International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Implementation of Intrusion Detection System Using Various Machine Learning Approaches with Ensemble learning

Ms. Pragati V. Pandit, Research Scholar, SJJTU Jhunjhunu, Rajasthan, pragativpandit2918@gmail.com

Dr. Shashi Bhushan, Research Guid, SJJTU Jhunjhunu, Rajasthan, tyagi_shashi@yahoo.com

Dr. Uday Patkar, Research Co-Guid, SJJTU Jhunjhunu, Rajasthan, udaypatkarcomp”gmail.com

ARTICLEINFO

Article History:

Accepted: 01 Aug 2023

Published: 08 Aug 2023

Publication Issue

Volume 9, Issue 4

July-August-2023

Page Number

461-466

ABSTRACT

Recent years have seen an increase in advanced threat attacks, yet feature filtering-based network intrusion detection systems have a number of shortcomings that make it challenging for security managers and analysts to identify and thwart network intrusions in their organizations. Information systems are routinely protected and damage is minimized using techniques for detecting intrusions. It protects against dangers and weaknesses in real-world and virtual computer networks. Effective intrusion detection systems are now typically created using machine learning techniques. Neural networks, statistical models, rule learning, and ensemble techniques are examples of machine learning techniques for intrusion detection. Machine learning ensemble techniques are renowned for their superior performance during the learning process. For the creation of a successful intrusion detection system, a suitable ensemble technique must be investigated. In this paper, we introduced a novel ensemble method for intrusion detection in the network along with a combination of decision tree, random forest, extra tree, and XGBoost algorithms. The suggested method was created utilizing the Python programming language and aids in improving detection accuracy. Utilizing the CICIDS2017 dataset, the constructed system is evaluated based on numerous evaluation criteria, including precision, recall, and f1-score. The ensemble approach significantly raises the detection accuracy.

Keywords—Intrusion detection, Machine Learning, ML algorithms, ensemble learning, Random Forest, Decision Tree, XgBoost, Extra Tree.

I. INTRODUCTION

Machine Learning techniques is been applied in many intrusion detection systems due to their freely available qualities, which may permit them to logically understand complicated harmful and normal patterns. Previous research indicates that the majority

of machine learning-based intrusion detection systems have accuracy issues, with high false alarm and poor detection rates [1]. More crucially, it was shown that the most majority of early ML-based IDS techniques failed to work properly into the real-world because of the insufficient dataset utilised in creating such models. Previous approaches relied on datasets that were widely criticised as being out of date and failing to reflect current network trends and the sophistication of everevolving incursions [2]. Intrusion Detection Systems is very important parameter design for a secure network [3]. The learning model and dataset quality are closely connected to an IDS's efficiency. Many research have relied on datasets that have recognised flaws [4]. The issues involve, but are not limited to traffic of the attack, privacy depend on ethics, redundancy and traffic which is simulated that is not from a genuine production network, and a lack of traffic variety and an all-encompassing dataset [4]. It supplements firewall services by observing network traffic packets, analysing and comparing previously stored routine events with suspect patterns in order to increase network and information resource security [5]. Through supervising the entering and also existing traffic of the network and recording any of the malicious activity the intrusion detection system controls, protects and many times helps to prevent malicious attacks [6][7]. The popular Machine Learning technique which combines the capabilities of different basic classifiers to create a classification model with improved overall categorization or prediction strength is known to be Ensemble learning. This method has been proved to perform better for IDS than individual classification models known as base learners. By understanding the properties of ensemble learning technique we used it in our proposed system to combine various algorithms. The model's performance is assessed using the CICIDS2017 dataset, which is one of the most recent and reliable intrusion detection datasets. This types of dataset represents the real-world network scenario. We develop these model to find the intrusion in network system by utilising the ensemble method. The ensemble method combines the properties of various ML algorithms. The proposed ML algorithms are DT, RF, ET and XgBoost is used to improve the performance of our proposed model.

II. LITERATURE SERVEY

With an emphasis on machine learning techniques for intrusion detection systems, we assess some related literature. The authors of [9] used Association Based Classification in the construction of the intrusion detection system. The Apriori method can run more quickly without losing any information by removing unnecessary steps from the rule induction process. The fuzzy association rules are applied to build descriptive models of various classes. The recommended classifier is efficient for categorising large datasets and can handle symbolic attributes. Unsupervised machine learning has been proposed by the authors of [10] as a method for classifying network data and locating applications. The authors used a feature selection technique to choose the best collection of flow features. This statistical characteristic of flow is used in a network to classify and identify packets. It is also determined how certain traits affect learning. Researchers have published a framework in [11] for the experimental assessment of a classifier for an intrusion detection system. A single class v-SVM classifier was addressed by the authors using an RBF kernel.

They advised adding simulated attack samples to the training data to make discriminative SVM classifiers more secure. The authors of [12] have built an enhanced incremental HMM stochastic process for an intrusion detection system. The pre-processing of the data speeds up a hidden Markov model. The system is a system call-based anomalous intrusion detection system. [13] has suggested a clustering-based classification strategy for network anomaly detection. Both supervised classification and unsupervised

incremental clustering are applied during training. By segmenting a labelled training dataset into several clusters, this technique may reveal the profile of both typical and anomalous packets. Supervised classification is used to test the tagging of objects with cluster characteristics. [14] has presented the SbSVM approach for intrusion detection. For situations where the class distribution lacks the imbalance required by SVM algorithms, an autonomous labelling approach The autonomous labelling system was created by SNORT.

The proposed SbSVM is more trustworthy. The intrusion detection system was constructed using the min-max normalisation technique by researchers in [15]. The KDD99 intrusion data is adjusted before being sent to the SVM. They found that normalisation can speed up computation and provide a classifier good performance. In this work, a few normalising methods are investigated and simulated. They find that min-max normalisation works better than other normalising methods in terms of precision and speed. The authors of this article [16] have created an intrusion detection system using an enhanced incremental HMM stochastic process. A GANN-based intrusion detection system was proposed by researchers in [17]. Genetic Engineering The Weight Extraction Algorithm is used to gather and enhance the weights between the neurons in order to accurately identify intrusions. [18] have suggested utilising unsupervised machine learning to categorise traffic. The authors used a feature selection technique to choose the best collection of flow features. A decision tree algorithm used in data mining has been proposed by [19]. The classification technique is used to inductively learn a model from the pre-classified data set. Each item of data is defined by the values of the attributes. A mapping from a set of qualities to a particular class can help you understand classification. The decision tree classifies the given data item based on the values of its characteristics. METHODOLOGY A. Block diagram The use of the ensemble approach in intrusion detection has grown in importance since it aids in determining new assaults in addition to recognised ones. As a result, an ensemble-based IDS based on Machine learning methods is provided in this suggested system. The suggested system's block diagram displays the many stages of the process.

The experiment was run using the CICIDS2017 dataset. For evaluating the effectiveness of the proposed approach these dataset is used. We used Python programming language to implement our proposed system. By comparing the Decision Tree, Random Forest, Extra Tree and XgBoost algorithm the accuracy or the proposed system is measured. Previously, this approach was utilised to construct collaborative IDS, through which the properties may flexibly choose to be measured and improves the power of prediction. Using the given following parameters the performance of the proposed system is measured.

- The precision is the parameter which is used to find the ability of the classifier for accurately identifying the instances which are not a false negative and which are negative.
- The parameter precision is the classifier's ability to accurately identify a negative occurrence and not FN.
- The capability of the classifier to accurately identify all positives examples is the value of recall. • After adding the value of accuracy and recall the result obtain is known to be f1-score parameter.

Our proposed intrusion detection system initially imports the ML libraries using Python programming. The CICIDS2017 dataset is read in the next step. Then the intrusion testing of the dataset is done. After detecting the intrusions of the dataset the empty or null values is get filled by zero. After that the pre-processing and the training of the dataset is done. The SMOTE library is used to learn the imbalanced in the system. After that the DT, RF, ET and XgBoost algorithms test the dataset through the ensemble

learning. Each algorithm of the proposed system individually train and test the dataset. The model performs collecting of data, selecting the features, pre-processing, and model for training, testing, and validation. The output of these algorithms is get merge through ensemble learning. The final performance of the system is get measured through precision, accuracy, f1-score and recall parameters. After all of these the final result of the proposed model is get printed.

Training and Testing: 1. DT (Decision Tree): A popular technique of categorization that can efficiently categorise the data is DT [8]. A DT is made up of combining the terminal and non-terminal nodes. The root node is the first characteristic of the DT. The root node have test condition of splitting every input data record towards the each internal node. This is based on the characteristics of data record. Before it can categorise fresh or untrained data, the DT must first be trained with known data. The Decision Tree is built throughout the training phase [9] by specifying the qualities and values that will be used to assess the input data at each internal node. The tree may predict or categorise incoming data after training. By traversing internal nodes based on test condition features until it reaches a leaf node with a class response [10]. By assuming that total count of leaf nodes in DT T is |T|, and that t is one of them. This node has Nt samples, and the number of sample points of k is Nk. Ht is the leaf node's entropy, and (0) is an optional parameter connected to the punishment term. As a result, Decision Tree T's loss function is as follows: ∝ (T) = (1) In which the entropy is calculated as follows: τ (T) = (2)

RF (Random Forest): The RF approach of Machine Learning method is used for categorizing and declining issues. The RF approach is made up of on the concept of ensemble learning which includes combining the various algorithms which have ability to solve major complications and helps to increase the models performance. We will use Python to construct the Random Forest Algorithm tree. K decision trees serve as the foundation for the random forest model. Each tree votes on which class a given independent variable X belongs to, and the class with the most votes wins [11][12][13]. The K decision trees are described as follows: (3) From the equation (3), the symbol {Øk} represents the randomly distributed identical random vectors, and each tree. At input x, there is a vote for the most renowned class. The nature and dimensions of Ø are determined by on its application in tree construction [14].

ET (Extra Tree algorithm): A fully random tree classification is known to be extra tree classification [15]. The ET is a type of method which creates many decision trees. But here the sampling of each tree is random where the samples cannot be replaced. It determines the best cutoff point for each and every node for random feature [16]. It lessens the computational strain on typical trees and forests in determining optimal cut-offs. ι = (4) The equation [F= {(x) =wq(x)} (q:Rm→T,w € RT)] is defined as the space that is a part of the regression trees. The equation (4), defines the tree structure which is represented by q. The T represents the number of leaves in the tree in the equation (4), whilst each f k corresponds to a tree structure, that is, q, and leaf weights, that is, w. 4. XGBoost Algorithm: XGBoost is a sort of algorithm that was created to increase the systems speed and performance. It has dominated in the field of applied ML. To be very efficient, adaptable, and portable XGBoost is a distributed gradient boosting toolkit that has been developed [17].

For constructing the ML algorithms the XgBoost method uses the Gradient Boosting structure. The weight is the important factor in the XgBoost method. For predicting the result the weight is given to the all independent variable which are then provided to DT. In boosting various models are sequentially created. The model was first created by guessing simply random variables, the second model was created by using the residuals. And then the new model was created by combining the first and second model. By

assuming that having a feature matrix X and a target variable Y.

Also, supposing the f0(x) is a model initially created only through random guessing. By using the difference between the goal of the outcome and the gusses which are random the second model is constructed which a nothing but a residual. Which is defined as: (5) In the tree for every leaf the average of the residual is computed. The second model was obtained as h0(x) = Mean of residuals or the model which is fitted on residuals as a result. Then the new model f1(x) will be, (6) Many iterations can be executed in the same way as the initial iteration was, (7) (8) 5. Ensemble Learning: The various approaches are get combined to give the solution which is very efficient more that any single algorithm is defined to be Ensemble learning method. The benefits of ensemble learning are that it is very efficient and not particular to individual assaults. Figure 2 depicts the ensemble technique process. In order to improve overall performance, we adopted ensemble learning approach. We integrated the outputs of all classifiers in training and testing phases by extracting the most valuable information from all classifiers.

III.CONCLUSION

In this study, we combined the DT, RF, ET, and XgBoost algorithms to present a novel ensemble intrusion detection system. Combining the classification capabilities of the algorithms used in the proposed system improves the MLbased IDS model, as evidenced by the average probability of the algorithms' predictive abilities. The CICIDS2017 dataset was one of the datasets used for the evaluation of the new model and was judged to be trustworthy and accurately reflect the current state of various threats. The suggested system is assessed using the f1score, recall, precision, and accuracy metrics. The results section demonstrates that the ensemble technique works better than the most advanced system. Future improvements can be made to enhance the effectiveness of intrusion detection attacks.

IV. REFERENCES

[1]. V. V. Mandhare, D. R. Pede, and P. S. Vikhe, “Network Intrusion Detection using a Deep Learning Approach,” Int. J. Recent Technol. Eng., vol. 9, no. 3, pp. 59–64, 2020, doi: 10.35940/ijrte.b4086.099320.

[2]. C. Sinclair, L. Pierce, and S. Matzner, “An application of machine learning to network intrusion detection,” Proc. - Annu. Comput. Secur. Appl. Conf. ACSAC, vol. Part F1334, no. 0293, pp. 371–377, 1999, doi: 10.1109/CSAC.1999.816048.

[3]. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, “An Investigation on Intrusion Detection System Using Machine Learning,” Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018, pp. 1684–1691, 2019, doi: 10.1109/SSCI.2018.8628676.

[4]. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy, 2018, vol. 2018-Janua, pp. 108–116, doi: 10.5220/0006639801080116.

[5]. S. M. Mehibs and S. H. Hashim, “Proposed Network Intrusion Detection System Based on Fuzzy c Mean Algorithm in Cloud Computing Environment,” J. Univ. BABYLON Pure Appl. Sci., vol. 26, no. 2, pp. 27–35, 2017, doi: 10.29196/jub.v26i2.471.

[6]. S. Rizvi, G. Labrador, M. Guyan, and J. Savan, “Advocating for Hybrid Intrusion Detection Prevention System and Framework Improvement,” in Procedia Computer Science, 2016, vol. 95, pp. 369–374, doi: 10.1016/j.procs.2016.09.347.

[7]. L. Sellami, D. Idoughi, A. Baadache, and P. Tiako, “A Novel Detection Intrusion Approach for Ubiquitous and Pervasive Environments,” in Procedia Computer Science, 2016, vol. 94, pp. 429–434, doi: 10.1016/j.procs.2016.08.066.

[8]. Z. Pan, S. Chen, … G. H.-P. of the 2003, and undefined 2003, “Hybrid neural network and C4. 5

for misuse detection," ieeexplore.ieee.org, 2003, doi: 10.1109/ICMLC.2003.1259925.

[9]. J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/bf00116251.

[10]. S. Chiu and D. Tavella, "Introduction to Data Mining," in Data Mining and Market Intelligence for Optimal Marketing Returns, 2020, pp. 151–206.

[11]. B. N. Kagara and M. Md Siraj, "A Review on Network Intrusion Detection System Using Machine Learning," Int. J. Innov. Comput., vol. 10, no. 1, pp. 27–34, 2020, doi: 10.11113/ijic.v10n1.252.

[12]. N. Farnaaz and M. A. Jabbar, "Random Forest Modeling for Network Intrusion Detection System," in Procedia Computer Science, 2016, vol. 89, pp. 213–217, doi: 10.1016/j.procs.2016.06.047.

[13]. Y. Y. Aung and M. M. Min, "An analysis of K-means algorithm based network intrusion detection system," Adv. Sci. Technol. Eng. Syst., vol. 3, no. 1, pp. 496–501, 2018, doi: 10.25046/aj030160.

[14]. [14] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[15]. B. S. Bhati and C. S. Rai, "Ensemble based approach for intrusion detection using extra tree classifier," in Advances in Intelligent Systems and Computing, 2020, vol. 1125, pp. 213– 220, doi: 10.1007/978-981-15-2780-7_25.

[16]. S. Patil et al., "Explainable Artificial Intelligence for Intrusion Detection System," Electron., vol. 11, no. 19, 2022, doi: 10.3390/electronics11193079.

[17]. X. Xu, W. Chen, and Y. Sun, "Over-sampling algorithm for imbalanced data classification," J. Syst. Eng. Electron., vol. 30, no. 6, pp. 1182–1191, 2019, doi: 10.21629/JSEE.2019.06.12