iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52167

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Intrusion Detection Prediction Using Data Science Technique

T. Sanjay[1], K. Rahulkaran[2], B. Naveen Kumar[3], K. Anitha[4]
Computer Science Engineering Department, Anna University, 12, Sardar Patel Road, Guindy, Chennai, Tamilnadu 600025

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 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 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, and Specificity.***
Index Terms: Machine learning, Instrusion Detection System, network security, Evaluation

I. INTRODUCTION

An intrusion detection system (IDS) is a device or software application that monitors a network for malicious activity or policy violations. Any malicious activity or violation is typically reported or collected centrally using a security information and event management system. The goal is to develop a machine learning model for intrusion detection Prediction, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.
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 scans a 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 security information and event management (SIEM) system. 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. It means properly setting up the intrusion detection systems to recognize what normal traffic on the network looks like as compared to malicious activity.
Intrusion prevention systems also monitor network packets inbound the system to check the malicious activities involved in it and at once send the warning

A. Natural Language Processing (NLP)

Natural language processing (NLP) allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of natural language processing include information retrieval, text mining, question answering and machine translation. Many current approaches use word co-occurrence frequencies to construct syntactic representations of text. "Keyword spotting" strategies for search are popular and scalable but dumb; a search query for "dog" might only match documents with the literal word "dog" and miss a document with the word "poodle". "Lexical affinity" strategies use the occurrence of words such as "accident" to assess the sentiment of a document. Modern statistical NLP approaches can combine all these strategies as well as others, and often achieve acceptable accuracy at the page or paragraph level.

©IJRASET: All Rights are Reserved

Beyond semantic NLP, the ultimate goal of "narrative" NLP is to embody a full understanding of common sense reasoning. By 2019, transformer-based deep learning architectures could generate coherent text.

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. 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. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labelling to learn data has to be labelled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points.

Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. . Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

Supervised Machine Learning **is the** majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output **is y = f(X).** The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include **logistic regression**, **multi-class classification**, **Decision Trees** and **support vector machines etc**. Supervised learning requires that the data used to train the algorithm is already labelled with correct answers.

Supervised learning problems can be further grouped into **Classification** problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification



Process Of Machine Learning

II. LITERATURE REVIEW

Title: A Prediction Model of DoS Attack’s Distribution Discrete Probability
Author: Wentao Zhao, Jianping Yin, Jun Long
Year: 2008

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 relation between traffics and attack amounts, and then builds up several prediction sub-models about DoS attack

Title: Adversarial Examples: Attacks and Defenses for Deep Learning
Author: Xiaoyong Yuan , Pan He, Qile Zhu
Year: 2019
With rapid progress and significant successes in a wide spectrum of applications, deep learning is being applied in many safety-critical environments. However, deep neural networks (DNNs) have been recently found vulnerable to well-designed input samples called adversarial examples. Adversarial perturbations are imperceptible to human but can easily fool DNNs in the testing/deploying stage. The vulnerability to adversarial examples becomes one of the major risks for applying DNNs in safety-critical environments. Therefore, attacks and defenses on adversarial examples draw great attention. In this paper, we review recent findings on adversarial examples for DNNs, summarize the methods forgenerating adversarial examples, and propose a taxonomy of these methods. Under the taxonomy, applications for adversarial examples are investigated. We further elaborate on countermeasures for adversarial examples.

Title: Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network
Author: Preetish Ranjan, Abhishek Vaish
Year: 2014
Social network analysis is a basic mechanism to observe the behavior of a community in society. In the huge and complex social network formed using cyberspace or telecommunication technology, the identification or prediction of any kind of socio-technical attack is always difficult. This challenge creates an opportunity to explore different methodologies, concepts and algorithms used to identify these kinds of community on the basis of certain pattern, properties, structure and trend in their linkage. This paper tries to find the hidden information in huge social network by compressing it 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.

Title: New Attack Scenario Prediction Methodology
Author: Seraj Fayyad, Cristoph Meinel
Year: 2013: Intrusion detection system generates significant data about malicious activities run against network. Generated data by IDS are stored in IDS database. This data represent attacks scenarios history against network. Main goal of IDS system is to enhance network defense technologies. Other techniques are also used to enhance the defense of network such as Attack graph. 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.

III.EXISTING SYSTEM

Enhancing Network Intrusion Detection Systems (NIDS) with supervised Machine Learning (ML) is tough. ML-NIDS must be trained and evaluated, operations requiring data where benign and malicious samples are clearly labelled. Such labels demand costly expert knowledge, resulting in a lack of real deployments, as well as on papers always relying on the same outdated data. The situation improved recently, as some efforts disclosed their labelled datasets. However, most past works used such datasets just as a ‘yet another’ test bed, overlooking the added potential provided by such availability. Despite many successes, the integration of supervised Machine Learning (ML) methods in Network Intrusion Detection Systems (NIDS) is still at an early stage. This is due to the difficulty in obtaining comprehensive sets of labelled data for training and evaluating an ML-NIDS. The recent release of labelled datasets for ML-NIDS was appreciated by the research community; however, few works noticed the opportunity that such availability provides to the state-of-the-art.

A. Disadvantages

1) They are not predicting the classification of attack types.
2) They are not mentioning the accuracy.
3) They are using machine learning technique only for analysing purpose.

IV. PROPOSED SYSTEM

The proposed model is to build a machine learning model for anomaly detection. Anomaly detection is an important technique for recognizing fraud activities, suspicious activities, network intrusion, and other abnormal events that may have great significance but are difficult to detect. The machine learning model is built by applying proper data science techniques like variable identification which is the dependent and independent variables. Then the pre-processing and visualisation of the data is done .The model is build based on the previous dataset where the algorithm learn data and get trained different algorithms are used for better comparisons. The performance metrics are calculated and compared.

A. Advantages

1) We are implementing the machine learning algorithm for classification purpose.
2) More than two machine learning algorithms are used for comparison of getting the best accuracy.
3) Deployment is done for getting result.

V. METHODOLOGY

The below 4 different algorithms are compared:

- SVM
- Random Forest
- Voting
- AdaBoost

A. SVM

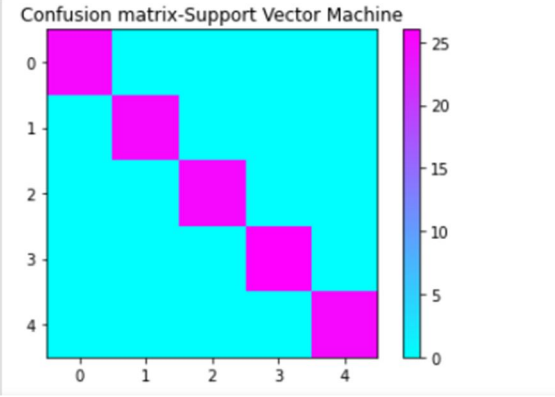
Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

```
s = SVC()

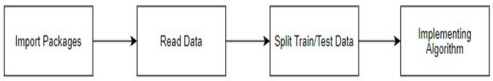
s.fit(X_train,y_train)

predictS = s.predict(X_test)
```

```
print("")
print("Accuracy Result of Support Vector Machine is:",accuracy.mean() * 100)
svc=accuracy.mean() * 100
```



MODULE DIAGRAM



Given Input Expected Output
Input : data
Output : getting accuracy

B. Adaboost Classifier:

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners .It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference.

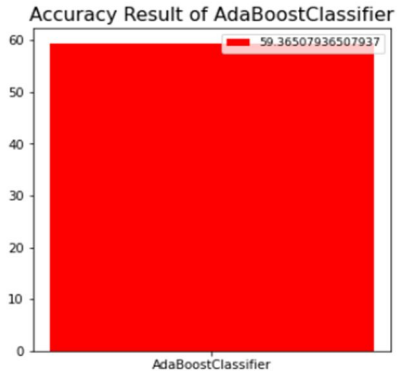
```
ADC = AdaBoostClassifier()

ADC.fit(X_train,y_train)

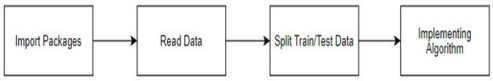
predictRF = ADC.predict(X_test)
```

```
accuracy = cross_val_score(ADC, x_ros, y_ros, scoring='accuracy')
print('Cross validation test results of accuracy:')
print(accuracy)


print("")
print("Accuracy Result of AdaBoostClassifier is:",accuracy.mean() * 100)
adc=accuracy.mean() * 100
```



Module Diagram

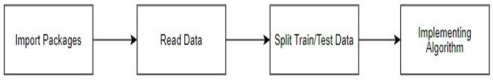


Given Input Expected Output
Input : data
Output : getting accuracy

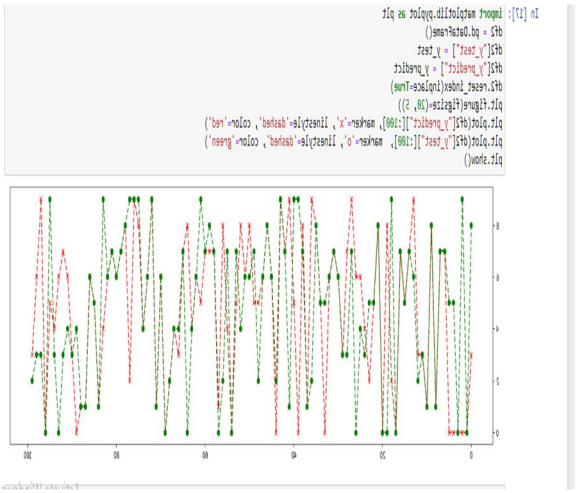
C. Random Forest Classifier

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 is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Module Diagram



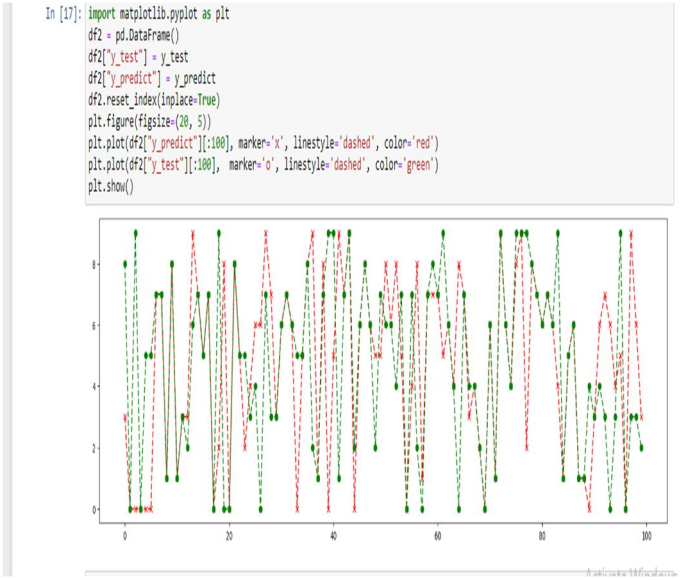
The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



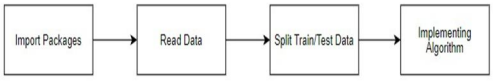
Given Input Expected Output
Input : data
Output : getting accuracy

D. 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 creating separate 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.

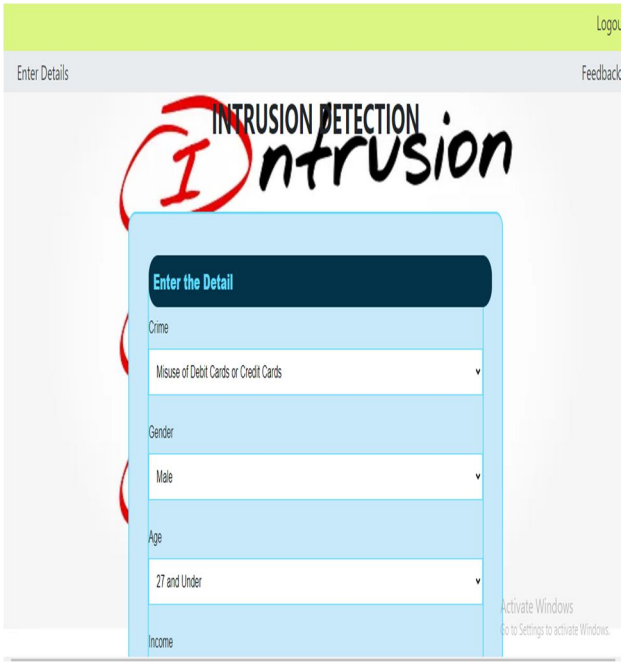


MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT
input : data
output : getting accuracy.

VI. RESULT



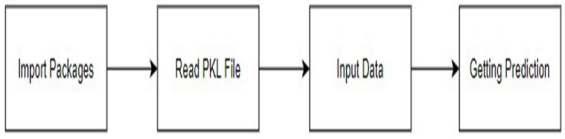
Thus using all four algorithm we can predict and prevent intruders. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy, precision, Recall ,Score ,Sensitivity, and Specificity.

VII. DEPLOYMENT

A. Deploying the model predicting output

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 heart attack.

Module Diagram



Given Input Expected Output
Input : data values
Output : predicting output

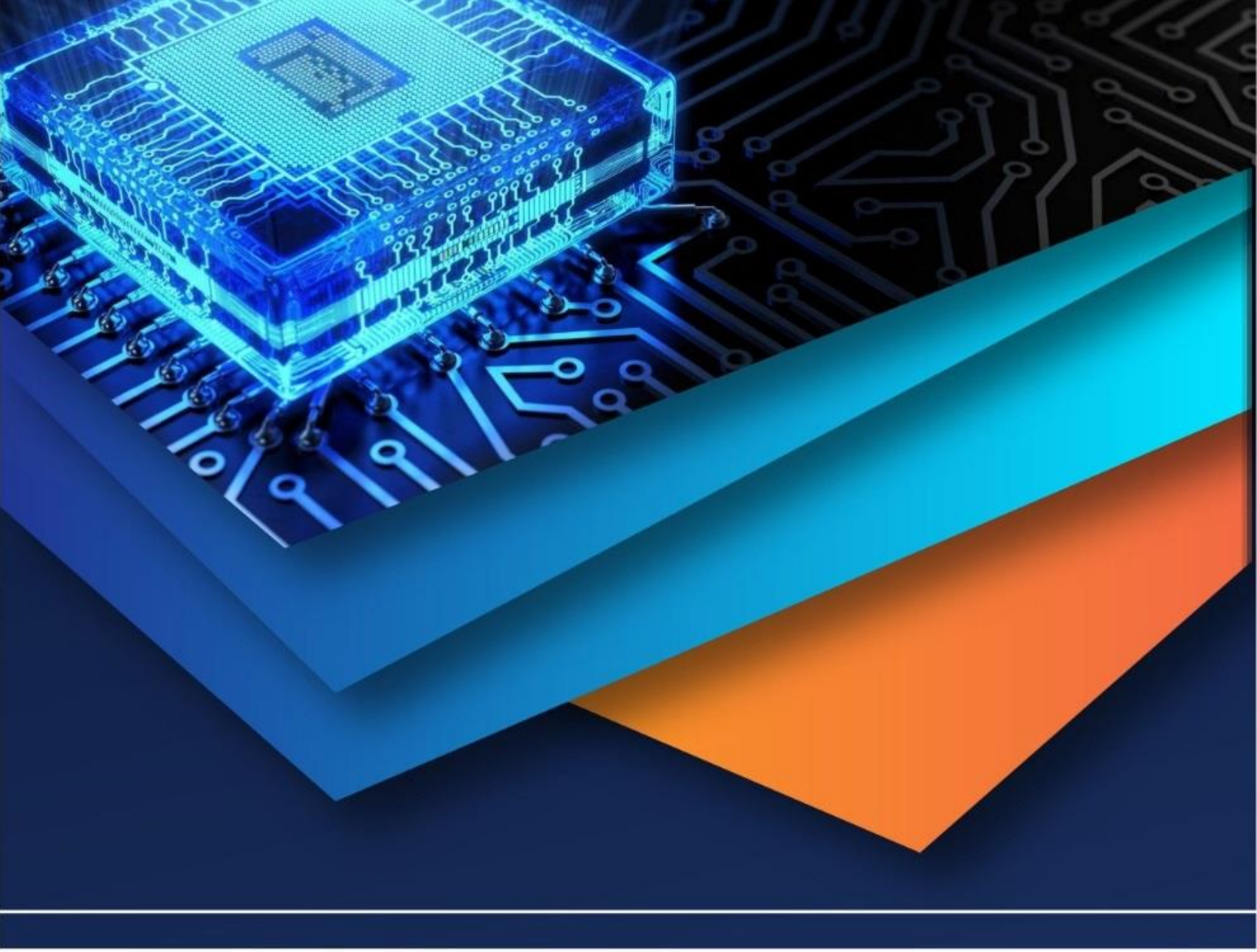
VIII. CONCLUSION

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] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC medical research methodology, vol. 19, no. 1, pp. 1–18, 2019.
[2] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia et al., "Machine learning at facebook: Understanding inference at the edge," in Proc. IEEE Int. Symp. High Perf. Comp, Arch., 2019, pp. 331–344.
[3] K. Bresniker, A. Gavrilovska, J. Holt, D. Milojicic, and T. Tran, "Grand challenge: Applying artificial intelligence and machine learning to cybersecurity," Computer, vol. 52, no. 12, pp. 45–52, 2019.
[4] W. Fleshman, E. Raff, R. Zak, M. McLean, and C.
[5] Nicholas, "Static malware detection & subterfuge: Quantifying the robustness of machine learning and current anti-virus," in Proc. IEEE Int. Conf. Malicious Unwanted Soft., 2018, pp. 1–10.
[6] G. D'Angelo, M. Ficco, and F. Palmieri, "Malware detection in mobile environments based on Autoencoders and API-images," Elsevier J. Parallel Distrib. Comp., vol. 137, pp. 26–33, 2020.
[7] B. Liang, M. Su, W. You, W. Shi, and G. Yang, "Cracking classifiers for evasion: a case study on the google's phishing pages filter," in Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 345–356.
[8] Darktrace, "Machine Learning in the Age of Cyber AI," Tech. Rep., 2020. [Online]. Available: https://www.darktrace.com/es/resources/ wp-machine-learning.
[9] Lastline, "Using AI to detect and contain Cyberthreats," Tech. Rep., 2019. [Online]. Available: https://www.lastline.com/wp-content/ uploads/2020/01/Lastline WP AI Done Right web.pdf
[10] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symp. Secur. Privacy, 2010, pp. 305–316.
[11] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," IEEE T. Knowl. Data. Eng., vol. 26, no. 4, pp. 984–996, 2013.
[12] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the effectiveness of machine and deep learning for cybersecurity," in Proc. IEEE Int. Conf. Cyber Conflicts, May 2018, pp. 371–390.
[13] B. Miller, A. Kantchelian, M. C. Tschantz, S. Afroz, R. Bachwani, R. Faizullabhoy, L. Huang, V. Shankar, T. Wu, G. Yiu et al., "Reviewer integration and performance measurement for malware detection," in Proc. Int. Conf. DIMVA, 2016, pp. 122–141.
[14] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," Computers & Security, vol. 86, pp. 147–167, 2019.
[15] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," IEEE Comm. Surv. Tut., vol. 21, no. 1, pp. 686–728, 2018.
[16] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in Proc. IEEE Int. Conf. Inf. Syst. Secur. Privacy, 2018, pp. 108–116.
[17] S. Garcia, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," Elsevier Comput. Secur., vol. 45, pp. 100–123, 2014.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)