IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 10, October 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

Intrusion Detection System using Machine Learning

Mr.G.Dhanasekar[1], Ms.A.Vani[2], Ms.G.Varalakshmi[3], Ms.K. Kavitha[4]

Assistant Professor, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[1]

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[2]

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[3]

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[4]

ABSTRACT: Today, world has come closer due to rapid increase internet. As technology has been developed many threats are emerged for the data security which is not at all good for sensitive data transactions, but as we know that the network security also possess equal importance in the computer infrastructure. Because of the intruders the security of the network has become serious problem. Thus to overcome this we are proposing this paper which is based on machine learning algorithm for intrusion detection using SVM and Random Forest, which is based on probabilistic model. This algorithm performs balance detections and keeps false positive rate at acceptable level for different types of real time networking attacks. We demonstrate the design, implementation, and evaluation of Citrus: a novel intrusion detection framework which is adept at tackling emerging threats through the collection and labelling of live attack data by utilising diverse Internet vantage points in order to detect and classify malicious behaviour using graph-based metrics as well as a range of machine learning (ML) algorithms. Citrus considers the importance of ground truth data validation and its flexible software architecture enables both the real-time and offline profiling, detection and classification of emerging cyber-attacks under optimal computational costs. Thus, establishing it as a viable and practical solution for next generation network defence and resilience strategies. In this, the system is trained by arranging the data attributes in a characterised format which eliminates the redundancy resulting in the reduction of complexity.

KEYWORDS: Intrusion Detection, Machine Learning, SVM and Random Forest classifierEtc…

.

I. INTRODUCTION

With the tremendous growth of network-based services and sensitive information on networks, network security is becoming more and more important than ever before. Intrusion detection techniques are the last line of defences against computer attacks behind secure network architecture design, firewalls, passwords, encryptions and personal screening. Despite number of intrusion prevention techniques available, attacks against computer systems are still successful. Thus, intrusion detection systems (IDSs) play a vital role in real time network security. An intrusion is defined as any set of actions that attempt to harm or damage the data which includes a deliberate unauthorised access to the information, manipulate information, or make a system unreliable. Intrusion detection system is a model designed to detect attacks among the various type of packets. It is the process of examining the events which occurs in a computer system or network and analysing them for presence of intrusions.

The decision making process can be categorized as either misuse or anomaly detection. Misuse detection uses predefined attack patterns using signatures of known malware. As a result, novel attacks exploiting unknown vulnerabilities bypass this approach as no corresponding signature exists [2]. Misuse detection relies upon a database of known attack signatures, which is necessarily large and requires constant updates to keep abreast of developing threats. Anomaly detection techniques leverage a description of normal behaviour which is learned through observations in training data, and any future observation which deviates from the normal baseline is labelled as malicious. Consequently, attacks which were not previously encountered are still possible to detect. Nonetheless, the task of profiling normal behaviour is highly challenging due to its dependency on the establishment of ground truth data which in many cases are not validated either statistically or even empirically. Hence, anomaly detection solutions can also be problematic especially when large volume of information is processed, analysed, and statistical normality is not thoroughly assessed. The requirements distilled by the nature of the emerging cyber threat landscape demonstrate that

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

attacks are now conducted on a large scale, thus IDS-based solutions need to maintain vast quantities of either signatures or training data used for decision making. Therefore a high throughput processing framework is required to adequately analyse the data in a reasonable time. Recently, novel frameworks have emerged which possess the ability to handle the large amount of data required for contemporary intrusion detection. Providing anomaly detection algorithms the ability to scale with the influx of emerging threats is crucial to overcome the challenges brought about by the rapid growth of connected devices, innovative infection techniques, and large volumes of data [3]. When using supervised statistical algorithms to determine abnormality in systems telemetry, it is essential to train the models on data which includes benign and relevant attack behaviour. There exist challenges with contemporary intrusion detection data sets, including containing dated and nonrelevant attacks [4], and manually injected attacks which do not accurately reflect the current threat landscape [5]. Honeypot telemetry provides improved training data that better represents the emerging threat landscape [1], as the data contained within is a partial view of the malicious actions currently propagating throughout the Internet. Furthermore, one of the most crucial features in supervised machine learning algorithms is is the ground truth represented as target labels. Current data sets do not either include this feature [6], or engineer it in a non-standard and often problematic manner [7], [8].

This paper is organized in five sections. After this introduction, in Section II, literature survey discussed of the paper, section III about the Existing system, Section IV about Proposed System, as well as the novel feature of the proposed method. Finally, Sections V and VI provide the simulation results and the conclusions and Future work, respectively.

.

II. RELATED WORK

We have reviewed various papers of researchers. The contribution of researchers has been discussed:

Jonathan Palmer [7] have proposed Intrusion detection is an important part of network infrastructures but rule based IDSs can sometimes be difficult and time consuming to maintain. Many machine learning algorithms for intrusion detection have been researched over the past few years to help solve this problem. These techniques require a training data set and many researchers use the KDD '99 data set. In order for these techniques to perform well in real world environments, they must be trained with realistic data sets. This paper set out to determine if the KDD '99 data set was indeed suitable for this application. After attempting to design and perfrom an experiment to test the validity of the KDD '99 dataset, the direct results were inconclusive but lead to a further investigation of the data. Upon reviewing the data set itself and comparing its structure to the structure of some simulated attacks, it would appear that the KDD '99 data set is not the best choice for training machine learning algorithms that would be used in real world applications.

Salem Benferhat, Abdelhamid Boudjelida, Habiba Drias [8] presented.The most adapted Bayesian classification model for intrusion detection is naive Bayes. They present several advantages due to their simple structure. Bayesian naïve networks construction is very simple; it is always easy to consider new scenarios (updates facility). Inference is polynomial, while inference in Bayesian networks with general structures is known to be a hard problem.

Amjad Hussain Bhat, Sabyasachi Patra, Dr. Debasish Jena [9] have proposed Tree algorithm is used to classify network connections into intrusion and normal data based on a labelled training dataset that helps it in building classification patterns. In the second, anomaly detection part, we use hybrid approach of NB Tree and Random forest algorithm is used based on the similarity of connections features. There are challenges in anomaly detection, one challenge is the imbalance between intrusion types in real network connections datasets which are used as training data to our detection system. Some types of intrusions like denial of service (DoS) have many network connections than other intrusion types like user to root (U2R) and normal have more than connection among all; so, any data mining approach will be interested in decreasing the overall error rate of the system regardless of intrusion types, which causes increasing the error rate of the minority attacks like U2R, although these attacks are very dangerous than majority attacks.

Upendra [10] have showed C4.5 selected 7 attribute technique for intrusion detection and performed feature set reduction and evaluated their performance. From the result, it is observed that after applying the feature selection from 41 attributes to 11 and 7 attributes.

NB. Mohan Banerjee, Roopali Soni [11] research on two learning algorithms of data mining i.e. K-means and Naive Bayes classifier. K-means is a clustering algorithm, which work to provide grouping to data sample on the basis of their similarities and dissimilarities. Naive Bayes classifier is classification algorithm which correctly classifies the intrusion/attack. The combination of these two algorithms used in order to improve accuracy, precision rate and reduce the false positive rate. In this paper, the apply one of the efficient data mining algorithms called kmeans clustering via naïve bayes classification for anomaly based network intrusion detection.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

Experimental results on the KDD cup'99 data set show the novelty of our approach in detecting network intrusion. It is observed that the proposed technique performs better in terms of Detection rate when applied to KDD'99 data sets compared to a naïve bayes based approach.

Neethu B [12] have proposed IDSs based on human experts, intrusion detection techniques using machine learning have attracted more and more interests in recent years. Machine learning is a field of study which provides the computers with the ability of learning from previous experience. Machine learning is based heavily on statistical analysis of data and some algorithms can use patterns found in previous data to make decisions about new data.

Dewan Md. Farid, Nouria Harbi, and Mohammad Zahidur Rahman [13] have paper introducing a new hybrid learning algorithm for adaptive network intrusion detection using naive Bayesian classifier and ID3 algorithm, which analyzes the large volume of network data and considers the complex properties of attack behaviours to improve the performance of detection speed and detection accuracy. In this paper we have concentrated on the development of the performance of naïve Bayesian classifier and ID3 algorithm. It has been successfully tested that this hybrid algorithm minimized false positives, as well as maximize balance detection rates on the 5 classes of KDD99 benchmark dataset. The attacks of KDD99 dataset detected with 99% accuracy using proposed algorithm.

Mrutyunjaya Panda and Manas Ranjan Patra [14] have proposed a framework of NIDS based on Naïve Bayes algorithm. The framework builds the patterns of the network services over data sets labelled by the services. With the built patterns, the framework detects attacks in the datasets using the naïve Bayes Classifier algorithm. Compared to the neural network based approach, our approach achieve higher detection rate, less time consuming and has low cost factor. However, it generates somewhat more false positives. As a naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes. This poses a limitation to this research work. In order to alleviate this problem so as to reduce the false positives, active platform or event based classification may be thought of using Bayesian network. We continue our work in this direction in order to build an efficient intrusion detection model.

III. EXISTING SYSTEM

An Intrusion Detection System (IDS) is developed in a network to detect threats from monitoring packets transmitted though. IDSs detect anomalous and malicious activities from inside and outside intruders. An IDS need to deal with problems such as vast network traffic volumes and highly uneven data distribution. The primary function of an IDS is to monitor information sources, such as computers or networks, for unauthorised access activities. IDSs collect data from different systems and network sources and analyse the data for possible threats. IDSs are further developed into network intrusion detection systems (NIDS) and host-based intrusion detection systems (HIDS). Figure 1 shows a general overview of IDSs based on the implemented detection techniques and the deployment environment. Intrusion detection system can be implemented using different methods and techniques. A number of detection mechanisms have been developed to detect abnormalities, which are categorized into statistical methods, data-mining methods and machine learning based methods. NIDS can be implemented using three detection techniques: the signature based detection and the anomaly based detection. A signature based NIDS is limited to detecting from known malicious threats. A combination of the packet header and packet content inspection rules are applied to the detection system from the anomalous traffic flows through signature specification. Anomaly detection techniques are designed to automatically understand attacks which are unknown and unpredictable for signature-based NIDS. Machine learning methods are one of the examples of anomaly based intrusion detection techniques.

IV. PROPOSED METHOD

When using supervised statistical algorithms to determine abnormality in systems telemetry, it is essential to train the models on data which includes benign and relevant attack behaviour. There exist challenges with contemporary intrusion detection data sets, including containing dated and non-relevant attacks, and manually injected attacks which do not accurately reflect the current threat landscape. Honeypot telemetry provides improved training data that better represents the emerging threat landscape, as the data contained within is a partial view of the malicious actions currently propagating throughout the Internet. Furthermore, one of the most crucial features in supervised machine learning algorithms is is the ground truth represented as target labels. Current data sets do not either include this feature, or engineer it in a non-standard and often problematic manner.

The main contributions within this paper are:

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

- An experimental study has been conducted to investigate the accuracy rate, execution time, and memory consumption of existing anomaly detection algorithms.
- A composite clustering algorithm and a novel framework has been proposed based on machine learning technologies.
- The proposed framework has been evaluated in terms of accuracy, memory consumption, and execution time against the existing approaches.

A. System Architecture

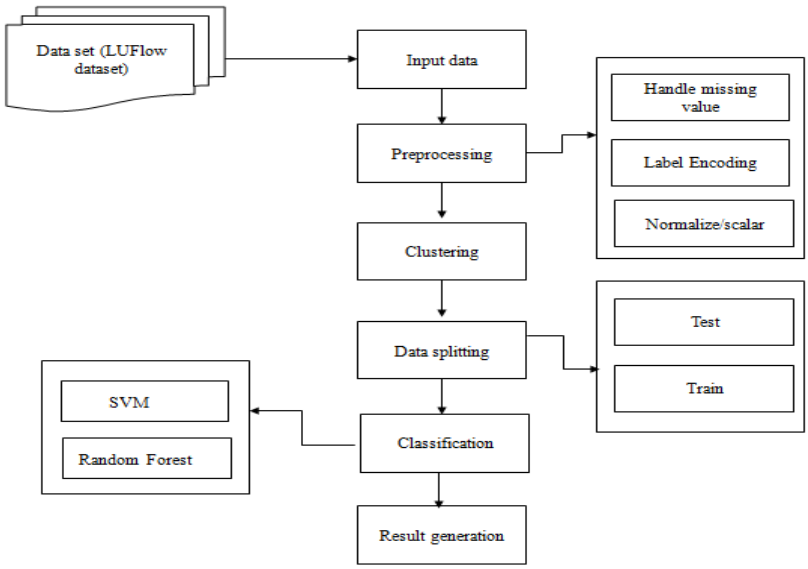


Fig.1:System Architecture

As shown in Fig.1We are proposing the Intrusion Detection System using SVM and Random Forest algorithm which mainly implies the detection of abnormal packets using past experience of the system. Here the incoming packets are analysed and categorised according to values of the attributes to produce dataset. Using this data set the next arriving packets are detected as normal or abnormal packets. If abnormal packets are detected reporting can be done.

B. Implementation

Modules:

- Data selection
- Data preprocessing
- Data clustering
- Data splitting
- Classification
- Result Generation

C. Modules Description:

Data Selection Module:

- The input data was collected from dataset repository.
- In our process, the LFlow dataset is used.
- Data selection is the process of selecting the appropriate data set for processing.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

- Each of the record consists of 16 features and one marked as attack.
- The LUFlow Dataset is used for detecting the intrusion. All the data's are selected and loaded into the database for detecting the intrusion.

Data Pre-processing:

- Data pre-processing is the process of removing the unwanted data from the dataset.
- Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning.
- This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makes it more efficient.
- Missing data removal
- Encoding Categorical data
- Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.
- Missing and duplicate values were removed and data was cleaned of any abnormalities.
- Encoding Categorical data: That categorical data is defined as variables with a finite set of label values.
- That most machine learning algorithms require numerical input and output variables.

Data Clustering:

- Here our dataset will be Clustered into two categories like,
- (i). Benign (ii). Malicious.
- **Benign** is to suggest it is not dangerous or serious.In general, a benign grows slowly and is not harmful.
- **Malicious** attacks on the basis of the specific patterns such as number of bytes or number of 1's or number of 0's in the network traffic. It also detects on the basis of the already known malicious instruction sequence that is used by the malware.

Data Splitting:

- During the machine learning process, data are needed so that learning can take place.
- In addition to the data required for training, test data are needed to evaluate the performance of the algorithm in order to see how well it works.
- In our process, we considered 70% of the LUFlow dataset to be the training data and the remaining 30% to be the testing data.
- Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
- One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.

Result Generation

- The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,
- The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,
- Accuracy
- Precision
- Specificity
- Recall

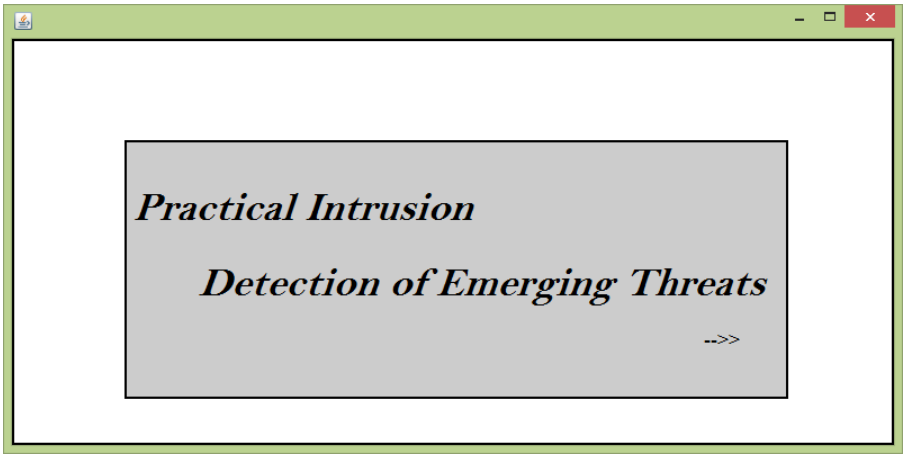
| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

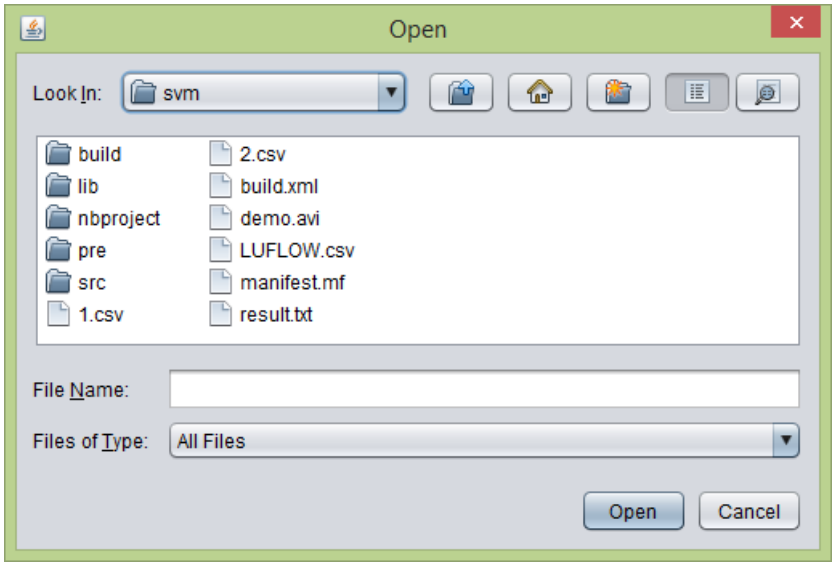
||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

V. SIMULATION RESULTS

Data selection



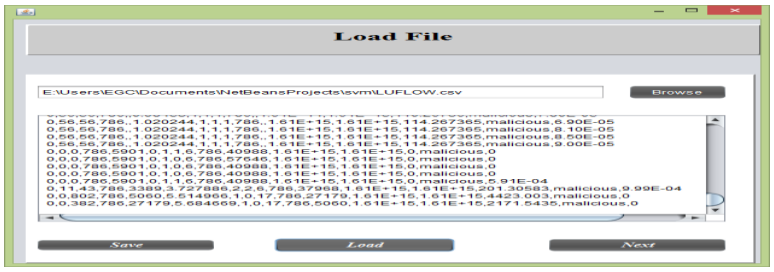


| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

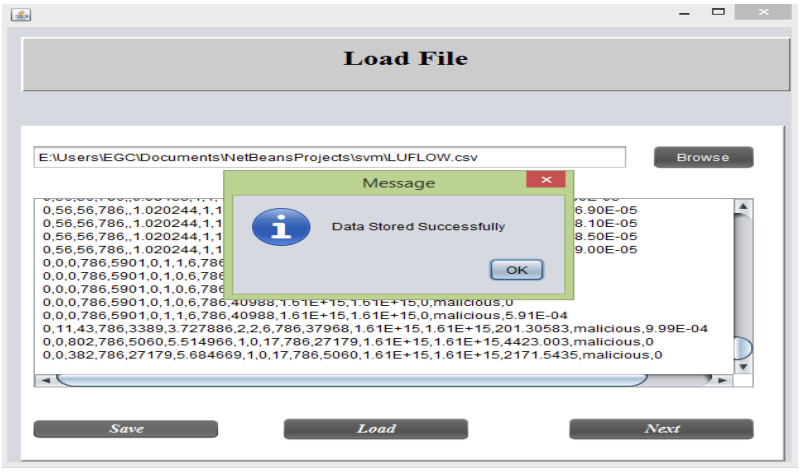
||Volume 11, Issue 10, October 2022||

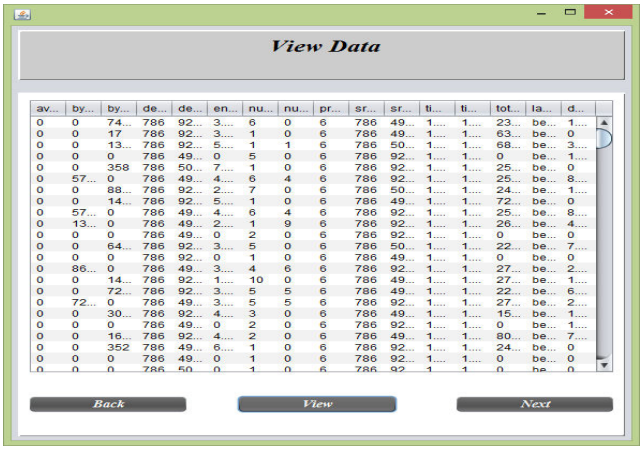
| DOI:10.15680/IJIRSET.2022.1110056 |

Data Loading



Saving the Data



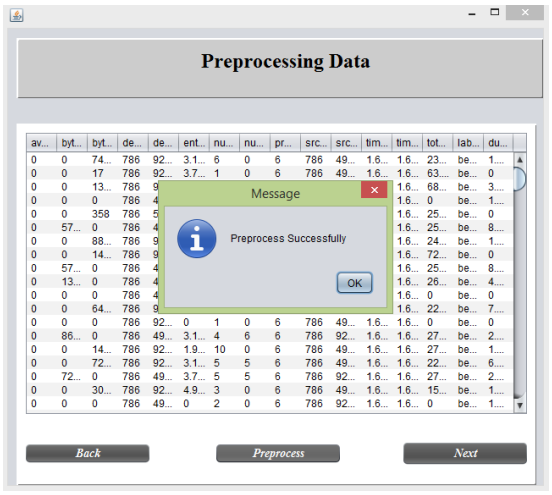


| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

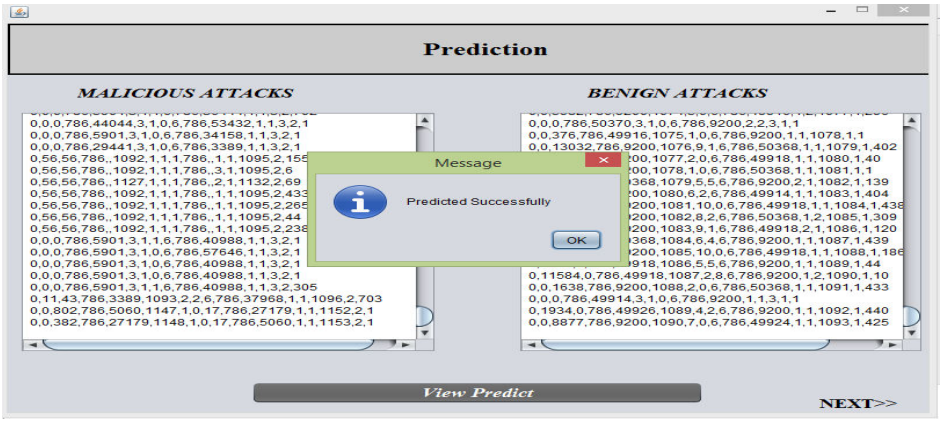
| DOI:10.15680/IJIRSET.2022.1110056 |

Pre-processing:



Classification:

Data Splitting

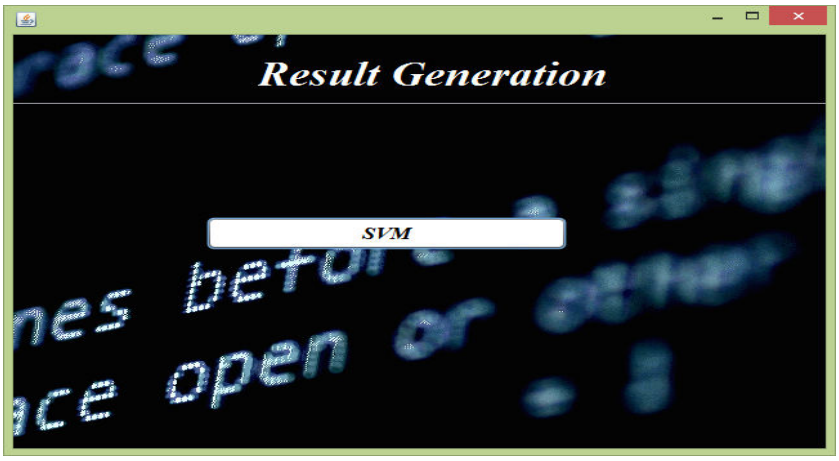


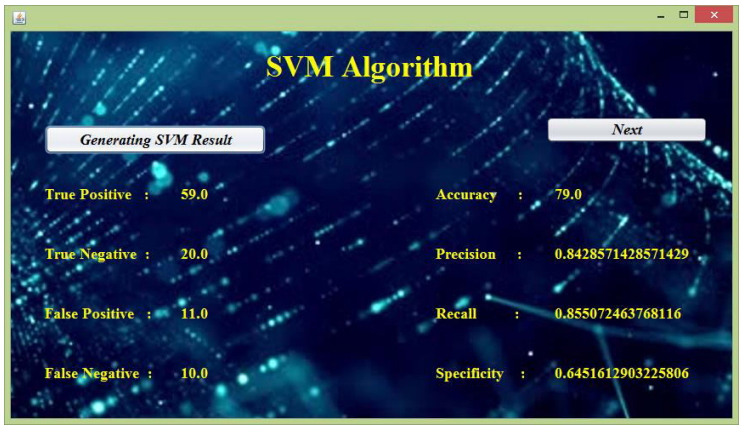
| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

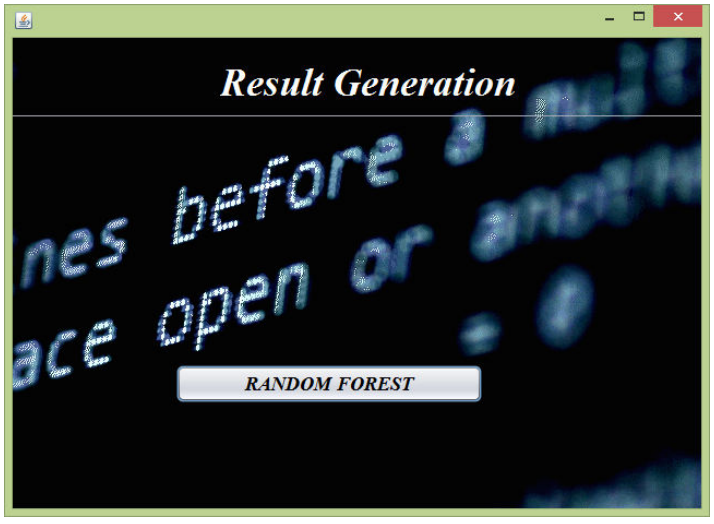
||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

Result Generation:



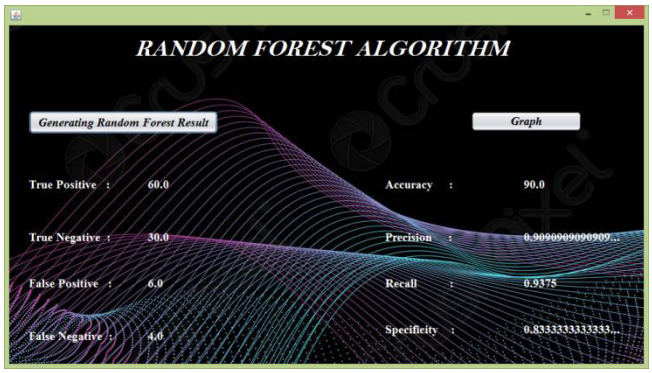


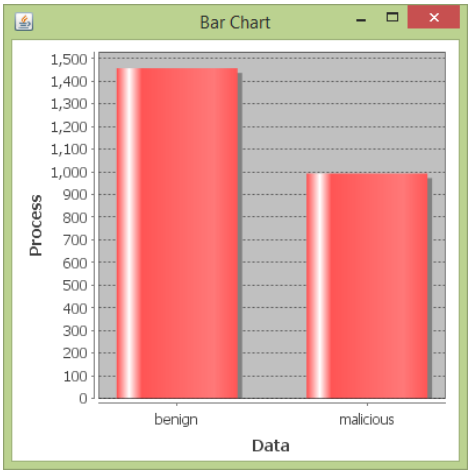


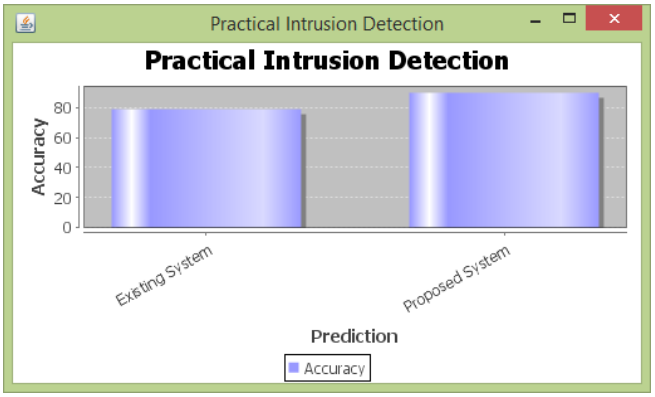
| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |







VI. Conclusion And Future Work

In this article, presents the design, implementation, and evaluation of a practical solution for intrusion detection. An evaluation of the labelling approach is conducted to ensure the accuracy of the identified attack super nodes using Machine learning Algorithms such as SVM and Random Forest. In the proposed model, the accuracy of the detection capabilities, and demonstrates the efficiency of parallel computation through the evaluation and comparison of model training implementations.

Future Work

As a future work, it would be interesting to evaluate the performance of some unsupervised algorithms. In the future, we should like to combine different machine learning and deep learning algorithms as a multi-layered model to

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 10, October 2022||

| DOI:10.15680/IJIRSET.2022.1110056 |

improve the detection performance. Future extension of this article includes extending the proposed model to multivariate time series including machine states and external sensors data.

REFERENCES

[1] Prof. D.P.Gaikwad and Dr.R.C.Thool, Architecture Taxonomy and Product of IDS,International Conference on Computer Applications,Computer Application-II,doi:10.3850/978-981-08-7304-2_0382.
[2] Christine Dartigue, Hyun Ik Jang, and Wenjun Zeng Computer Science Department University of Missouri-Columbia Columbia, Missouri, USA , Services Research Conference on A New Data-Mining Based Approach for Network Intrusion Detection, 2009 IEEE Seventh Annual Communication Networks and Services Research Conference.
[3] Sandhya Peddabachigari, Ajith Abraham, Johnson Thomas,Department of Computer Science, Oklahoma State University, USA paper on Intrusion Detection Systems Using Decision Treesand Support Vector Machines.
[4] G.Prashanth, V.Prashanth, P.Jayashree, N.Srinivasan, IEEE-International Conference on Signal processing, Communications and Networking ,Using Random Forests for Network-based Anomaly detection at Active routers ,Madras Institute of Technology, Anna University Chennai India, Jan 4-6, 2008. Pp93-96.
[5] Mehdi MORADI and Mohammad ZULKERNINE, paper on A Neural Network Based System for Intrusion Detection and Classification of Attacks.
[6] Amira Sayed A. Aziz,Mostafa A. Salama,Aboul ella Hassanien,Sanaa ElOla Hanafi paper on Artificial Immune System Inspired Intrusion Detection System Using Genetic Algorithm.
[7] Jonathan Palmer, international paper on Naive Bayes Classification for Intrusion Detection Using Live Packet Capture.
[8] Salem Benferhat, Abdelhamid Boudjelida, Habiba Drias, research on An Intrusion Detection Approach Based on Tree Augmented Naive Bayes and Expert Knowledge.
[9] Amjad Hussain Bhat, Sabyasachi Patra, Dr. Debasish Jena, IEEE paper on Machine Learning Approach for Intrusion Detection on Cloud Virtual Machines, june 2013.
[10] Upendra,Assistant Professor, CSE Department, NIT Raipur, C.G., India, International Journal of Emerging Trends & Technology in Computer Science on An Efficient Feature Reduction Comparison of Machine Learning Algorithms for Intrusion Detection System, Volume 2, Issue 1, January – February 2013.
[11] Mohan Banerjee, Roopali Soni, MTech (CSE) Scholar, Department of Computer Science & Engineering, Thakral College of Technology, Bhopal HOD & AP, Department of Computer Science & Engineering, Thakral College of Technology, Bhopal, International Journal of Science, Engineering and Technology Research on Design and Implementation of Network Intrusion Detection System by using K-means clustering and Naïve Bayes , Volume 2, Issue 3, March 2013.
[12] Neethu B,,Department of Computer Science, Amrita University,International Journal of Electronics and Computer Science Engineering on Classification of Intrusion Detection Dataset using machine learning Approaches.
[13] Dewan Md. Farid1, Nouria Harbi1, and Mohammad Zahidur Rahman2,Department of Computer Science and Engineering,Jahangirnagar University, International Journal of Network Security & Its Applications (IJNSA) on Combining Naïve Bayes and Decision Tree For Adaptive Intrusion Detection.
[14] Mrutyunjaya Panda and Manas Ranjan Patra, Department of E &TC Engineering, G.I.E.T., Gunupur, India, Department of Computer Science, Berhampur University, Berhampur, India, International Journal of Computer Science and Network Security on Network Intrusion Detection Naïve Bayes, VOL.7 No.12, December 2007258,Manuscript received December 5, 2007,Manuscript revised December 20, 2007.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details