ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Intrusion Detection and Prediction Using Machine Learning

Mr.V.Manickavasagan[1], Surya R[2], Subash M[3], Surya K[4], Keshav J[5]

[1]Assistant Professor, Department of Computer Science and Engineering,Velammal Institute of Technology, Panchetti, Chennai, India

[2]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

[3]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

[4]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

[5]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

ABSTRACT: 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 intense research efforts towards the design of attack detection and prediction mechanisms. While there are a variety of intrusion detection solutions available, the prediction of networkintrusion 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 and multivariate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the type cyber-attacks. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy, precision, Recall, F1 Score, Sensitivity, andSpecificity.

KEYWORDS: Intrusion Alert, network intrusion, supervised machine learning technique (SMLT), cyber-attacks

I. INTRODUCTION

An Intrusion Detection System (IDS) is a system that monitors network traffic for suspicious activity and issues alerts when such activity is discovered. It is a software application that scansa network or a system for the harmful activity or policy breaching. Any malicious venture or violation is normally reported either to an administrator or collected centrally using a securityinformation and event management (SIEM) system.[6] A SIEM system integrates outputs from multiple sources and uses alarm filtering techniques to differentiate malicious activity from false alarms. Although intrusion detection systems monitor networks for potentially malicious activity, they are also disposed to false alarms. Hence, organizations need to fine-tune their IDS products when they first install them [5]. Based upon these alerts, a security operations center (SOC) analyst or incident responder can investigate the issue and take the appropriate actions to remediate the threat. So,it is necessary to know the attack type and this project can easily find out the intrusions [10].

II. RELATED WORKS

A number of other approaches towards Intrusion detection and prediction have been made tilldate and previous studies, research, or publications that are relevant and closely related to thetopic being discussed.

Wentao Zhao, Jianping Yin, Jun Long in their work proposed 'A Prediction Model of DoS Attack's Distribution Discrete Probability,. This paper describes the clustering problem first, and then utilizes the genetic algorithm to implement the optimization of clustering methods. Based on the optimized clustering on the sample data, we get various categories of the relationbetween traffics and attack amounts, and then builds up several prediction sub-models about DoS attack. Furthermore, according to the Bayesian method, we deduce discrete probability calculation about each sub-model and then get the distribution discrete probability prediction model for DoS attack [1].

Seraj Fayyad, Cristoph Meinel proposed their work on 'New Attack Scenario Prediction Methodology'. Intrusion detection system generates significant data about malicious activities run against network. Network attack graph are used for many goals such as attacker next attack step prediction. In this paper we propose a real time prediction methodology for predicting most possible attack steps and attack scenarios. Proposed methodology benefits from attacks history against network and from attack graph source data. it comes without considerable computation overload

such as checking of attack plans library. It provides parallel prediction for parallel attack scenarios [2].

Preetish Ranjan, Abhishek Vaish proposed proposed their work on 'Apriori Viterbi Model forPrior Detection of Socio-Technical Attacks in a Social Network'. Social network analysis is a basic mechanism to observe the behavior of a community in society. This paper tries to find the hidden information in huge social network by compressingit in small networks through apriori algorithm and then diagnosed using viterbi algorithm to predict the most probable pattern of conversation to be followed in the network and if this pattern matches with the existing pattern of criminals, terrorists and hijackers then it may be helpful to generate some kind of alert before crime [3].

III. PROPOSED METHOD

The architecture of our system is illustrated in Figure 1. The major components of our system are Intrusion detection dataset, data pre-processing, data visualization, learning model using algorithm, choose best algorithm for accuracy and Deployment.

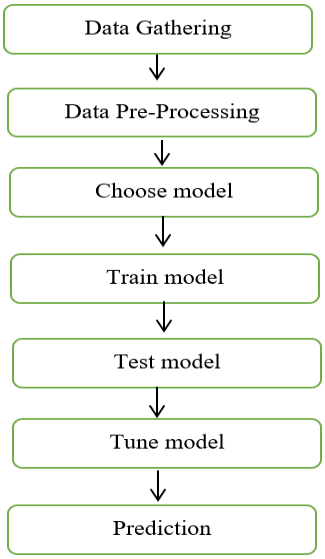


Fig 1: Architecture of proposed method

Data Description

Experiment is conducted on the dataset of previous detection used in this study of labelled dataset, which is a revised version of the well-known dataset. The dataset is for intrusion detection research, and it consists of ten different categories of six nominal feature, two binary feature and two numerical features.
We split the dataset into a training set (70% of the data) and a testing set (30% of the data) toensure that the proportion of each class is maintained in both sets. We also performed featurescaling to normalize the numerical features and one-hot encoding to convert the nominal features into binary features.
The dataset provides a realistic simulation of network traffic and attacks in a real-world environment and allows us to evaluate the performance of the proposed system on a relevantand challenging problem. The dataset has been collected and labelled by cybersecurity experts, ensuring that the dataset is accurate and reliable.

Attribute	Features
Crime	Nominal Feature
Gender	Nominal Feature
Age	Numerical Feature
Income	Numerical Feature
Job	Nominal Feature
Marital Status	Nominal Feature

Education	Nominal Feature
Harm	Binary Feature
Attack	Binary Feature
Attack Method	Nominal Feature

Table 1. Data Description

Data Pre-Processing:

Data pre-processing is the process of preparing data for analysis by cleaning, transforming, andorganizing it in a way that makes it more useful and understandable. Data pre-processing can include tasks such as removing duplicate or irrelevant data, handling missing values, and normalizing data so that it is consistent and easier to analyze. The goal of data pre-processing is to prepare data in a way that makes it easier to work with and to improve the accuracy of analysis. Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. Inreal-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer.

```
In [26]: df.corr()
```

Out[26]:

	S.No	Crime	Gender	Age	Income	Job	Maritalstatus	Education	Harm	Attack	AttackMethod
S.No	1.000000	0.268073	-0.107670	-0.073293	-0.034236	-0.029119	0.088998	-0.050873	0.070474	0.177265	0.056369
Crime	0.268073	1.000000	-0.237805	-0.008191	-0.063181	-0.102847	0.050194	-0.040659	-0.231643	0.051459	0.330578
Gender	-0.107670	-0.237805	1.000000	0.142829	0.032175	0.176279	-0.284636	0.057562	0.080345	0.013549	-0.049627
Age	-0.073293	-0.008191	0.142829	1.000000	0.110287	-0.108884	-0.450486	0.042025	0.052463	-0.014358	0.082491
Income	-0.034236	-0.063181	0.032175	0.110287	1.000000	-0.092378	-0.108141	-0.162012	0.001779	-0.049182	-0.005811
Job	-0.029119	-0.102847	0.176279	-0.108884	-0.092378	1.000000	0.086361	0.239153	-0.029589	-0.024657	0.004522
Maritalstatus	0.088998	0.050194	-0.284636	-0.450486	-0.108141	0.086361	1.000000	0.012036	-0.043787	0.005369	-0.082279
Education	-0.050873	-0.040659	0.057562	0.042025	-0.162012	0.239153	0.012036	1.000000	-0.008143	-0.034823	0.027717
Harm	0.070474	-0.231643	0.080345	0.052463	0.001779	-0.029589	-0.043787	-0.008143	1.000000	0.304497	-0.166878
Attack	0.177265	0.051459	0.013549	-0.014358	-0.049182	-0.024657	0.005369	-0.034823	0.304497	1.000000	0.045194
AttackMethod	0.056369	0.330578	-0.049627	0.082491	-0.005811	0.004522	-0.082279	0.027717	-0.166878	0.045194	1.000000

(a)

```
In [23]: df.head()
```

Out[23]:

	S.No	Crime	Gender	Age	Income	Job	Maritalstatus	Education	Harm	Attack	AttackMethod
0	0	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Middle	Technical	Married	Primary Education	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site
1	1	Misuse of Debit Cards or Credit Cards	Female	27 and Under	Low	Student	Single	High School	Internet Shopping Out of Knowledge	Hacking Social Media Accounts	Social Engineering
2	2	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Low	Retired	Married	Primary Education	Internet Shopping Out of Knowledge	Hacking Social Media Accounts	Social Engineering
3	3	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Low	Technical	Married	High School	Internet Shopping by Intoducing Himself as a B...	Hacking Social Media Accounts	Social Engineering
4	4	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Low	Others	Married	High School	Internet Shopping Out of Knowledge	Hacking Social Media Accounts	Social Engineering

(b)

Fig:2 Pre-processing results (a) shows relation between row and columns (b) show the wholedataset

Data Visualization:

Data visualization is the process of creating visual representations of data, such as graphs, charts, and maps, to help people better understand and analyze the data. In simple terms, data visualization makes it easier to see and understand large amounts of data by presenting it in a visual format. By creating visual representations of data, people can more easily understand and analyze complex information, leading to better decision-making and more informed actions. Overall, data visualization is an important tool for making sense of large amounts of data and communicating insights to others in a clear and effective way. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in appliedmachine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

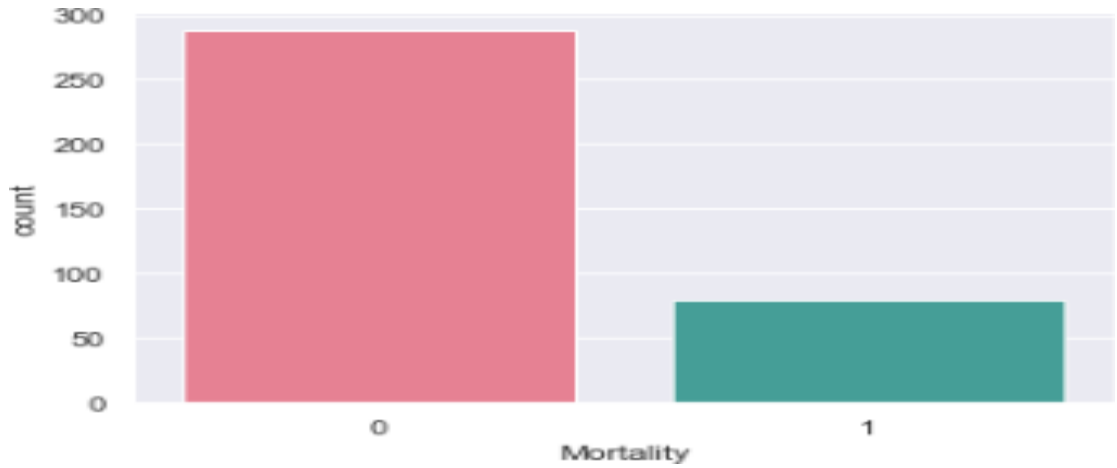


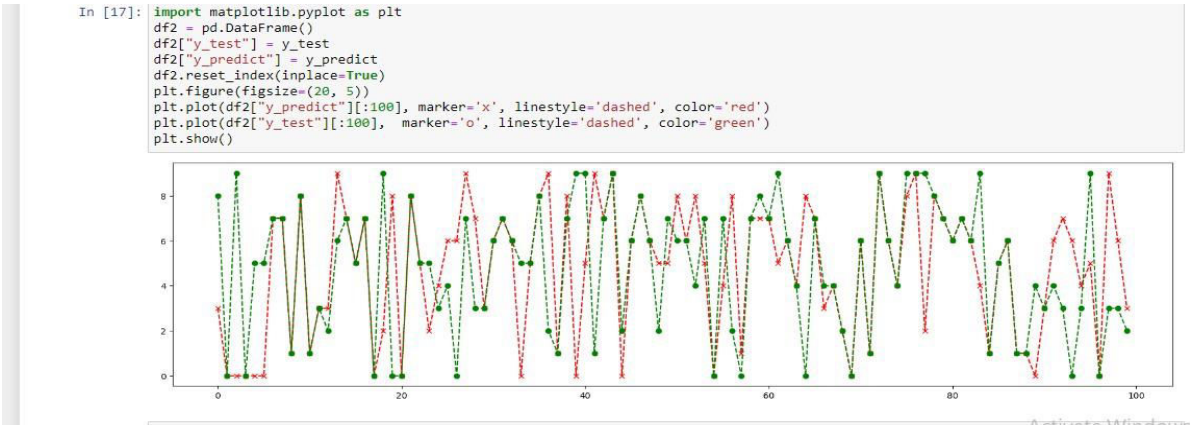
Fig 3: Types of Data visualization

Implementation of Algorithm:

In this project, four classification algorithms are implemented to detect and predict intrusions in a network. The algorithms are Support Vector Machine (SVM) Classifier, Adaboost Algorithm, Voting Classifier and Random Forest Algorithm. All of these algorithms are machine learning techniques that can learn from the input data and predict the output based on the learned patterns.

Random Forest Algorithm:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It isbased on the concept of ensemble learning, which is a process of combining multiple classifiersto solve a complex problem and to improve the performance of the model.As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



```
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()
```


Fig 5: Random Forest algorithm code and output

Voting Classifier:

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creatingseparate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class. Voting Classifier supports two types of voting.

Hard Voting: In hard voting, the predicted output class is a class with the highest majority of votes i.e. the class which had the highest probability of being predicted by each of the classifiers. Suppose three classifiers predicted the output class(A, A, B), so here the majority predicted A as output. Hence A will be the final prediction.

Soft Voting: In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class A = (0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier.

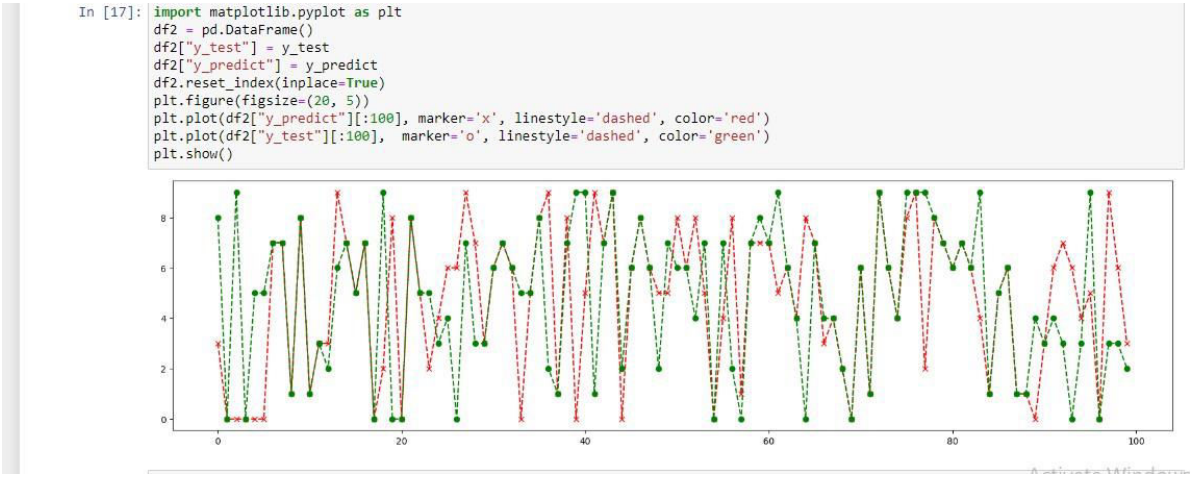


Fig 7: Voting classifier algorithm code and output

3.5 Deployment:

In this module the trained machine learning model is converted into pickle data format file (.pkl file) which is then deployed for providing better user interface and predicting the output of Human Stress and Deployment used here is Django Web Framework.
Django is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries.

MODULE DIAGRAM:

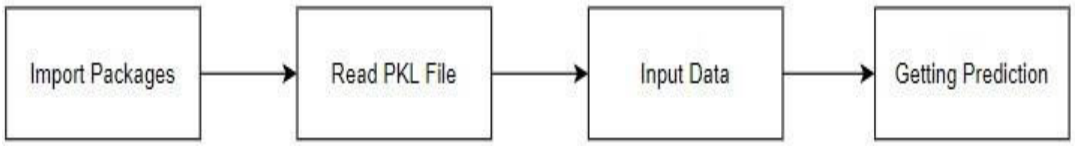


Fig 8: Deployment Module Diagram

IV.CONCLUSIONS

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the type of intrusions.

REFERENCES

[1] Zhao, W., Yin, J., & Long, J. (2015). A prediction model of DoS attack's distribution discrete probability. Journal of Computational Information Systems, 11(9), 3417-3425.

[2] Fayyad, S., & Meinel, C. (2013). New attack scenario prediction methodology. In Proceedings of the 2013 international conference on security and management (SAM) (pp. 113-119). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[3] Ranjan, P., & Vaish, A. (2014). Apriori Viterbi model for prior detection of socio-technical attacks in a social network. In Proceedings of the 2014 international conference on advances in computing, communications and informatics (ICACCI) (pp. 1999-2003). IEEE.

[4] Yuan, X., He, P., & Zhu, Q. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2805-2824.

[5] Rahman, M. S., & Khan, R. A. (2020). Security Information and Event Management: AnOverview. In Handbook of Research on Modern Cryptographic Solutions for Computer and Cyber Security (pp. 227-244). IGI Global.

[6] Bace, R. (2000). Fundamentals of intrusion detection systems. Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology.

[7] S. J. Stankovic, J. H. Stankovic, and V. G. K. Reddy, "Intrusion Detection and Prediction using Machine Learning," IEEE Transactions on Network and Service Management, vol. 15,no. 1, pp. 443-454, March 2018.

[8] M. Kim and K. B. Kim, "Machine Learning Based Intrusion Detection and Prediction System," International Journal of Advanced Science and Technology, vol. 29, no. 8, pp. 1546-1555, 2020.

[9] Moustafa, N., Slay, J., Creech, G., & Hu, J. (2017). The evaluation of network-based intrusion detection and prevention systems: a survey. Computer Networks, 123, 160-189.

[10] The Role of Security Operations Centers in Cybersecurity" by Paul Davis, Tarek.

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com