Network Intrusion Prediction using Machine Learning

Mrs. K. Nithiya[1] and Sathya Jayasri. S[2], Vijithra. P[3]
Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3]
Anjalai Ammal Mahalingam Engineering College, Thiruvarur, India
nithyak045@gmail.com[1] and sivasiva131094@gmail.com[2], vijithrapandiyan@gmail.com[3]

Abstract: *Intrusion Detection Systems are designed to safeguard thesecurity needs of enterprise networks against cyber-attacks. However, networks suffer from several limitations, such as generating a high volume of low-quality alerts. While there area variety of intrusion detection solutions available, the prediction of network intrusion events is still under active investigation. Over the past, statistical methods have dominated the design of attack prediction methods. The analysis of dataset by Supervised Machine Learning Technique(SMLT) to capture several information's like, variable identification, univariate analysis, bivariate andmultivariate analysis, missing value treatments etc. Then, we going to implement a machine-learning algorithm to get betteraccuracy. After that, we can build looks like a web page by using Django Framework. Once the user passed the information then the Admin can see the output, and after he will send the response to the user. The result shows the effectiveness of the machine learning metrics which are accuracy, precision,Recall, F1 Score, Sensitivity, and Specificity.*

Keywords: Network Intrusion, Machine Learning algorithms, SVM, RF, Adaboost

I. INTRODUCTION

An Intrusion Detection machine (IDS) is a device that monitors community visitors for suspicious hobby and problem signals while such interest is discovered. it's miles a software program application that scans a community or a device for harmful hobby or coverage breaching. Any malicious undertaking or violation is generally pronounced either to an administrator or amassed centrally through the usage of a protection records and event management (SIEM) system. A SIEM gadget integrates outputs from more than one source and makes use of alarm filtering techniques to differentiate malicious interest from false alarms.

Even though intrusion detection structures reveal networks for potentially malicious activity, they may be additionally disposed to fake alarms. subsequently, businesses want to nice-track their IDS merchandise when they first set up them. It means nicely setting up the intrusion detection structures to understand what normal visitors on the network look like as compared to the malicious hobby.

Intrusion prevention systems additionally display community packets inbound to the gadget to check the malicious sports concerned in it and without delay send the caution notifications.

Data science is an interdisciplinary subject that uses scientific strategies, procedures, algorithms, and systems to extract knowledge and insights from dependent and unstructured records, and apply knowledge and actionable insights from statistics throughout a broad variety of software domain names.

Artificial intelligence (AI) is intelligence tested by using machines, in preference to the natural intelligence displayed using human beings or animals.

natural language processing (NLP) permits machines to study and understand human language. A sufficiently powerful natural language processing machine could permit herbal-language user interfaces and the purchase of know-how without delay from human-written sources, together with newswire texts.

device getting to know is to are expecting the future from past data. Machine Learning (ML) is a type of Artificial intelligence (AI) that provides computer systems with the ability to analyze without being explicitly programmed.

DOI: 10.48175/IJARSCT-9513

Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python.

Supervised system gaining knowledge is the majority of sensible device mastering makes use of supervised mastering. Supervised getting to know is wherein input variables (X) and an output variable (y) and use a set of rules to research the mapping characteristic from the input to the output is y = f(X). The goal is to approximate the mapping feature so well that if you have new enter records (X) you can predict the output variables (y) for that data.

II. RELATED WORK

[1] This system predicts attacks depending on the type of attacks that they are experiencing. Here the system takes a minimum of 1 and maximum of 5 types of attacks that the user has and tests those using 4 different types of supervised algorithms known as the Random forest, Ada booster, voting classifier, and The raw data collection is taken from an excel sheet (a CSV file). The CSV file contains a list of different types of attacks which they might cause. The vast data of the CSV file is then used for further testing and different analyzing purposes. The CSV file which was read by the panda module of python will now be analyzed by the NumPy.

[2] The data will be then used by the different algorithms that the system has used Random forest Algorithm, Ada boost, and voting classifier algorithm. This paper's main goal is to predict the types of attacks. In this paper, ML models are built using various pre-processing techniques to balance the unbalanced data and predicted using the algorithm. For implementation Anaconda 4. Anaconda has various ML libraries in python and R programming languages. Jupiter notebook in Anaconda was used to run the models since data is set in imbalanced various over-sampling and under-sampling techniques are used to balance the data. In sampling samples from minority classes are duplicated and in under-sampling, samples from major classes are detected. Random forest is a supervised ML model which is used for classification and regression problems.

[3] It consists of many decision trees on various parts of the given data set and prediction is done by taking the average of the prediction from each decision tree. The larger the number of trees RF leads to greater performance RF (Random Forest) Intrusion detection is the process of identifying the attacks on any sample towards the number of attack classes. Any attack sample would contain several features and the process has to measure the similarity of sample features with classified features of different attack classes. According to the similarity value measured, this method would perform attack prediction. In elementary cases, attack prediction is performed by suiting or matching each feature for the availability of each attack belonging to a specific attack.

[4] The challenge in this system is that most attacks share a common website. For example, if the user uses a purchasing website and makes an online payment based on that attack will happen. The third party may collect the online payment credentials from the user and access the card details and the attack will occur. So, while measuring the similarity among features, it is necessary to consider the maximum parameter. Also, the performance of the attack prediction is depending on the number of samples available in the training class. The purpose of this model is to determine the risk of having network attacks based on the user usage of the website. Machine learning is used for the same. Firstly, data clearing is done to transform raw data into a useful data set. Followed by an analysis of data to calculate the significance of each feature. In this process, the features are identified and converted into ML acceptance forms.

[5] The above processes are performed for every model to predict attacks related to security and websites. The dataset is uncleansed and since it has missing values, it can't be used directly. Thus, features are age, website, country, location, education, qualification, employment, etc... The missing and null value dataset is not used. The features used in the models are Age, Gender, Income, Job, and Attack Method. various Machine Learning models were applied after the dataset models are cleansed and analyzed. All the dataset models are worked on by using the logistic regression model. Thereafter the prediction is made based on the logistic regression models. This helps in improved training of the model. In this paper various machine learning algorithms are used like logistic regression. Random forest, AdaBoost, voter classifier. This system consists of a network attack dataset.

[6] The dataset is explored in the python environment along with a data dictionary of the attribute involved the prediction accuracy of our proposed method reaches 87.1% in intrusion detection. The future scope and improvement of

the project involved, the project is deployed in the cloud and optimizing the work to implement in the IOT system. The use of pipeline structure for data pre-processing cloud further helps in achieving improved results.

III. WORKING

Download and deploy anaconda and get the most beneficial bundle for devices getting to know Python.
Load a dataset and understand its structure using statistical summaries and information visualization.
machine studying models, pick the great and construct self-assurance that the accuracy is dependable.
Python is a popular and effective interpreted language. unlike R, Python is an entire language and platform you could use for research and development and manufacturing structures. There are also several modules and libraries to choose from, presenting a couple of approaches to each challenge. it can sense overwhelming.
The satisfactory way to get commenced the usage of Python for Machine learning is to finish a mission.

- It will force you to install and start the Python interpreter (at the very least).
- It will come up with a fowl's eye view of a way to step via a small undertaking.
- It will give you confidence, maybe to go on to your small projects.

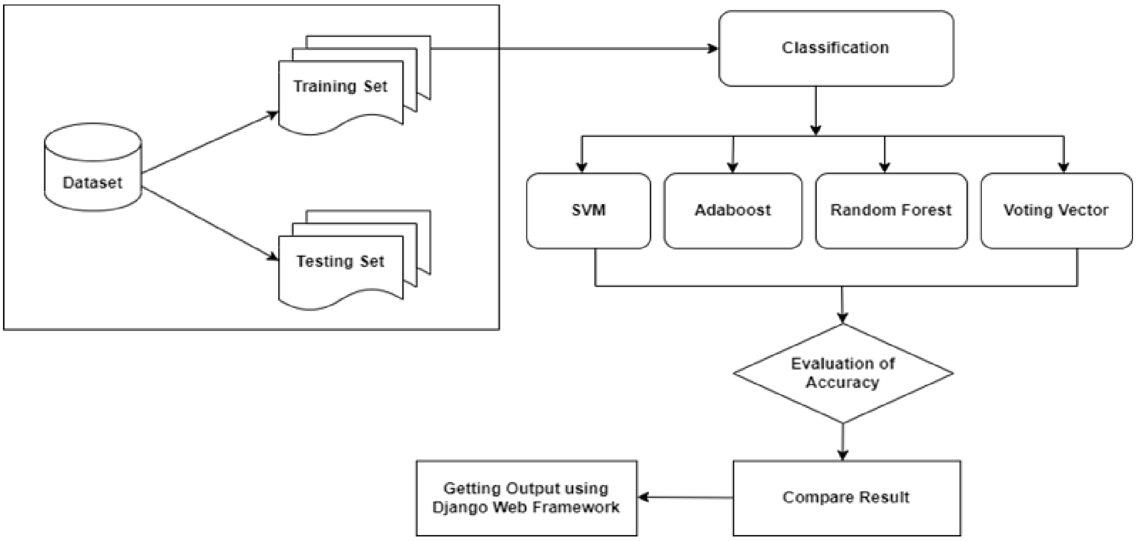
While you are making use of gadgets to gain knowledge of your very own datasets, you're running on a project. A system studying undertaking may not be linear, but it has some steps:

- Define the Problem.
- Prepare Data.
- Evaluate Algorithms.
- Improve Results.
- Present Results.

The first-class way to virtually come to phrases with a brand-new platform or tool is to work via a system get to know the venture give it up-to-end and cover the necessary thing steps. namely, from loading facts, summarizing statistics, evaluating algorithms, and making some predictions.
right here is an overview of what we are going to cover:

1. Installing the Python anaconda platform.
2. Loading the dataset.
3. Summarize the dataset.
4. Visualizing the dataset.
5. Evaluating some algorithms.
6. Making some predictions.

IV. SYSTEM MODEL



DOI: 10.48175/IJARSCT-9513

We will select the algorithm with the high accuracy rate for each attack after applying our knowledge. After that, we constructed a pickle file for each Attack and merged it with the Django framework for the model output displayed on a webpage.

V. METHODOLOGY

Different machine learning algorithms are used in this Intrusion Detection system. The methodology used is the SVM algorithm, AdaBoost Classifier, Random Forest classifier, and voting Classifier. The best accuracy-yielding algorithm is used for reciting the specific attack.

5.1 Support Vector Machine

Step 1: Support Vector Machine (SVM) is a supervised machine getting-to-know algorithm used for both category and regression. even though we say regression issues as nicely it's great proper for class.

Step 2: The objective of the SVM set of rules is to discover a hyperplane in an N-dimensional area that surprisingly classifies the records factors.

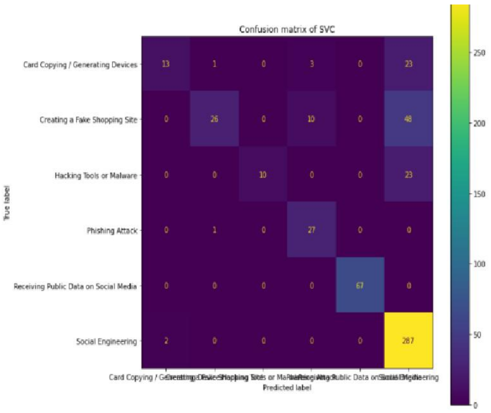
Step 3: SVM works by mapping statistics to a high-dimensional function space so that statistics factors can be categorized, even if the facts aren't in any other case linearly separable.

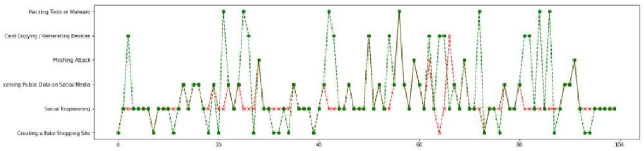
Step 4: A separator between the kinds is found, then the statistics are converted in this type manner so that the separator can be drawn as a hyperplane.

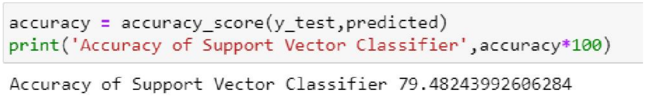
Step 5: SVMs are used in programs like handwriting recognition, intrusion detection, face detection, electronic mail classification, gene classification, and in web pages.

Step 6: This is one of the motives we use SVMs in device learning. it may take care of both classification and regression on linear and non-linear facts.

s= SVC ( )







```
accuracy = accuracy_score(y_test,predicted)
print('Accuracy of Support Vector Classifier',accuracy*100)
```

```
Accuracy of Support Vector Classifier 79.48243992606284
```

5.2 AdaBoost Classifier

DOI: 10.48175/IJARSCT-9513

Step 1: An AdaBoost classifier is a meta-estimator that starts evolving with the aid of fitting a classifier at the unique dataset after which fits extra copies of the classifier at the identical dataset however wherein the weights of incorrectly categorized times are adjusted such that the next classifiers recognize greater on difficult cases.

Step 2: AdaBoost can be used to reinforce the overall performance of any device mastering algorithm. it's far more pleasant used with weak novices. those are models that attain accuracy simply above random threat on a classification problem.

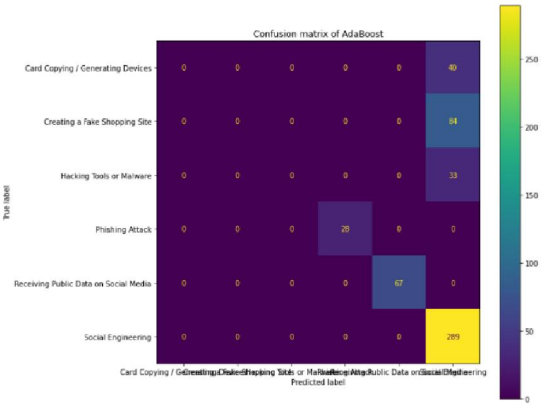
Step 3: The most proper and therefore maximum commonplace set of rules used with AdaBoost is decision bushes with one degree. How does the AdaBoost set of rules paintings explain?

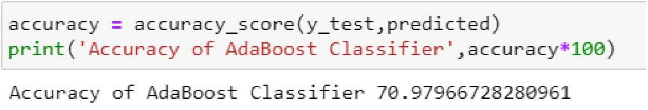
Step 4: It works on the precept of inexperienced persons developing sequentially. except for the primary, every next learner is grown from previously grown novices.

Step 5: In simple words, susceptible newbies are converted into robust ones. The AdaBoost set of rules works on the equal principle of boosting with a moderate distinction.

ADC = AdaBoost(classifier()

ADC. fit(X_train,y_train)





```
accuracy = accuracy_score(y_test,predicted)
print('Accuracy of AdaBoost Classifier',accuracy*100)
```

```
Accuracy of AdaBoost Classifier 70.97966728280961
```

5.3 Random Forest

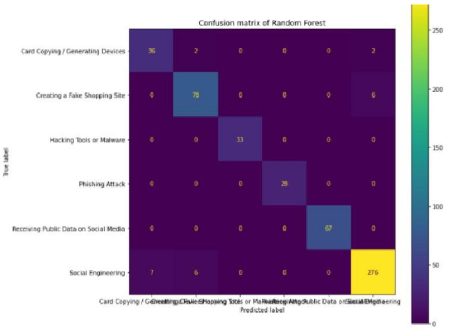
Step 1: Random Forest is a famous device studying algorithm that belongs to the supervised gaining knowledge of approach.

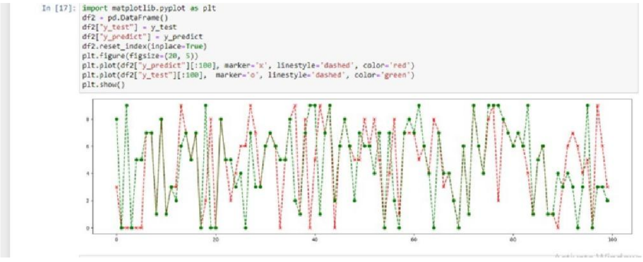
Step 2: It may be used for each class and Regression problems in ML. it's far primarily based on the concept of ensemble studying, that's a technique of mixing a couple of classifiers to resolve complicated trouble and to improve the overall performance of the version.

Step 3: As the call indicates, "Random Forest is a classifier that consists of some of the decision timber on diverse subsets of the given dataset and takes the commonplace to enhance the predictive accuracy of that dataset." instead of relying on one choice tree, the random woodland takes the prediction from every tree and based totally on the majority votes of predictions, and it predicts the final output.

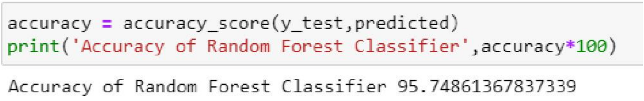
Step 4: The more range of trees in the wooded area results in higher accuracy and stops the trouble of overfitting.

DOI: 10.48175/IJARSCT-9513





```
In [17]: import matplotlib.pyplot as plt
         df2 = pd.DataFrame()
         df2["y_test"] = y_test
         df2["y_predict"] = y_predict
         df2.reset_index(inplace=True)
         plt.figure(figsize=(20, 5))
         plt.plot(df2["y_predict"][:100], marker='x', linestyle='dashed', color='red')
         plt.plot(df2["y_test"][:100],  marker='o', linestyle='dashed', color='green')
         plt.show()
```



```
accuracy = accuracy_score(y_test,predicted)
print('Accuracy of Random Forest Classifier',accuracy*100)
```

```
Accuracy of Random Forest Classifier 95.74861367837339
```

5.4 Voting Classifier

Step 1: A voting Classifier is a device learning model that trains on an ensemble of several models and predicts an output (elegance) based totally on their maximum opportunity of selecting elegance as the output.

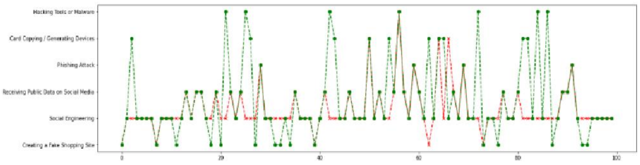
Step 2: It truly aggregates the findings of every classifier surpassed into vote casting Classifier and predicts the output elegance based totally on the best majority of balloting.

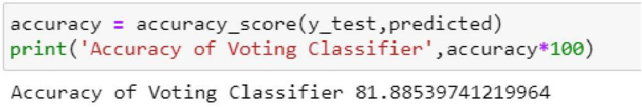
Step 3: The idea is rather than creating separate devoted fashions and finding the accuracy for every them, we create a single version that trains with the aid of those fashions and predicts output based totally on their mixed majority of balloting for every output class.

Step 4: vote casting Classifier helps two types of voting.

1. Hard voting: In hard voting, the anticipated output magnificence is a class with the very best majority of votes i.e., elegance which had the best chance of being expected with the aid of each of the classifiers. think 3 classifiers expected the output class (A, A, B), so right here the general public expected A as output. for this reason, A might be the final prediction.

2. Soft Voting: In soft voting, the output class is the prediction primarily based on the common probability given to that elegance. think given some input to a few models, the prediction opportunity for class A = (0.30, zero. forty-seven, zero. fifty-three) and B = (0.20, 0.32, zero. forty). So, the average for sophistication A is 0.4333 and B is 0.3067, the winner is truly magnificent A as it had the highest possibility averaged by using each classifier.





```
accuracy = accuracy_score(y_test,predicted)
print('Accuracy of Voting Classifier',accuracy*100)
```

```
Accuracy of Voting Classifier 81.88539741219964
```

DOI: 10.48175/IJARSCT-9513

VI. CONCLUSION

Intrusion Detection Systems are designed to safeguard the security needs of enterprise networks against cyber-attacks. However, networks suffer from several limitations, such as generating a high volume of low-quality alerts. The study has reviewed the state-of-the-art cyber-attack prediction based on Intrusion Alert, its models, and limitations. The ever-increasing frequency and intensity of intrusion attacks on computer networks worldwide have intense research efforts toward the design of attack detection and prediction mechanisms. While there are a variety of intrusion detection solutions available, the prediction of network intrusion events is still below active research. over the past, statistical strategies have dominated the layout of assault prediction strategies.

The analysis of the dataset by supervised machine learning technique (SMLT) to capture several information like variable identification, univariate analysis, bivariate and multivariate analysis, missing value treatments, etc. A comparative examination between machine studying algorithms was accomplished on the way to determine which set of rules is the most accurate in predicting the type of cyber assaults. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with the best accuracy, precision, Recall, F1 Score, Sensitivity, and Specificity. The analytical system commenced with data cleansing and processing, missing value, exploratory evaluation, and ultimately model building and evaluation The best accuracy on a public test set of higher accuracy score algorithms will find out. The founded one is used in the application which can help to find the type of intrusions.

REFERENCES

[1]. E. D. Alalade, "Intrusion Detection System in Smart Home Network Using Artificial Immune System and Extreme Learning Machine Hybrid Approach," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, 2020, pp. 1-2, doi: 10.1109/WF-IoT48130.2020.9221151.

[2]. J. A. Abraham and V. R. Bindu, "Intrusion Detection and Prevention in Networks Using Machine Learning and Deep Learning Approaches: A Review," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-4, doi: 10.1109/ICAECA52838.2021.9675595.

[3]. G. Abdelmoumin, D. B. Rawat and A. Rahman, "On the Performance of Machine Learning Models for Anomaly-Based Intelligent Intrusion Detection Systems for the Internet of Things," in IEEE Internet of Things Journal, vol. 9, no. 6, pp. 4280-4290, 15 March15, 2022, doi: 10.1109/JIOT.2021.3103829.

[4]. N. Tran, H. Chen, J. Bhuyan and J. Ding, "Data Curation and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection," in IEEE Access, vol. 10, pp. 121900-121923, 2022, doi: 10.1109/ACCESS.2022.3211313.

[5]. M. A. Siddiqi and W. Pak, "Tier-Based Optimization for Synthesized Network Intrusion Detection System," in IEEE Access, vol. 10, pp. 108530-108544, 2022, doi: 10.1109/ACCESS.2022.3213937.

[6]. G. Pu, L. Wang, J. Shen and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," in Tsinghua Science and Technology, vol. 26, no. 2, pp. 146-153, April 2021, doi: 10.26599/TST.2019.9010051.

[7]. W. Wang, X. Du, D. Shan, R. Qin and N. Wang, "Cloud Intrusion Detection Method Based on Stacked Contractive Auto-Encoder and Support Vector Machine," in IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 1634-1646, 1 July-Sept. 2022, doi: 10.1109/TCC.2020.3001017.

[8]. J. Lansky et al., "Deep Learning-Based Intrusion Detection Systems: A Systematic Review," in IEEE Access, vol. 9, pp. 101574-101599, 2021, doi: 10.1109/ACCESS.2021.3097247.

[9]. R. Conde Camillo da Silva, M. P. Oliveira Camargo, M. Sanches Quessada, A. Claiton Lopes, J. Diassala Monteiro Ernesto and K. A. Pontara da Costa, "An Intrusion Detection System for Web-Based Attacks Using IBM Watson," in IEEE Latin America Transactions, vol. 20, no. 2, pp. 191-197, Feb. 2022, doi: 10.1109/TLA.2022.9661457.

[10]. O. Alkadi, N. Moustafa, B. Turnbull and K. -K. R. Choo, "A Deep Blockchain Framework-Enabled Collaborative Intrusion Detection for Protecting IoT and Cloud Networks," in IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9463-9472, 15 June15, 2021, doi: 10.1109/JIOT.2020.2996590.

[11]. P. Barnard, N. Marchetti and L. A. DaSilva, "Robust Network Intrusion Detection Through Explainable Artificial Intelligence (XAI)," in IEEE Networking Letters, vol. 4, no. 3, pp. 167-171, Sept. 2022, doi: 10.1109/LNET.2022.3186589.

[12]. J. Gao et al., "Omni SCADA Intrusion Detection Using Deep Learning Algorithms," in IEEE Internet of Things Journal, vol. 8, no. 2, pp. 951-961, 15 Jan.15, 2021, doi: 10.1109/JIOT.2020.3009180.