IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 6, June 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |

Identification of Malware Attacks in Network Using Machine Learning

Mrs. Akula Silpa[1] Mr. J. Rakesh Babu[2]

[1]PG Scholar, Department of CSE, Priyadarshini Institute of Technology & Management, GUNTUR, India

[2]Assistant Professor, Department of CSE, Priyadarshini Institute of Technology & Management, GUNTUR, India

ABSTRACT: Contrasted with the past, improvements in PC and correspondence innovations have given broad and propelled changes. The use of new innovations gives incredible advantages to people, organizations, and governments, be that as it may, messes some up against them. For instance, the protection of significant data, security of put away information stages, accessibility of information and so forth. Contingent upon these issues, digital fear based oppression is one of the most significant issues in this day and age. Digital fear, which made a great deal of issues people and establishments, has arrived at a level that could undermine open and nation security by different gatherings, for example, criminal association, proficient people and digital activists. Along these lines, cyber attack detection systems has been created to maintain a strategic distance from digital assaults. With the development of the fifth-generation networks and artificial intelligence technologies, new threats and challenges have emerged to wireless communication system, especially in cyber security. In this project, we offer a review on attack detection methods involving strength of machine learning techniques. Specifically, we firstly summarize fundamental problems of network security and attack detection and introduce several successful related applications using machine learning techniques. Afterwards, we present some benchmark datasets with descriptions and compare the performance of representing approaches to show the current working state of attack detection methods with machine learning structures. Finally, we summarize this project by detecting the type of attack that the dataset has prone to and discuss some ways to improve the performance of attack detection under thoughts of utilizing machine learning structures.

KEYWORDS: Cyber Attack, machine Learning, Artificial Intelligence.

I. INTRODUCTION

Cyber-attacks are increasing within the cyber world. There ought to be some advanced security measures taken to scale back or avoid the amount of cyber-attacks. There are various attacks like D-Dos attacks, Man within the middle, information escape, PROBE, User-To Root, Remote-To Local. These attacks are utilized by the hackers or intruders to realize the unauthorized access to any non-public network, websites, information or perhaps in our personal computers. Therefore, outside or internal hackers use using advanced techniques or finding ways to tickle or break any defense systems to shield the sensitive information, information, money data. Sensible intrusion munitions ought to stop or try and manage varied innovative attacks created or programmed by the hackers. Cyber security refers to the science of technologies, processes, and practices designed to shield networks, devices, programs, and information from attacks, damage, or unauthorized access. Cyber security can also be stated because it's security, within the year 2016, witnessed several advancements in machine learning techniques like self-driven cars, linguistic communication process, health sector, and sensible virtual assistant. They need to be used for locating helpful data from varied audit datasets, which are applied to the matter of intrusion detection. With the assistance of Machine learning technology, we will deploy these ideas in cyber security to boost the protection measures within the intrusion detection system.

II.LITERATURE SURVEY

An IDS generally has to deal with problems such as large network traffic volumes, highly uneven data distribution, the difficulty to realize decision boundaries between normal and abnormal behaviour, and a requirement for continuous adaptation to a constantly changing environment . In general, the challenge is to efficiently capture and classify various behaviours in a computer network. Strategies for classification of network behaviours are typically divided into two categories: misuse detection and anomaly detection. Misuse detection techniques examine both network and system

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |

activity for known instances of misuse using signature matching algorithms. This technique is effective at detecting attacks that are already known. However, novel attacks are often missed giving rise to false negatives. Alerts may be generated by the IDS, but reaction to every alert wastes time and resources leading to instability of the system. To overcome this problem, IDS should not start elimination procedure as soon as the first symptom has been detected but rather it should be patient enough to collect alerts and decide based on the correlation of them. Some research statistics with regards to the impact of cyber security to businesses, organizations, and individuals include: In recent years, cybercrime has been responsible for more than $400 billion in funds stolen and costs to mitigate damages caused by crimes. It has been predicted that a shortage of over 1.8 million cybersecurity workers will be experienced by 2022. It's been predicted that organizations globally will spend at least $100 billion annually on cyber security protection. Attackers currently make over $1 billion in annual revenue from Ransomware attacks, such as Wannacry and Crypto Wall attacks.

2.1Machine Learning

Machine learning is one of the applications of artificial intelligence (AI) that provides computers, the ability to learn automatically and improve from experience instead of explicitly programmed. It focuses on developing computer programs that can access data and use it to learn from themselves. The main aim is to allow computers to learn automatically without human intervention and also adjust actions accordingly.

Machine Learning methods
Machine learning algorithms are often categorized as supervised and unsupervised.

Supervised machine learning algorithms can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to makepredictions about the output values.

Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systemscan infer a function to describe a hidden structure from unlabeled data.

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabeled data for training –typically a small amount of labelled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy.

III.EXISTING SYSTEM

Blameless Bayes and Principal Component Analysis (PCA) were been used with the KDD99 dataset by Almansob and Lomte [9].Similarly, PCA, SVM, and KDD99 were used Chithik and Rabbani for IDS [10]. In Aljawarneh et al's. Paper, their assessment and examinations were conveyed reliant on the NSL-KDD dataset for their IDS model [11] Composing inspects show that KDD99 dataset is continually used for IDS [6]– [10].There are 41 highlights in KDD99 and it was created in 1999. Consequently, KDD99 is old and doesn't give any data about cutting edge new assault types, example, multi day misuses and so forth. In this manner we utilized a cutting-edge and new CICIDS2017 dataset [12] in our investigation.

The present systems of malware detection are reliant on three major methods

- Linguistic Based Methods
- Behaviors-Based Methods
- Graph-Based Methods

Disadvantages

- Strict Regulations.
- Difficult to work with for non-technical users.
- Restrictive to resources.
- Constantly needs Patching.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |

- ❖ Constantly being attacked.

IV.PROPOSED SYSTEM

Instead of using sophisticated and hybrid models, this study relies on relatively simple classification algorithms to solve this problem like Support Vector Machine, KNN, Decision Tree and Random Forest. The dataset is in the form of Excel files which areconverted into plaintext during text pre-processing. This paper has used two feature setsto find the most optimal feature set and respective models. Hence, the data is converted into a compressed sparse row matrix format for modelling. A perfect (or best) model should be the one that reduces underfitting or overfitting. Inproposed work to prevent underfitting and overfitting, the modelling results are evaluated first through a cross-validation score, and then evaluated by evaluation metrics of classification.
Important steps of the algorithm are given in below:-

- Normalization of every dataset.
- Convert that dataset into the testing and training.
- Form IDS models with the help of using RF, KNN, Decision Tree and SVM algorithms.
- Evaluate every model's performances.

Advantages

- Protection from malicious attacks on your network.
- Deletion and/or guaranteeing malicious elements within a preexisting network.
- Prevents users from unauthorized access to the network**.**

V. SYSTEM DESIGN

Introduction of System Analysis

The Systems Development Life Cycle (SDLC), or Software Development Life Cycle in systems engineering, information systems and software engineering, is the process of creating or altering systems, and the models and methodologies that people use to develop these systems. In software engineering the SDLC concept underpins many kinds of software development methodologies. These methodologies form the framework for planning and controlling the creation of an information system the software development process.

Software Model or Architecture Analysis

Structured project management techniques (such as an SDLC) enhance management's control over projects by dividing complex tasks into manageable sections. A software life cycle model is either a descriptive or prescriptive characterization of how software is or should be developed. But none of the SDLC models discuss the key issues like Change management, Incident management and Release management processes within the SDLC process, but it is addresses in the overall project management. In the proposed hypothetical model, the concept of user-developer interaction in the conventional SDLC model has been converted into a three-dimensional model which comprises of the user, owner and the developer. The drawback of addressing these management processes under the overall project management is missing of key technical issues pertaining to software development process that is, these issues are talked in the project management at the surface level but not at the ground level.

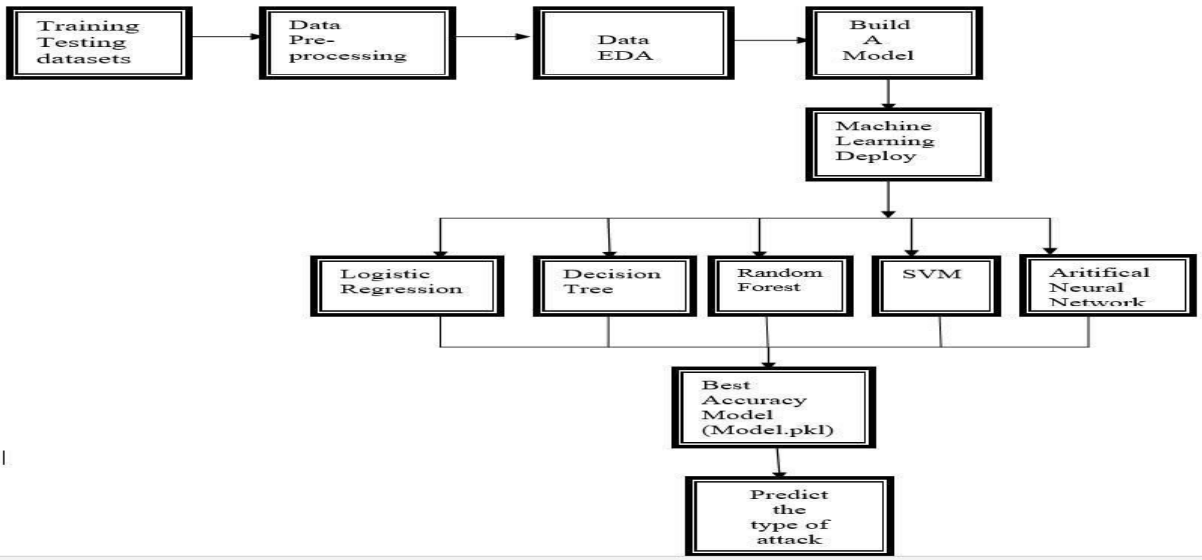
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |

System Architecture



Data Pre-processing

Before feeding data to an algorithm we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better resultsusing a model in Machine Learning, data format has to be in a proper manner. The datashould be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values haveto be managed using raw data.

Data Description

The dataset contains 20 columns. Some of the columns of dataset include flag, protocol type, src bytes, service. The dataset contains a total of 13000 records.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	duration	protocol_t	service	flag	src_bytes	dst_bytes	land	wrong_fra	urgent	hot	num_faile	logged_in	num_com	root_shell	su_attem	num_root	num_file_	num_shel	num_acce	num_
2	1	0	tcp	other	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	2	0	tcp	http	SF	54540	8314	0	0	0	2	0	1	1	0	0	0	0	0	0	
4	3	0	tcp	other	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	4	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	5	0	tcp	other	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	6	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	7	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	8	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	9	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	10	0	tcp	http	SF	54540	8314	0	0	0	2	0	1	1	0	0	0	0	0	0	
12	11	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	12	0	tcp	other	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	13	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	14	0	tcp	other	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	15	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	16	0	tcp	private	RSTR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	17	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	18	0	tcp	private	RSTR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	19	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	20	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	21	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	22	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	23	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	24	0	tcp	other	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	25	0	tcp	http	SF	54540	8314	0	0	0	2	0	1	1	0	0	0	0	0	0	
27	26	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Fig: Data description

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |

Missing values

Filling missing values is one of the pre-processing techniques. The missing valuesin the dataset is represented as '?' but it a non-standard missing value and it has to be converted into a standard missing value NaN. So that pandas can detect the missing values. A missing value can signify a number of different things in your data. Data cleaning should be done on missing data and erroneous data. Data cleaning can be doneby filling missing values manually or by attribute mean or median or the most probablevalue.

Removing Duplicates

Removing duplicates is one of the pre-processing techniques. By having duplicates in the dataset it does not look unique dataset. So, to make the dataset uniquewe need to remove the duplicates from the data set. First we need to check the duplicatesand remove them from the dataset.

Data Exploration

The bar graph given below depicts the frequency of occurrence of attacks in thegiven dataset.

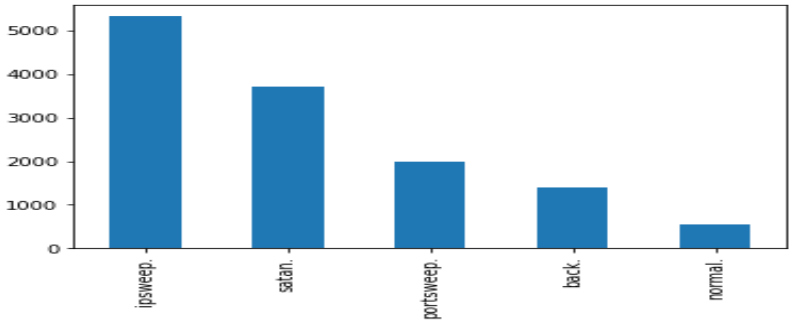


Fig: Count of types of attack

Data Cleaning

Data cleaning is the process of ensuring data is correct, consistent and usable. Youcan clean data by identifying errors or corruptions, correcting or deleting them, or manually processing data as needed to prevent the same errors from occurring. Most aspects of data cleaning can be done through the use of software tools, but aportion of it must be done manually. Although this can make data cleaning an overwhelming task, it is an essential part of managing company data.

Vectorization

Vectorization is a technique by which you can make your code execute fast. It is avery interesting and important way to optimize algorithms when you are implementing itfrom scratch. Now, with the help of highly optimized numerical linear algebra libraries in C/C++, Octave/Matlab, Python, …etc. We can make our code run efficiently.

Feature scaling

It is nothing but data normalization in data processing used to standardize the range of independent variables. This feature is very useful in data pre-processing step.The Standard Scaler will assume that the data is normally distributed with in each attribute and will scale them in such a way that the distribution is now centered on 0, witha standard deviation is of 1.

The mean and standard deviation are calculated for the feature and then the featureis scaled based on:

yi–mean(y)/stdev(y)

yi represents the values of attribute y

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |

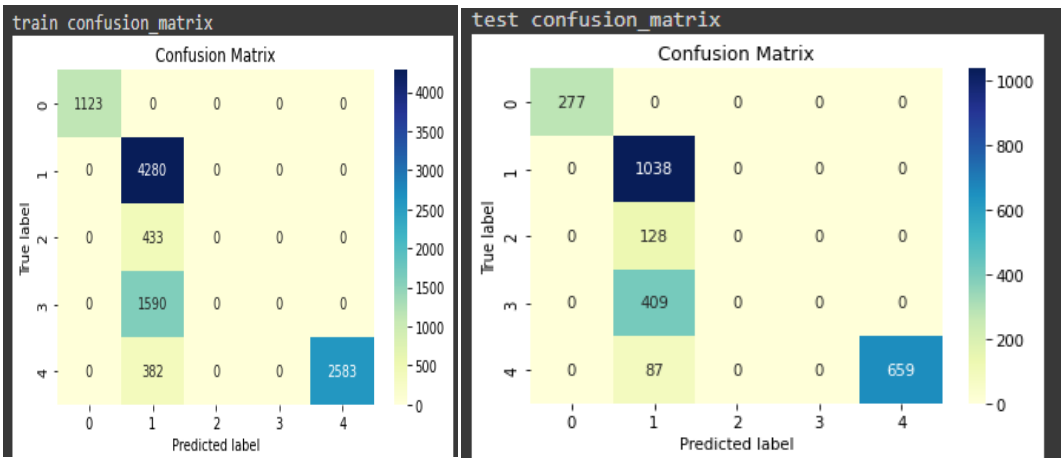


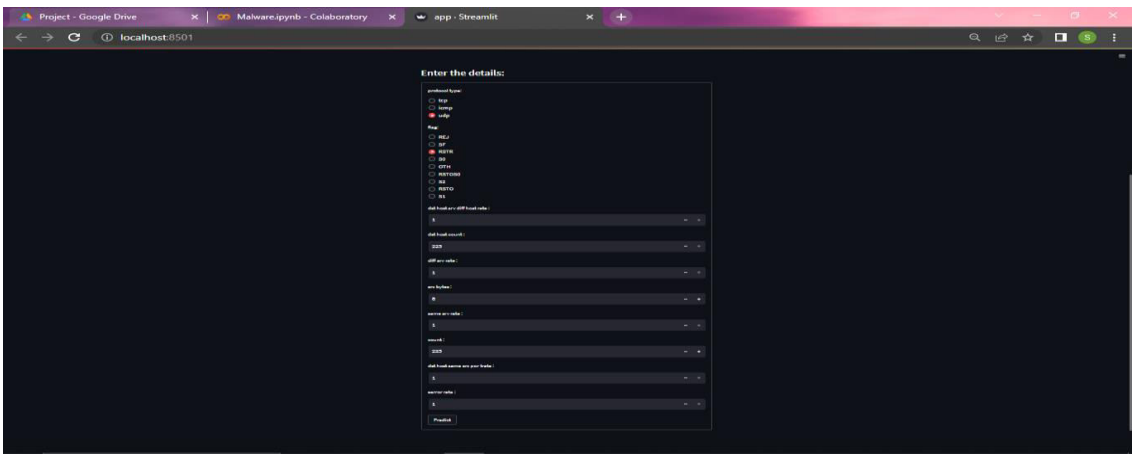
Fig: confusion matrix train and test dataset

Components of Confusion matrix:

A true positive(tp) is a result where the model predicts the positive class correctly. Similarly, a true negative(tn) is an outcome where the model correctly predicts the negative class.
A false positive (fp) is an outcome where the model incorrectly predicts the positive class.And a false negative (fn) is an outcome where the model incorrectly predicts the negativeclass.
Accuracy reflects the total proportion of individuals that are correctly classified. ACC=(tp+tn) / (tp+tn+fp+fn).
F1 score is the harmonic mean of precision and sensitivity F1=2tp/ (2tp+fp +fn).

VI.IMPLEMENTATION

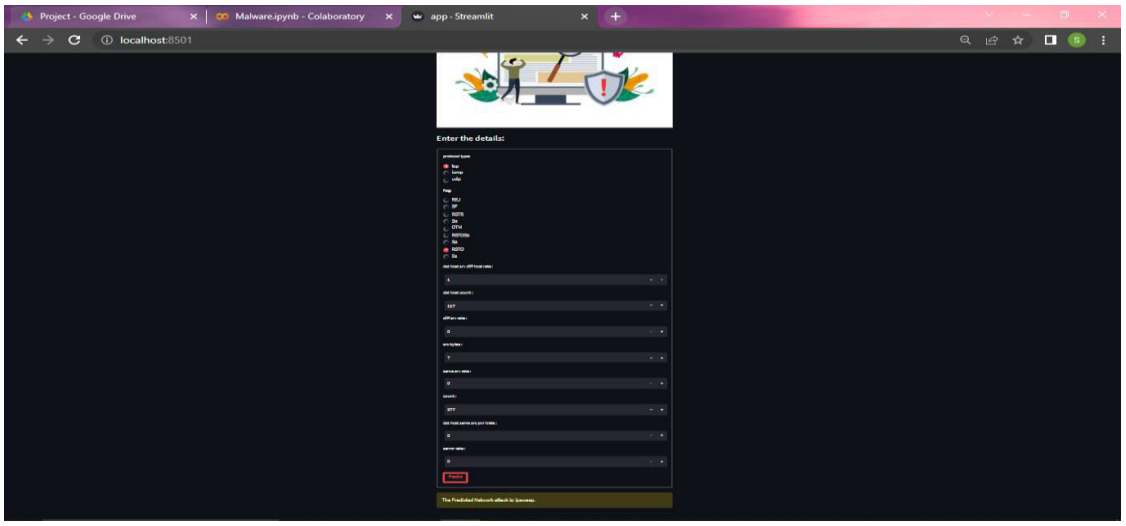
The system is implemented by using ANACONDA software , Anaconda is the world's most popular data science platform and the foundation of modern machine learning. We pioneered the use of Python for data science, champion its vibrant community, and continue to steward open-source projects that make tomorrow's innovations possible. Our enterprise-grade solutions enable corporate, research, and academic institutions around the world to harness the power of opensource for competitive advantage, ground breaking research, and a better world. Providing powerful open source tools in a centralized, collaborative, and version controlled environment Offering a package repository Giving the ability to monitor data science activity via auditing, versioning, and logging Automating model training and deployment on scalable, container-based infrastructure and some monitoring levels have been tested.



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |



VIII.CONCLUSION

Right now, estimations of support vector machine, KNN, Random Forest and decision tree classifier calculations are dependent on modern CICIDS2017 dataset. Results show that the decision tree classifier performed fundamentally preferable outcomes over SVM, RF and KNN. All these calculation helps us to detect the cyber attack in network. It happens in the way that when we consider long back years there may be so many attacks happened so when these attacks are recognized then the features at which values these attacks are happening will be stored in some datasets.
So by using these datasets we are going to predict whether cyber attack is done or not. These predictions can be done by four algorithms like SVM, KNN, RF, Decision tree classifier. This paper helps to identify which algorithm predicts the best accuracy rates which helps to predict best results to identify the cyber attacks happened or not.

VIII.FUTURE ENHANCEMENTS

In future, the model can be optimized to handle imbalanced datasets from various sources and domains. Also, the model can be modified for applying on Hadoop Map Reduce platform. Efficient detection of cyber attacks in malware plays a crucial role. Using Decision tree classifier model cyber attack detection gives the malware patterns, non-malware patterns and general patterns which easily identify the whether there is occurrence of cyber attack or not. The current method does not include the general patterns. It gives the general patterns of which user can decide to determine the type of attack. The current proposed system is for English language sites but as future scope we can design the system for multiple languages.

REFERENCES

1. K. Graves, Ceh: Official certified ethical hacker review guide: Exam 312-50.John Wiley & Sons, 2007.
2. R. Christopher, “Port scanning techniques and the defense against them,” SANSInstitute, 2001.
3. S. Staniford, J. A. Hoagland, and J. M. McAlerney, “Practical automateddetection of stealthy portscans,” Journal of Computer Security, vol. 10, no. 1-2, pp105–136, 2002.
4. S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, “Surveillance detection inhighbandwidth environments,” in DARPA Information Survivability Conferenceand Exposition, 2003. Proceedings, vol. 1. IEEE, 2003, pp. 130–138.
5. K. Ibrahimi and M. Ouaddane, “Management of intrusion detection systemsbasedkdd99:
6. Analysis with lda and pca,” in Wireless Networks and MobileCommunications (WINCOM), 2017 International Conference on. IEEE, 2017, pp.1–6.
7. N. Moustafa and J. Slay, “The significant features of the unsw-nb15 and thekdd99 data set for network intrusion detection systems,” in Building AnalysisDatasets and GatheringExperience Returns for Security (BADGERS), 2015 4th International Workshop onIEEE, 2015, pp. 25–31.
8. L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang,
9. “Detectionandclassification of malicious patterns in network traffic using benford’s law,” inAsia-Pacific Signal and Information Processing Association Annual Summit andConference (APSIPA ASC), 2017. IEEE, 2017, pp.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106025 |

864–872.

10. S. M. Almansob and S. S. Lomte, "Addressing challenges for intrusion detectionsystem using naive bayes and pca algorithm," in Convergence in Technology(I2CT), 2017 2nd International Conference for. IEEE, 2017, pp. 565–568.
11. M. C. Raja and M. M. A. Rabbani, "Combined analysis of support vectormachine and principle component analysis for ids," in IEEE InternationalConference on Communication and Electronics Systems, 2016, pp. 1–5.
12. S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusiondetection system through feature selection analysis and building hybrid efficientmodel," Journal of Computational Science, vol. 25, pp. 152–160, 2018.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details