop distributed file processing system for processing the data by using multiple data nodes connected to name nodes to achieve the parallelism in data processing. By using, HDFS we process the input data by using different machine learning techniques to achieve the parallelism for increasing the speed of the system which helps to achieve the timeliness. The structure of the HDFS is distributed with thousands of name nodes and each name node is connected to thousands of data storage and processing nodes, hence system structure is encouraging the decentralized approach making the system fault tolerant.

1.4 TP-IDS (Two Phase - Intrusion Detection System): A complete solution using Machine Learning Techniques

The important thing for any intrusion detection system is the accuracy with which it detects every intrusion entering attempt and protects the system from unwanted access. With this intent, and from the previous attempts of designing intrusion detection system by using machine learning, it is realized that, to achieve the accuracy in intrusion detection, use of any of the standalone machine learning techniques is not sufficient, the combination of the machine learning techniques is important. From the comparison with respect to parameters like speed of execution, accuracy etc. it is observed that, four machine learning algorithms can provide better results in intrusion detection system. The algorithms are Support Vector Machine (SVM), k nearest neighbor (kNN), Decision Tree (DT), Naïve Bayes (NB). These four algorithms are used in two phases to verify the accuracy of the results of detection in Phase I by using Phase II, hence the system model designed is named as Two Phase – Intrusion Detection System (TP-IDS). The system is designed to detect both known and unknown attacks in the network. The results obtained in the accuracy and timeliness are found significant and proving the importance of the implemented model.

In this research article, the researchers are explaining the innovation with related work of existing systems and their drawbacks, objectives of the research, definitions and concepts used in the model designing, solution methodology, data set and data preprocessing, post implementation results and discussion and finally concluding the research conclusion.

In the results and discussion part, the detailed statement about solution accuracy is explained. The significant accuracy is obtained by the researchers, which is 99.93% for multi attack type identification. The time required for execution of the algorithm is also very less, as because of HDFS the speed of the system increases to huge amount.

2. RELATED WORK

The useful literature survey is carried out with regards to background work implemented in the field of intrusion detection system and architectures. We have studied the different existing designs of the intrusion detection systems and their performance issues associated with accuracy, timeliness and fault tolerance. The detailed study is presented here as follows.

Han et al. [6] have presented the work of intrusion detection in cyber physical systems. The work presented is a survey study carried out by authors, concluding that, cyber physical systems have the characteristics like timeliness, fault tolerance and effectively can be very useful for intrusion detection. The work is related to assumption and does not have any practical implementation base, hence cannot be considered as valid arguments as conclusive part. This study provides the importance of timeliness and fault tolerance in Security systems.

Otoum et al. [7]. have stated the intrusion detection system using deep learning-based approach. The Clustered Restricted Boltzmann Machine-Intrusion Detection System (RBC-IDS) is the name given by the authors to their implemented system. The results are compared with adaptive machine learning based intrusion detection system. The system uses 3 hidden layers and resultant accuracy is 99.12% approximately as stated in the result section. At the same time, it is also found that, time taken by RBC-IDS is double that of the required by adaptive machine learning based intrusion detection. So, the system does not satisfy the timeliness characteristics and hence cannot be implemented in information sensitive

networks.

Pu et al. [8] have presented an unsupervised anomaly detection algorithm using subspace clustering and one class support vector machine to detect intrusions. The system is named as SSC-OCSVM. The dataset used is NSL-KDD dataset and all the attributes are used for obtaining results. The important thing that is noted here, all the attributes are not associated with attack type attribute and hence, results obtained cannot be considered as accurate results as irrelevant attributes are also a part of input data. The results are compared with other clustering approaches and found better by authors. But these results are not verified, as it is important to verify the results obtained by the clustering techniques. The sequential approach is used for algorithm execution and hence, timeliness along with fault tolerance is not considered here in this research work.

Wang et al. [9] have presented the work of explainable model of intrusion detection system using machine learning techniques. The existing systems that we use now a days, are not explainable systems, as the reason of the specific decision taken by the system is not visible and it is a blind output for the users. In this work, authors have proposed theoretical foundation for this system using SHAP which helps in every possible output with explanation. The significance of the decisions taken with explanation is to improve the accuracy of the system and the same is stated by the authors through this work. It also concludes that, how we can achieve better results by using machine learning techniques with such type of approach.

Othman et al. [10] have presented the intrusion detection system model using machine learning and big data underlying infrastructure. The authors have proposed idea of using Spark Apache big data environment for the execution of Support Vector Machine (SVM). It is concluded that, because of the use of the spark apache the speed of execution is increased significantly. Also, they have used feature selection for appropriate use of dependent and independent variables to increase the speed and accuracy. But they have not explicitly stated any conclusion about effective accuracy rate, as used only one standalone machine learning technique for classification. The input is classified to either normal or attack type i.e. binary classification is executed. The data used is NSL KDD dataset for this work. This study is helpful to understand the importance of distributed file processing architecture.

Bankapalli et al. [11] have presented the architecture of intrusion detection system using deep learning techniques. The authors have used the NSL KDD dataset for the execution and model building. The deep neural network with auto encoder hidden layers are the model techniques which are compared with the regular deep neural networks and classical machine learning techniques. The accuracy observed and claimed by the authors is better than the accuracy rates of the deep neural network and classical machine learning techniques. But no comment is stated in the conclusion and the experimental work about the timeliness and fault tolerance of the deep neural network takes more training time and execution time, which is not accepted in the time sensitive and information sensitive applications.

Taher et al [12] have carried out the research work for designing of the intrusion detection system using classification techniques. Authors have implemented IDS with Decision tree classifier. They have used the NSL KDD data set and used preprocessing of the data with feature selection. The decision tree algorithm C4.5 is used for the decision tree construction and the results of the same are compared with the naïve bayes classifier. The accuracy claimed is more for C4.5 which is approximately 99% and more than the naïve bayes classifier. The drawback of the system is, the given results are only for detecting the known attack patterns and classification technique like decision tree does not provide better results for the unknown attack patterns. Also, no statements are concluded regarding timeliness and fault tolerance of the intrusion detection system.

Jabez and Muthukumar [13] have presented the intrusion detection system using supervised machine learning techniques. The authors have implemented different models with different machine learning techniques like SVM, ANN etc. Among all, the accuracy for ANN is observed as 94.02% and better as compared to all other techniques. The accuracy observed can be improved by better factors, the work is just the comparative analysis by using either of the machine learning techniques. Also, the work is for detecting known attacks only and no conclusive part is given about unknown attacks. The characteristics like timeliness and fault tolerance are not commented and considered by the authors.

Awad [14] has presented the intrusion detection system using the classification techniques and association rule mining. The authors have claimed the higher accuracy rates with low false positives for known as well as unknown attack detection. The authors have considered the decision tree classifier as classification technique and association rule mining for unknown attack classification. The time required for execution is not considered in this work, where the combination of these two techniques have higher time complexity and also the algorithm executed has a sequential structure which ensures longer execution time.

The important highlights of the survey are the two important characteristics such as timeliness and fault tolerance are not addressed by the different researchers while designing and implementing the intrusion detection system. Keeping in mind these two important characteristics, we have proposed and implemented the significant intrusion detection architecture using combination of machine learning techniques and big data environment like Hadoop.

3. OBJECTIVES

Following are the objectives of this research work:

(1) The design of the Two phase- Intrusion Detection System (TP-IDS) for increasing accuracy with second phase as verification phase of the model using machine learning techniques.

(2) The implementation of TP-IDS phase I using SVM and kNN, phase II using Decision Tree (DT) and Naïve Bayes (NB), the effective IDS model for increasing accuracy and reducing FPR, FNR.

(3) The TP-IDS model building and execution using Hadoop Distributed File System (HDFS) for increasing speed of intrusion detection model, achieving timeliness in detection and making TP-IDS fault tolerant.

4. DEFINITIONS AND CONCEPTS

In the proposed architecture the combination of different machine learning techniques is used. Also, keeping in mind

the important improvement in the values of properties like timeliness and fault tolerance the Hadoop Distributed File System (HDFS) is used. This section is introducing all these techniques and HDFS.

Here, we are introducing four important machine learning techniques and a HADOOP system as follows:

1. Support Vector Machine (SVM)
2. k Nearest Neighbor (kNN)
3. Decision Tree
4. Naïve Bayes
5. HADOOP.

4.1 Support vector machine

Support Vector Machine is a supervised machine learning algorithm used as a classification algorithm. It is abbreviated as SVM. SVM uses the hyper plane for classifying the data points into separate classes [15]. The data points which are near the hyper plane and boundary points of the differentiating classes are called as support vectors, as these are used to maximize the distance between the hyper plane and the support vectors to bring the accuracy in the classification of the data [16]. It is one of the successful algorithms of machine learning and implemented in different classification problems by different researchers. SVM supports both the binary as well as multiclassification of the data points.

4.2 k nearest neighbour

K nearest neighbor is the machine learning techniques used or classification. Sometimes, it is also used as kNN clustering, because of its working structure [17]. kNN is one of the faster classification techniques. It is popular because of negligible training time requirement. In this technique the algorithm trains itself with every input value passed for the classification, hence no separate training phase is required resulting in faster execution of the algorithm [18]. During classification of the input data, the closeness of the data item to be classified is measured as a distance with other classified data items of different classes and the data item will be classified to the class, where the distance is minimum with more number of data points in a particular class. Hence, kNN works based on similarity measure of the data points and very useful in classifying the unknown data points in most of the cases.

4.3 Decision tree

Decision Tree is the supervised machine learning technique used for multiclassification of the input data. Decision tree are constructed with decision nodes as internal nodes and the class labels as the leaf nodes of the tree. Each internal node is defined for the decision for classification path, based on the values of different features used for classification as input variables [19]. There are number of algorithms used for construction of the decision tree such as ID3, C4.5, VFDT, etc. During training phase, with the available input and output data values, the decision tree is constructed. During testing phase, based on the input values of the features, the path of the output class is decided and finally the input is classified [20]. Decision trees are popular for their multiclassification property, as well as speed of the execution with accuracy in results.

4.4 Naïve Bayes

Naïve Bayes is the probability distribution classifier machine learning technique. Naïve Bayes calculates the prior and posterior probability values of the data items for classification purpose [21]. It calculates the likelihood that, the data item will be classified to a specific class. It is popular because of the probability based classification, which provides the base for the accuracy improvements in classification [22]. It is very useful when the data items to be classified have independent existence for their class and does not have any dependency with other objects of the same class.

These four machine learning techniques can be very effective when used in combination for developing a model for intrusion detection system. To compensate the time complexity of execution of these techniques and achieve the timeliness, a big data environment called as HADOOP is used in this research work [23]. It works as given in the paragraph below.

4.5 HADOOP

HADOOP is the distributed data processing architecture. To manage and process the big data which is generated with a huge speed in today's world, this massively parallel data storage and processing architecture is used. It stores the data in distributed file system. Each file is divided in number of blocks which are stored in distributed manner in different data storage nodes. The file system used for storing and managing data is called as Hadoop Distributed File System (HDFS) [24]. HDFS mainly consists of the two important types of nodes name node and data nodes. Name node acts like a master node which manages the number of data nodes. Data nodes are responsible for storing and processing the data in terms of file blocks. The name nodes are responsible to divide file into blocks and distribute blocks in available data nodes [25]. The Figure 2 shows the analogy of TP-IDS using HDFS architecture.

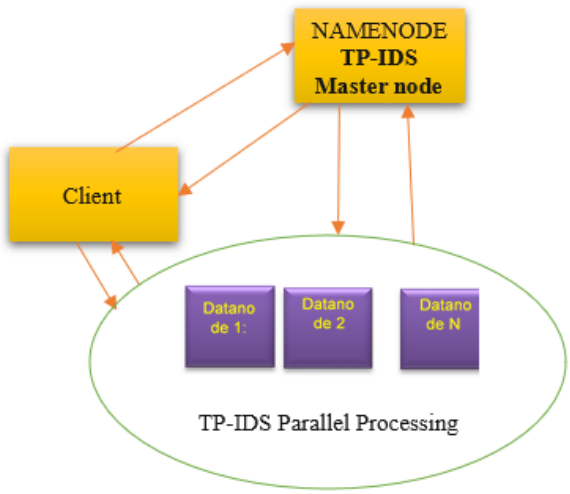

Figure 2. TP-IDS Parallel Design using HDFS

Name node manages and stores the Meta data of these data nodes block storage. Always the request of data storage or data retrieval is first sent to the name node, which in turn manages the execution of the request by using the Meta data available at name node. Meta data consists of the information of the data

nodes, availability of the memory at data nodes, mapping of files and blocks stored at different data nodes at specific memory locations. The general HDFS structure has thousands of such name nodes and each name node has thousands of data nodes for storage and processing of the data. Hence, it is a massively parallel data processing architecture. It reduces the execution time to greater extent and hence can be effectively used in time sensitive applications like intrusion detection system.

5. SOLUTION METHODOLOGY

The Figure 3 shows the designed architecture of the intrusion detection system in two phases, called as Two Phase - Intrusion Detection System (TP-IDS). The model of the TP-IDS consists of machine learning techniques such as Support Vector Machine (SVM), k Nearest Neighbor (kNN), Decision Tree (DT) and Naïve Bayes (NB) in two phases. In Phase I, SVM and kNN are used, in Phase II, DT and NB are used. The incoming connection request is first sent to Phase I techniques SVM and kNN. These algorithms are executed in parallel in distributed and parallel environment of HDFS, which ensures faster execution of these techniques. The incoming connection request is passed as input to SVM and kNN to classify it as either attack or normal connection request. If both the algorithms i.e. SVM and kNN classify it as a normal connection request, then access to the network is allowed and connection request is accepted. Else, if either of the techniques SVM or kNN detects incoming connection request as malicious one, then it will be blocked and will be sent to Phase II classification techniques.

In Phase II, we have used DT and NB machine learning techniques. This phase is used to verify the results of Phase I of TP-IDS, when either of the SVM or kNN detects incoming connection request as attack traffic. This Phase II as verification phase, helps us to reduce the false negatives and false positives in intrusion detection. After receiving the input from Phase I, in Phase II, algorithms DT and NB are executed in parallel in parallel HDFS environment. Again, due to distributed nature of the HDFS, faster execution of the Phase II techniques is ensured. After execution, if DT or NB, either of these techniques detects incoming connection request as malicious one, then it will be classified as an attack and access to the network will be blocked for that node. If, both of the Phase II techniques i.e. DT and NB, detects the incoming connection request as a normal connection request, then connection request is accepted and access to the networks is allowed. So, in Phase II, we verify that, whether the results classification obtained in Phase I are correct or not, and based on classification in Phase II, the decision is taken. With this structure, we ensure accuracy in intrusion detection.

The algorithms are executed in HDFS environment, which ensures that, though any of the processing node fails, there are other data nodes available to provide the data and processing for execution, achieving the property of fault tolerance. Hence, due to this distributed structure, system will always be up even in case of failures. So, HDFS helps us to achieve the timeliness and fault tolerance in intrusion detection model such as TP-IDS. This model is implemented with and without HDFS and significance of using HDFS is also observed. The significant results about the same are explained in results and discussion section.

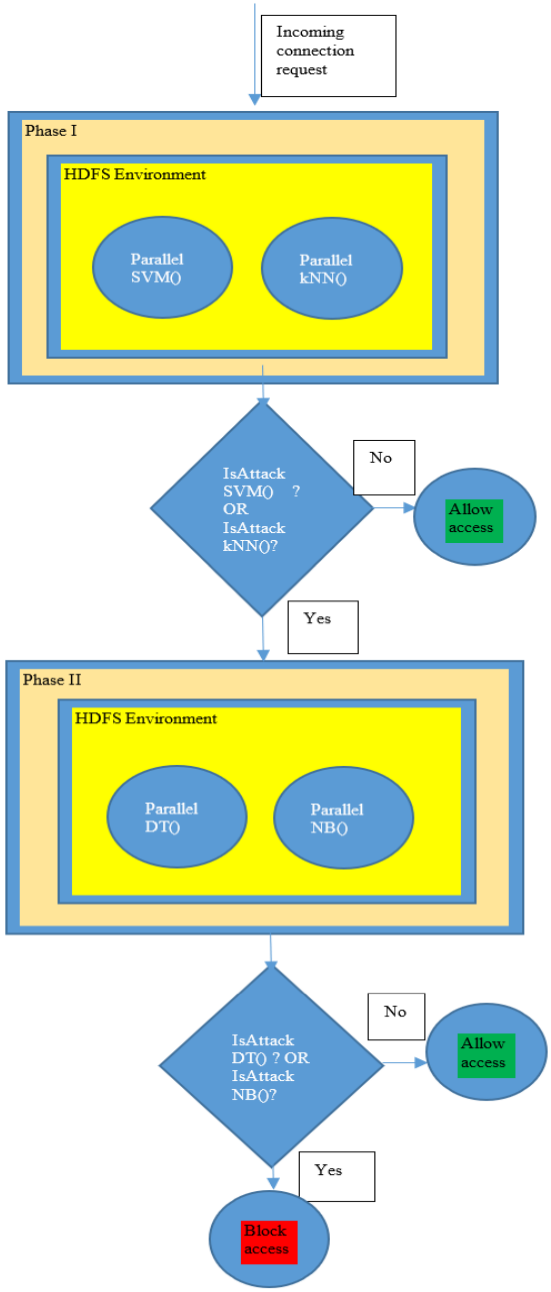

Figure 3. TP-IDS architecture using HDFS

6. DATA SET AND DATA PREPROCESSING

For this implementation of TP-IDS, the dataset used is NSL KDD dataset which consists of 43 attributes, including the output class label. It consists of data records of approximately 1, 25, 973 training samples, which are used for training of the TP-IDS model. It consists of 22,554 testing samples which are used for testing the accuracy of the TP-IDS model. The NSL KDD dataset does not have redundant record entries and consists of entries of all attack types [26]. NSL KDD consists of mainly records of Normal, DOS, Probe attack types, R2L, U2R. The Figure 4 shows the detail statistics of the entries available in NSL KDD dataset [27]. Figure 4 shows the attack types and number of records for each attack type in NSL KDD Dataset. The NSL KDD dataset is constructed to overcome the disadvantages of the KDD dataset, the advantages of NSL KDD dataset are as follows:

1. As this dataset does not have redundant entries, the classifiers will not be biased towards more frequent records.
2. With no duplicate records, learners are not biased by the better detection rate methods.
3. It helps to accurately evaluate different machine learning methods.
4. It helps to generate the efficient and consistent research results.

Every machine learning model algorithm should start with the feature selection as the data preprocessing part in the system [28]. Here, the feature selection allows to select only relevant features, hence increase the accuracy as well as speed of the model. To gain the desired accuracy in the model, it is important to reduce the irrelevant features from the dataset and use only highly correlated features for better model development [29].

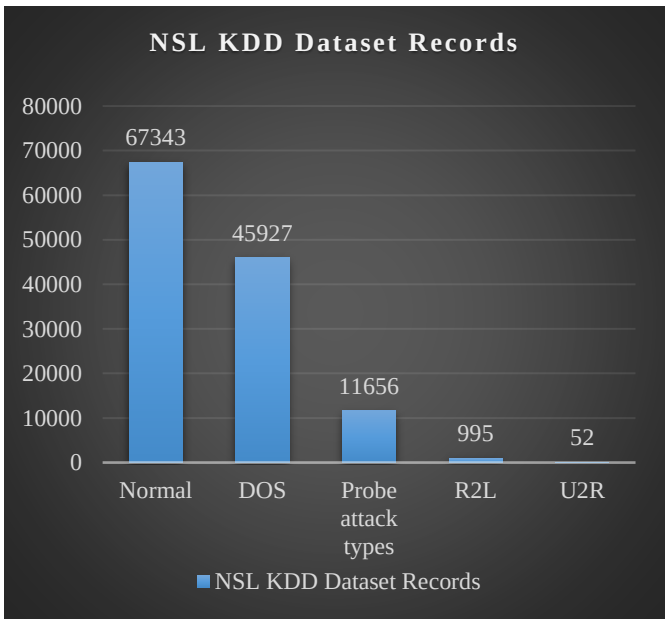

Figure 4. NSL KDD Dataset attack types and records

Hence, data preprocessing is important, before we proceed for model training and testing. As a part of data preprocessing, the correlation based feature selection can be used which helps for effective feature selection. Correlation-based feature selection (CFS) uses a heuristic assessment function based on correlations to rank attributes [30]. The function evaluates attribute vector subsets that are correlated with the class label but not with each other. The CFS algorithm considers that irrelevant features have a low correlation with the class and should thus be discarded [31]. The CFS algorithm is applied to the NSL KDD dataset with 41 attributes, out of which 29 attributes have shown high correlation with the output variable and thus remaining attributes having low correlation are eliminated from the used dataset. The results obtained by the researchers who have used feature selection on NSL KDD dataset are proving the importance of feature selection, as results are with increased accuracy [32]. Machine learning techniques like classification and clustering gives good results after using the feature selection process. The performance of the machine learning techniques is directly proportional to the attribute correlation [33]. More the number of highly correlated attributes, better is the performance. Hence, here the use of correlation based feature selection is promoted.

7. RESULTS AND DISCUSSION

The model is built and executed in RHIPE environment. RHIPE is R and Hadoop Integrated Platform environment. It is observed from the results that, the execution time is significantly reduced when we use the HDFS as underlying storage as well as processing environment for TP-IDS system. The accuracy observed during the testing of the model is almost 99.93%.

Table 1. TP-IDS Accuracy comparison with existing IDS models

Sr. No.	IDS Name	Accuracy in (%)
1	RBC-IDS [7]	99.12
2	SSC-OCSVM [8]	98.56
3	IDS – Decision Tree [12]	99
4	IDS – ANN [13]	94.02
5	**TP – IDS [Our Model]**	**99.93**

The accuracy obtained is significantly important, as it has reduced the false positives and false negatives both. The results of the accuracy are as shown in Table 1. The table shows the correct identification of the records according to their type and better accuracy results as compared to existing solutions.

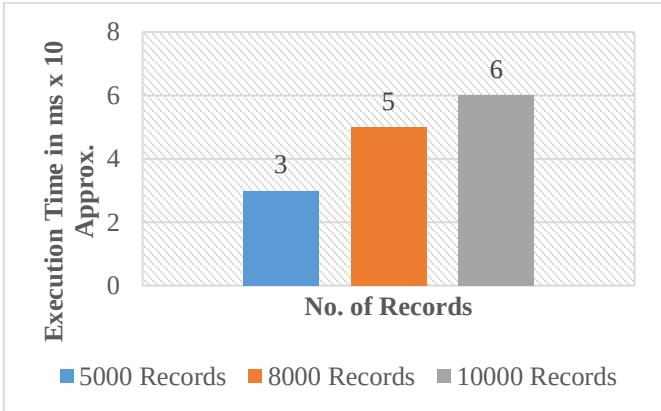

Figure 5. Execution Time Vs No. of Records on HDFS

The time execution or speed of the system is as shown in Figure 5. The speed results can be better than what presented in the graph, as the results are generated in simulation environment of the HDFS creating multimode environment on a single node. Hence, in real systems this timeliness can be easily achieved, if we are using HDFS system. Also, the HDFS is enabling the fault tolerance of the system, as it is a multi-node system, removing the dependency of the centralized node in the TP-IDS system. The data is preprocessed for removing noisy data, the feature selection is used to select only relevant features and increase the accuracy of the system. The Single feature selection technique is used for the same. The NSL KDD data set with 41 features is used, out of which only 29 relevant features are used for the execution of the TP-IDS model. The results obtained here by TP-IDS model are very effective from every quality attribute perspective. The accuracy obtained here, is significant as compared to existing models of IDS. The two phase model provides benefit of second phase, as verification phase and increases the accuracy of the IDS model. The time required for execution is really negligible for even large number connection requests, because

of the use of HDFS as underlying distributed storage and parallel data processing structure.

Also, with the use of HDFS, we have primary name nodes which are connecting thousands of data nodes and managing these nodes effectively. With the effective strategy of the replication of the data into multiple data nodes, removes the dependency of the data on the centralized data nodes or single data node. So, even in absence of any data node, data will be available from other data nodes where same data is replicated, making availability of the data even in failure of few data nodes, enabling IDS system as fault tolerant TP-IDS. Also, each primary name node has a backup secondary name node which is the backup node of the primary name node. Hence, even in case of failure of name nodes, the backup nodes are available and keep the system up, which is the fault tolerant TP-IDS system. Using two phases is of prime importance, because second phase as verification phase, ensures the low rates of FPR and FNR increasing accuracy in intrusion detection using TP-IDS.

8. CONCLUSION AND FUTURE SCOPE

So, hereby it is concluded that, to achieve the timeliness and fault tolerance with high accuracy, the use of big data environment like HDFS is efficient. In this research work, the Two Phase Intrusion Detection Model is used, which helps in increasing the accuracy by using Phase II for verification of the classification results of Phase I. When the accuracy results are observed, the significant findings are found as compared to the existing systems using standalone machine learning techniques for IDS used in one phase only. With this Two Phase architecture the false positives and false negatives are reduced. With the use of HDFS, the execution speed is increased to the greater extent due to parallelism in data processing for attack detection. HDFS as it is distributed environment, also ensures fault tolerance in the implemented TP-IDS model. Here, the work is completed in the simulation based environment of HDFS and network data sets are used. In future, the real time environment can be created and algorithms can be executed to better understand the results and subsequent efforts can be made to improve the same. Also, the future work can be extended with replacement of few of these traditional machine learning techniques with deep learning or more advanced and latest techniques in the area.

ACKNOWLEDGMENT

We are thankful to our respected Research and Development Department Head, members and all respected faculty members of Department of Computer Science and Engineering, Koneru Lakshamaiah Education Foundation, Guntur, A. P., India.

REFERENCES

[1] Chkirbene, Z., Erbad, A., Hamila, R., Mohamed, A., Guizani, M., Hamdi, M. (2020). TIDCS: A dynamic intrusion detection and classification system based feature selection. IEEE Access, 8: 95864-95877. https://doi.org/10.1109/ACCESS.2020.2994931

[2] Dutt, I., Borah, S., Maitra, I.K. (2020). Immune system based intrusion detection system (IS-IDS): A proposed model. IEEE Access, 8: 34929-34941. https://doi.org/10.1109/ACCESS.2020.2973608.

[3] Pan, S., Morris, T., Adhikari, U. (2015). Developing a hybrid intrusion detection system using data mining for power systems. IEEE Transactions on Smart Grid, 6(6): 3104-3113. https://doi.org/10.1109/TSG.2015.2409775.

[4] Zhong, W., Yu, N., Ai, C. (2020). Applying big data based deep learning system to intrusion detection. Big Data Mining and Analytics, 3(3): 181-195. https://doi.org/10.26599/BDMA.2020.9020003.

[5] Zhang, J., Zulkernine, M., Haque, A. (2008). Random-forests-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(5): 649-659. https://doi.org/10.1109/TSMCC.2008.923876.

[6] Han, S., Xie, M., Chen, H.H., Ling, Y. (2014). Intrusion detection in cyber-physical systems: Techniques and challenges. IEEE Systems Journal, 8(4): 1052-1062. https://doi.org/10.1109/JSYST.2013.2257594.

[7] Otoum, S., Kantarci, B., Mouftah, H.T. (2019). On the feasibility of deep learning in sensor network intrusion detection. IEEE Networking Letters, 1(2): 68-71. https://doi.org/10.1109/LNET.2019.2901792.

[8] Pu, G., Wang, L., Shen, J., Dong, F. (2020). A hybrid unsupervised clustering-based anomaly detection method. Tsinghua Science and Technology, 26(2): 146-153. https://doi.org/10.26599/TST.2019.9010051.

[9] Wang, M., Zheng, K., Yang, Y., Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. IEEE Access, 8: 73127-73141. https://doi.org/10.1109/ACCESS.2020.2988359.

[10] Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T., Al-Hashida, A.Y. (2018). Intrusion detection model using machine learning algorithm on Big Data environment. Journal of Big Data, 5(1): 1-12. https://doi.org/10.1186/s40537-018-0145-4

[11] Bankapalli, V.R., Akkalakshmi. (2020). Network intrusion detection using Deep Learning techniques. International Journal of Advanced Science and Technology, 29(6): 8278-8287. http://sersc.org/journals/index.php/IJAST/article/view/25273.

[12] Taher, K.A., Jisan, B.M.Y., Rahman, M.M. (2019). Network intrusion detection using supervised machine learning technique with feature selection. In 2019 International conference on robotics, electrical and signal processing techniques (ICREST), 643-646. https://doi.org/10.1109/ICREST.2019.8644161

[13] Jabez, J., Muthukumar, B. (2015). Intrusion Detection System (IDS): Anomaly detection using outlier detection approach. Procedia Computer Science, 48: 338-346. https://doi.org/10.1016/j.procs.2015.04.191

[14] Awad, N.A. (2021). Enhancing network intrusion detection model using machine learning algorithms. CMC-Computers Materials & Continua, 67(1): 979-990.

[15] Bhargava, M.G., Vidyullatha, P, Rao, P.V., Sucharita, V. (2018). A study on potential of big visual data analytics in construction Arena. International Journal of Engineering and Technology (UAE), 7: 652-656. https://doi.org/10.14419/ijet.v7i2.7.10916

[16] Babu, S., Mohan, K., Bano, S. (2018). Navigation usability improvement by using actual and anticipated data. Journal of Advanced Research in Dynamical and

Control Systems, 10(4): 478-482. .
[17] Mehrotra, S., Kohli, S., Sharan, A. (2019). An intelligent clustering approach for improving search result of a website. International Journal of Advanced Intelligence Paradigms, 12(3-4): 295-304.
[18] Rama Rao, K.V.S.N., Sivakannan, S., Prasad, M.A., Agilesh Saravanan, R. (2018). Technical challenges and perspectives in batch and stream big data machine learning. International Journal of Engineering and Technology (UAE), 7(1): 48-51. https://doi.org/10.14419/ijet.v7i1.3.9225
[19] Jadhav, A.D., Pellakuri, V. (2021). Highly accurate and efficient two phase-intrusion detection system (TP-IDS) using distributed processing of HADOOP and machine learning techniques. J Big Data, 8: 131. https://doi.org/10.1186/s40537-021-00521-y
[20] Nadh, V.L., Prasad, G.S. (2018). Support vector machine in the anticipation of currency markets. Int. J. Eng. Technol, 7(2-7): 66.
[21] Jadhav, A.D., Pellakuri, V. (2020). Efficient intrusion detection systems using machine learning approach for sustainable IT development. Journal of Green Engineering, 2020, 10(10): 8298-8310.
[22] Vishwakarma, P.P., Tripathy, A.K., Vemuru, S. (2018). A layered approach to fraud analytics for NFC-enabled mobile payment system. In International Conference on Distributed Computing and Internet Technology, 127-131. https://doi.org/10.1007/978-3-319-72344-0_9
[23] Patel, A.K., Chatterjee, S., Gorai, A.K. (2019). Development of a machine vision system using the support vector machine regression (SVR) algorithm for the online prediction of iron ore grades. Earth Science Informatics, 12(2): 197-210. https://doi.org/10.1007/s12145-018-0370-6
[24] Vasantham, V., Haritha, D. (2018). A Survey on cost minimization techniques for big data processing. Journal of Advanced Research in Dynamical and Control Systems, 10(2): 547-551.
[25] Raghav, R.S., Amudhavel, J., Dhavachelvan, P. (2017). A survey on tools used in big data platform. Adv Appl Math Sci., 17(1): 213-229.
[26] Alzahrani, A.O., Alenazi, M.J. (2021). Designing a network intrusion detection system based on machine learning for software defined networks. Future Internet, 13(5): 111. https://doi.org/10.3390/fi13050111
[27] Thomas Rincy, N, Roopam, G. (2021). Design and development of an efficient network intrusion detection system using machine learning techniques. Wireless Communications and Mobile Computing, 2021: 9974270. https://doi.org/10.1155/2021/9974270
[28] Ravi, N., Ramachandran, G. (2020). A robust intrusion detection system using machine learning techniques for MANET. International Journal of Knowledge-based and Intelligent Engineering Systems, 24(3): 253-260. https://doi.org/10.3233/KES-200047
[29] Khraisat, A., Gondal, I., Vamplew, P. et al. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2: 20. https://doi.org/10.1186/s42400-019-0038-7
[30] Wosiak, A., Zakrzewska, D. (2018). Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. Complexity, 2018: 2520706. https://doi.org/10.1155/2018/2520706
[31] Kowshalya, A.M., Madhumathi, R., Gopika, N. (2019). Correlation based feature selection algorithms for varying datasets of different dimensionality. Wireless Personal Communications, 108(3): 1977-1993. https://doi.org/10.1007/s11277-019-06504-w
[32] Sstla, V., Kolli, V.K.K., Voggu, L.K., Bhavanam, R., Vallabhasoyula, S. (2020). Predictive model for network intrusion detection system using deep learning. Revue d'Intelligence Artificielle, 34(3): 323-330. https://doi.org/10.18280/ria.340310
[33] Nalavade, K.C. (2020). Using machine learning and statistical models for intrusion detection. International Journal of Computer Applications, 175(31): 14-21. https://doi.org/10.5120/ijca2020920854