JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# INTRUSION PREDICTION AND DETECTION USING SUPERVISED MACHINE LEARNING TECHNIQUE

**[1]Archana.C, [2]Aadhilakshmi.A, [3]KeerthanaG, [4]Nayagan.S**
[1]*UG Student,* [2]*UG Student,* [3]*UG Student,* [4]*Assistant Professor*
[1]*Department of Computer Science and Engineering*
[1]*Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College ,Avadi, Chennai, India*

Abstract-***Recently, technology has been blooming rapidly, and the requirement for security is much more important. And so for the security of the network enterprises, we are using the intrusion detection system. An intrusion detection system is a device or software app that overlooks the inbound and outbound network traffic, continuously analyses for activity changes and patterns, and alerts an administrator. In this paper, we have analysed various methods to detect the threat in real time. Hence, we used a variety of algorithms to detect the intrusion and prevent the attack. With the help of a variety of algorithms, it not only predicts the attack but also finds what type it is. The design of attack prediction tools has always been dominated by statistical methodologies. The supervised machine learning technique (SMLT) helps in analysing the datasets for the collection of information, which includes variable identification, univariate, bivariate, and multivariate analysis, missing values, etc. The most effective machine learning algorithm for predicting the types of cyberattacks has been determined through a comparison study of different algorithms. The outcomes may be compared to the highest levels of accuracy, precision, recall, F1 score, sensitivity, and specificity. Finally, the best accuracy algorithm is used to predict the attack and its type.***
***Keywords:Intrusion detection, Attack prediction, Supervised Machine Learning Technique (SMLT), Machine learning.***

## 1. INTRODUCTION

The traditional approach to computer security is no longer adequate for the expanding, multidimensional, and interconnected network environment due to the rapid advancement of technology. Internet security issues have become increasingly prevalent worldwide as a result of the ongoing development of technologies like the Internet of Things and cloud computing, as well as the dawn of the era of big data. In this way, utilising AI innovation to dissect a lot of organisation traffic to decide the interruption conduct is a compelling method for upgrading the security of the organisation.

An intrusion detection system is a device or software app that overlooks the inbound and outbound network traffic, continuously analysing for activity changes and patterns, and alerts an administrator. When it detects unusual behaviour or unauthorised activity, an administrator reviews the alert or alarm and takes adequate action to remove the threat.

Unsupervised learning uses clustering algorithms to generate a model from unlabeled data. In this way, they can distinguish malicious inputs from network traffic or host logs. According to their statistical properties, unsupervised methods analyse the data characteristics randomly, without any prior knowledge.

## 2. LITERATURE SURVEY

[1] Ghada Abdelmoumin et al. discussed the use of optimisation techniques to enhance the performance of single-learner AML-IDS, such as PCA and 1-SVM AML-IDS models, for building efficient, scalable, and distributed intelligent IDS for detecting intrusions in the IoT.

[2] Darshana Upadhyay et al. discussed According to experimental findings, our approach outperforms earlier ones in terms of accuracy, precision, recall, F1 score, and miss rate. Additionally, the model is assessed using precision recall (PR) and receiver operating characteristic (ROC) plots for the binary, three-class, seven-class, and multi-class class categories, as well as for each of these class categories separately.

[3] Sugandh Seth et al. discussed The proposed approach is based on building an ensemble by ranking the detection abilities of different base classifiers to identify various types of attacks. The F1-score of an algorithm is used to compute the rank matrix for different attack categories. The final prediction algorithm's output for an attack is only considered if the algorithm has the highest F1-Score in the rank matrix for the particular attack category.

[4] Chuang Ma, Jiajun Zhang, et al. The experimental results show that, compared with traditional network attack detection methods, the detection performance of this method combined with kernel principal component analysis and the decision tree algorithm is better than the method using only the decision tree algorithm or principal component analysis combined with the decision tree algorithm.

[5] Gurdip Kaur et al. discussed Image classification is proposed based on a deep neural model to classify various attacks by using two comprehensive datasets called CICIDS 2017 and CSE-CICIDS 2018. Secondly, we provide a list of the best network flow features to identify these attacks. We deploy a convolutional neural network model to classify and characterise different attacks, with promising evaluation results.

[6] Timothy Chadza et al. discussed the paper, which proposes a transfer learning (TL) approach that exploits already learned knowledge gained from a labelled source dataset and an unlabelled target dataset. Five unsupervised HMM techniques are developed using a TL approach and evaluated against conventional machine learning approaches.

## 3. PROPOSED METHOD

This experiment was conducted with an actual setup by applying different attacks on a target network. Next, we have to collect the traffic data logs. At the last step, we have to apply our chosen machine learning algorithms. At the final stage, we need to detect and classify the data traffic based on predefined rules using our selected algorithms and observe its performance. The time, however, was not on our side to simulate such assaults on a real system that we had developed in order to see how the suggested algorithms would perform on it. Hence, we had to choose a ready-to-use dataset that contains one attack type similar to the one that we decided to model on the target network. We have taken the readymade dataset from Kaggle. The selected dataset corresponds to the traffic logs. We have followed the experiment steps by applying a machine learning process algorithm.

IN MACHINE LEARNING, THE FIRST STEP IS TO MAKE THE ALGORITHMS FAMILIAR WITH THE DATA THEY WILL RECOGNIZE. FOR THIS PURPOSE, WE DIVIDED THE DATASET INTO TWO PARTS, NAMELY, TRAINING DATA AND TESTING DATA. IN THAT CASE, WE PARTITIONED THE DATA INTO 70% IN THE TRAINING DATASET AND 30% IN THE TESTING DATASET. WE USED ANACONDA PLATFORM TO PERFORM THE SELECTED ALGORITHMS. WE SELECTED THE ONES THAT PERFORMED BETTER IN ANOMALY DETECTION AND CLASSIFICATION BASED ON THE PERFORMANCE METRICS VALUES OF ACCURACY, TRUE POSITIVE RATE, FALSE POSITIVE RATE, PRECISION AND RECALL. WE EVALUATE THE PERFORMANCE OF THE SELECTED ALGORITHMS TESTING THEM ON A DATA SET THAT CONTAINS SEVERAL ANOMALIES.

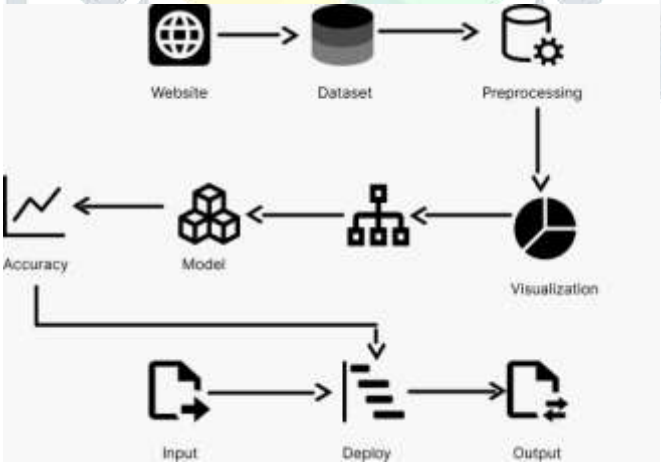


**Fig 3.1 The Experimental process**

We have used the below-mentioned algorithms for this experimental setup:

**SVM**: Support vector machine approach is supervised machine learning, which is a linear model for classification and regression issues. It is effective for many real-world issues and can solve both linear and non-linear problems. The fundamental goal is to establish a line, or hyperplane, that categorises the data into classes

**Random forest algorithm:** Random Forest is a machine learning method that belongs to the supervised learning technique. Its foundation is the idea of ensemble learning. To arrive at a single conclusion, it aggregates the results of multiple decision trees. Both classification and regression issues are supported.

**Adaboost Algorithm:** Adaptive Boosting is an ensemble approach of supervised machine learning. It can be applied to both classification and regression. One can use AdaBoost to improve an algorithm's performance. When teaching weak learners, it works best. These are models whose categorization accuracy is slightly better than random chance.

**Voting algorithm**: **A voting classifier is a supervised machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of choosing that class as the output.**

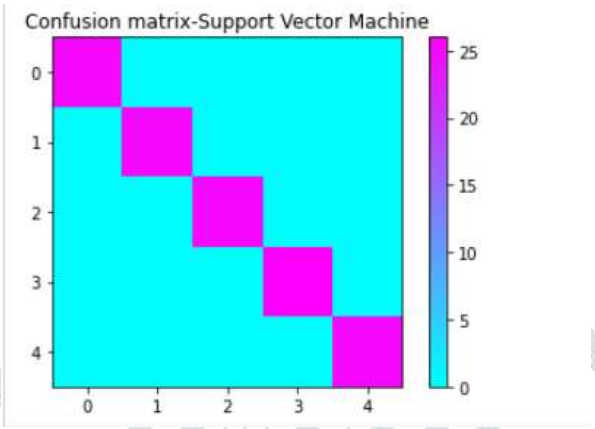


Fig 3.2 SVM algorithm

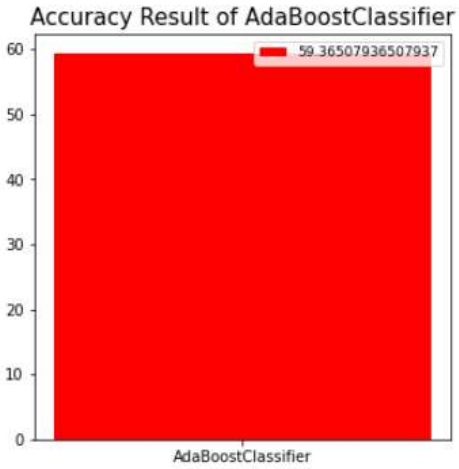


Fig3.3 Adaboost algorithm

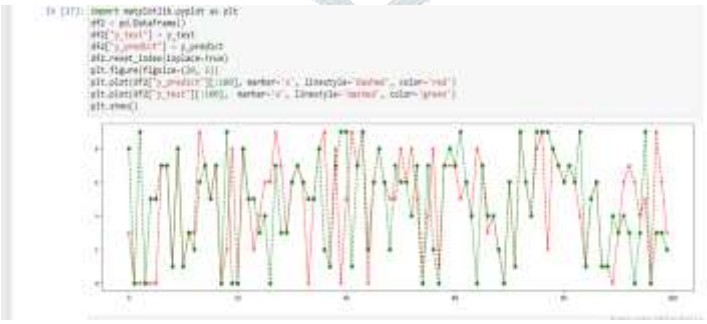

Fig3.4 Random forest algorithm

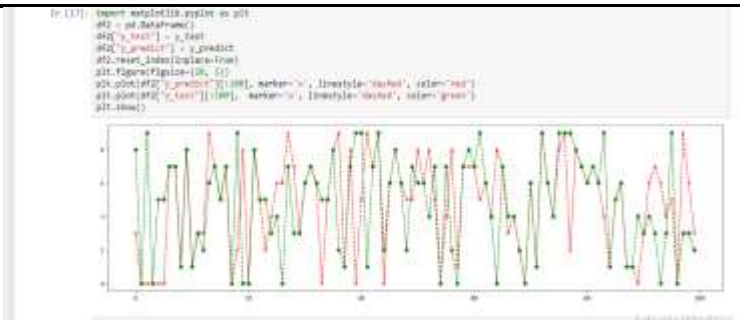

Fig 3.5 voting algorithm

The above-mentioned fig are the results of the algorithms we selected for the intrusion system.

**Process:**

**Step 1:** Downloading the Dataset

**Step 2:** Data Pre- processing

**Step 3:** Data Visualization

**Step 4:** Implementing Support Vector Machine Algorithm

**Step 5:** Implementing Random Forest Algorithm

**Step 6:** Implementing Ada Boost Algorithm

**Step 7:** Implementing Voting Classifier Algorithm

**Step 8:** The Algorithm ,with the best accuracy is implemented

**Step 9:** Deployment

**Step 10:** The coding part is coded in visual studio, an IP Address is created.

**Step 11:** Then the Login page gets opened. The user can login into the page

**Step 12:** When the user enter the details, the type of attack is predicted and the output is shown to the user.

## 4. RESULTS AND DISCUSSION

Performance evaluation metrics in machine learning algorithms evaluate the machine learning framework performance with the input data and help predict how well it will work on new data. A confusion matrix is a technique of visualizing the relationship between the current outcomes and the predicted ones. It is used to evaluate the prediction accuracy of an algorithm or classifier. We have defined the performance metrics definition below:

1. True Positive (TP): Attack is correctly classified as an attack.

2. False Positive (FP): Normal is incorrectly classified as an attack.

3. True Negative (TN): Normal is correctly classified as normal

4. False Negative (FN): Attack is incorrectly classified as normal.

The machine learning algorithm's performance is evaluated based on the below-mentioned metrics.

**Recall:** It is about the number of relevant instances detected. It addresses the proportion of positive prediction to the total number of positive predictions. In other words, it decides how much valuable data is available from any machine learning method. The recall formula is defined as:

Recall = T P /T P + F N

**Precision:** It is defined as the number of relevant instances among the detected instances. using given formula it can be calculated:

Precision = T F /T P + F P

**Accuracy:** Accuracy is one of the performance evaluation of any algorithm. It is the ratio of the total number of true positive and true negative that are correctly detected and divided by the total amount of the dataset positive and negative amount.

using given formula it can be calculated:

Accuracy = T P + T N /T P + T N + F P + F N

In this experimental setup, Random forest algorithm gives the best accuracy model with accuracy as 95.74. Hence we detect the attack and its type using Random forest algorithm.

| Classifier | Accuracy |
|---|---|
| SVM | 79.48243992606284 |
| Random forest | 95.74861367837339 |
| Voting | 81.88539741219964 |
| AdaBoost | 70.97966728280961 |

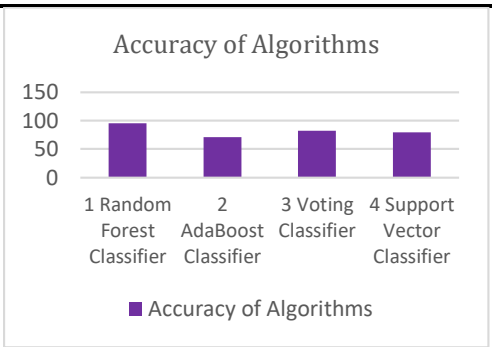


Fig 4.1 Comparison of Accuracy in Algorithm

**OUTPUT**

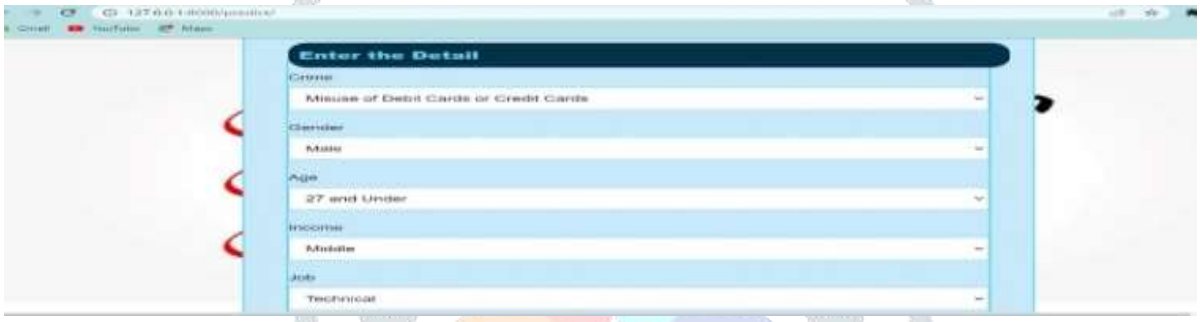


Fig 4.2 output


Fig 4.3 output


Fig4.4 output

## 5. CONCLUSION

THE PERFORMANCE DIFFERENCES BETWEEN EACH ALGORITHM ARE NOT SIGNIFICANT IN THE DETECTION EFFECT OF A SINGLE CLASSIFICATION TECHNIQUE. WHATEVER LEARNING ALGORITHM IS CHOSEN, A NUMBER OF TECHNIQUES CAN BE APPLIED TO BOOST THE EFFECTIVENESS OF DETECTION. OUR MODEL'S MAIN CONCEPT IS TO USE MACHINE LEARNING TO COMPILE THE BENEFITS OF SEVERAL TECHNIQUES. TO ENHANCE THE DETECTING EFFECT, WE EMPLOY MACHINE LEARNING. IT HAS BEEN DEMONSTRATED THAT, WHEN COMPARED TO OTHER RESEARCH ARTICLES, OUR METHODOLOGY SIGNIFICANTLY INCREASES DETECTION ACCURACY. THE ALGORITHM'S PRECISION IS THE ALGORITHM'S EFFECT IS UNQUESTIONABLY ENHANCED WHEN COMPARED TO OTHER ALGORITHMS OF A SIMILAR TYPE, AND IT HAS SIGNIFICANT PRACTICAL USEFULNESS.

## REFERENCES


[1] Ghada Abdelmoumin, Danda B. Rawat , Abdul Rahman, "On the Performance of Machine Learning Models for Anomaly-Based Intelligent Intrusion Detection Systems for the Internet of Things",IEEE VOL.9,NO.6,**March 15,2022.**

[2] Darshana Upadhyay , Jaume Manero , Marzia Zaman , and Srinivas Sampalli "Intrusion Detection in SCADA Based Power Grids:Recursive Feature Elimination Model With Majority Vote Ensemble Algorithm ieee transactions on network science and engineering, IEEE vol. 8, no. 3, **july-september 2021**

[3] Sugandh Seth,Kuljit Kaur Chalhal, Gurvinder Singh," A Novel Ensemble Framework For An Intelligent Intrusion Detection System", DOI 10.1109?ACCESS.2021.3116219,IEEE **2021**

[4] Chuang Ma, JIajun Zhang, Li Wang, Haisheng You "Network Attack Detection Based on Kernal Principle Component Analysis and Decision Tree", 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery(CyberC)|978-1-7281-8448-7/20/IEEE **2020**.

[5] Gurdip Kaur, Arash Habibi Lashkari, Abir Rahali,"Intrusion Traffic Detection and Characterization using Deep Image Learning", **2020** IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress.

[6] Timothy Chadza, Konstantinos G. Kyriakopoulos, Sangarapillai Lambotharan,"Learning to Learn Sequential Network Attacks Using Hidden Markov Models",Digital Object Identifier 10.1109/ACCESS.2020.3011293, IEEE **2020**

[7] Y. Zhang, J. Niu, G. He, L. Zhu, and D. Guo, "Network Intrusion Detection Based on Active Semi-supervised Learning," in Proc. IEEE/IFIP Int. Conf. Dep. Syst. Netw. Workshops, **2021**, pp. 129–135.

[8] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," T. Emerg. Telecomm. Techn.,IEEE vol. 32, no. 1, p. e4150, 2021.

[9] D. Upadhyay et al., "Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids," IEEE Trans. Netw. Service Manag., vol. 18, no. 1, pp. 1104–1116, **Mar. 2021.**

[10] Giovanni Apruzzese, Luca Pajola, Mauro Conti,"The Cross-evaluation of Machine Learning based Network Intrusion Detection Systems", DOI 10.1109/TNSM.2022.3157344, IEEE **2022.**

[11] M. Keshk, E. Sitnikova, N. Moustafa, J. Hu, and I. Khalil, "An integrated framework for privacy-preserving based anomaly detection for cyber-physical systems," IEEE Trans. Sustain. Comput., vol. 6, no. 1, pp. 66–79, **Jan.–Mar. 2021.**

[12] M. Bozdal, M. Samie, and I. K. Jennions, "WINDS: A wavelet-based intrusion detection system for Controller Area Network (CAN)," IEEE Access, vol. 9, pp. 58621–58633, **2021..**

[13] Nedhal Ahmad Hamdi Qaiwmchi, Haleh Amintoosi , Amirhossein Mohajerzadeh, "Intrusion Detection System Based on Gradient Corrected Online Sequential Extreme Learning Machine ",Digital Object Identifier 10.1109/ACCESS.2020.3047933, IEEE **2021.**

[14] Smitha Rajagopal , Poornima Panduranga Kundapur, Hareesha K. S.,"Towards Effective Network Intrusion Detection: From Concept to Creation on Azure Cloud",Digital Object Identifier 10.1109/ACCESS.2021.3054688, IEEE **2021.**

[15] Ayesha Siddiqua Dina, A. B. Siddique, D. Manivannan, " Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks",Digital Object Identifier 10.1109/ACCESS.2022.3205337, IEEE **2021.**

[16] Maheswari S , Arunesh K,"Unsupervised Binary BAT algorithm based Network Intrusion Detection System using enhanced multiple classifiers",Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC **2020**) IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

[17] S. Ghosh and S. Sampalli, "A survey of security in scada networks: Current issues and future challenges," IEEE Access, vol. 7, pp. 135812–135831, **2019.**

[18] Kai Zhang, Fei Zhao, Shoushan Luo, Yang Xin, Hongliang Zhu,"An Intrusion Action-Based IDS Alert Correlation Analysis and Prediction Framework",Digital Object Identifier 10.1109/ACCESS.2019.2946261, IEEE **2019.**

[19] Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, A. AlNemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," IEEE Access, vol. 7, pp. 41 525–41 550, **2019.**

[20] Xianwei Gao , Chun Shan , Changzhen Hu, Zequn Niu , And Zhen Liu,"An Adaptive Ensemble Machine Learning Model for Intrusion Detection",Digital Object Identifier 10.1109/ACCESS.**2019**.2923640.