iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45588

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 10 Issue VII July 2022- Available at www.ijraset.com

Comparison of Machine Learning and Deep Learning algorithms for Detecting Intrusions in Network

Bhargav H R[1], Chandan K S[2], Darpan N Karur[3], Thanmai D L[4], Dr. Jyoti Neeli[5]
[1, 2, 3, 4]*Student, Global Academy of Technology, Bengaluru*
[5]*Professor and HOD, Department of Information Science and Engineering, Global Academy of Technology, Bengaluru*

Abstract: ***Due to the introduction of the devices for networking with the fast internet development in earlier years, the safety of the networks has developed to be important in this contemporary age. Intrusion Detection Systems are used in identifying unapproved, unacquainted and traffic that is suspicious through networks. This project pursues the anomaly detection through the design of a hybrid model that classifies a network traffic first either as benign or intrusive. When the traffic is established as intrusive, the model additionally detects the intrusive traffic category traveling throughout the network. Furthermore, the research proposes deep learning and machine learning algorithm usage in determining the quickest and utmost precise algorithm for network intrusions detection.***
Keywords: Intrusion Detection Systems (IDS), Deep Learning Model, Machine Learning Model, Anomaly Detection

I. INTRODUCTION

With the wide spreading usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate. Intrusion detection is the first step to prevent security attack. Hence the security solutions such as Firewall, Intrusion Detection System (IDS), Unified Threat Modelling (UTM) and Intrusion Prevention System (IPS) are getting much attention in studies. IDS detects attacks from a variety of systems and network sources by collecting information and then analyses the information for possible security breaches. The network-based IDS analyses the data packets that travel over a network and this analysis are carried out in two ways. Till today anomaly-based detection is far behind than the detection that works based on signature and hence anomaly-based detection remains a major area for research. The challenges with anomaly-based intrusion detection are that it needs to deal with novel attack for which there is no prior knowledge to identify the anomaly. Hence the system somehow needs to have the intelligence to segregate which traffic is harmless and which one is malicious or anomalous and for that machine learning techniques are being explored by the researchers over the last few years. IDS however is not an answer to all security related problems. For example, IDS cannot compensate weak identification and authentication mechanisms or if there is a weakness in the network protocols.

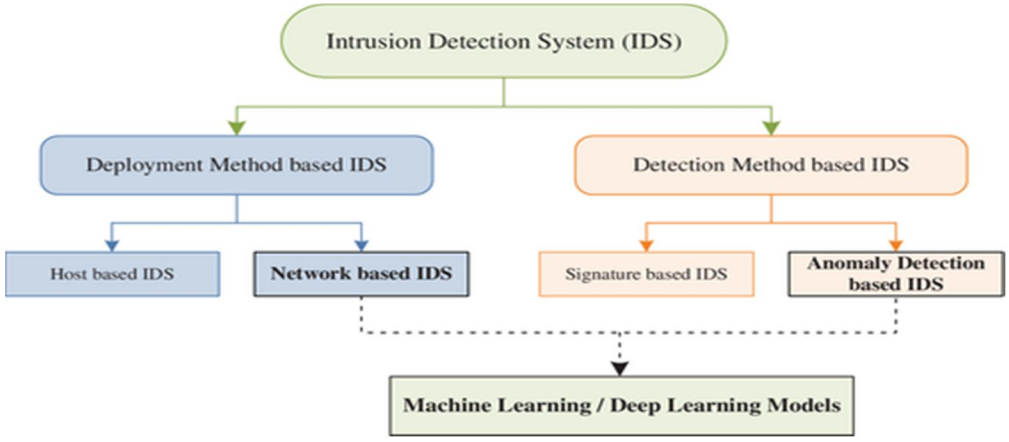


Figure 1: Types Of IDS

©IJRASET: All Rights are Reserved

Some common detection techniques include anomaly-based and signature-based approaches. While anomaly-based detection is a dynamic method with a reliable model that contrasts with harmful activity, signature-based detection is a static method with established patterns. A signature-based detection technique might not work against unidentified attackers. Anomaly-based detection trains a model based on typical activity to detect any anomalies. This model is made up of both machine learning (ML) and deep learning (DL) architectures. These anomalies from unidentified attacks can also be found by them.

The predictability of AI models could lean toward accuracy or interpretability. Both traits are important. Accurate or black-box models such as neural networks or complicated ensemble models provide high accuracy, but they do not provide feature importance. On the other hand, white-box models like linear regression or decision trees can provide feature engineering but might lack in accuracy for complex datasets. Hence, both ML and DL models should be taken into consideration

II. PROBLEM STATEMENT

The growth of the Internet and data traffic exhibited several problems regarding security management. The increase in the number of users around the world has introduced the need to incorporate access controls. Intruders are finding best methods to infiltrate or disrupt network traffic. They continue to adapt prevention mechanisms in place and continue to find ways to exploit the systems that are in place to prevent these intrusions from occurring. Current intrusion detection systems (IDSs) suffer to attain both the high detection rate and low false alarm rate. To address this issue, we propose an IDS using different machine learning (ML) and deep learning (DL) models. We present a comparative analysis of different ML models and DL models on NSL-KDD dataset.

III. METHODOLOGY

The study was conducted using the Knowledge Discovery in Databases (KDD) approach. This approach features procedures such as data selection/ transformation/, evaluation, pre-procession collection and modelling. Some of these processes are explained below:

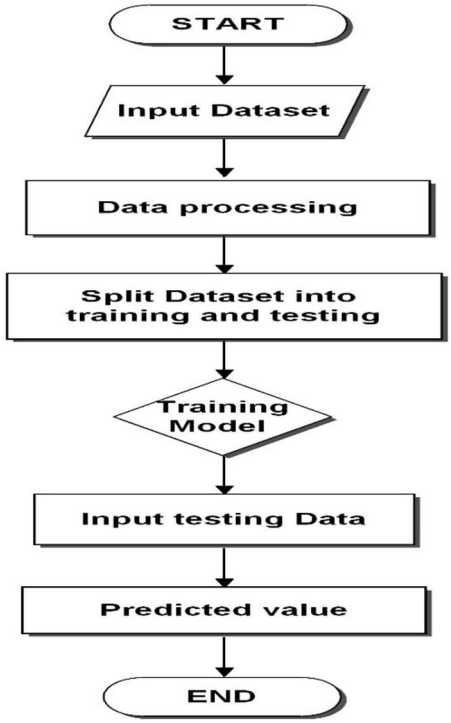


Figure 2: Flow Chart Diagram

A. Data Collection And Pre-Processing

1) The job here is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analysing results with the help of statistical techniques.

2) The purpose of pre-processing is to convert raw data into a form that fits machine learning. Structured and clean data allows to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

3) Some Dataset examples - CIC-IDS2017, NSL-KDD etc

.

B. Data Visualisation

1) Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.

2) Exploratory Data Analysis (EDA) is a process of describing the data by means of statistical and visualization techniques in order to bring important aspects of that data into focus for further analysis.

C. Data Splitting and Training

1) Data Splitting can be done in 3 ways, they are:

2) Training set: A training set can be used to train a model and define its optimal parameters.

3) Test set: A test set is needed for an evaluation of the trained model and its capability for generalization.

4) Validation Set: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

5) Model Training: A machine learning training model is a process in which a machine learning (ML) algorithm is fed with sufficient training data to learn from.

D. Prediction Methodology

1) In this method, the test data without label must give prediction module which generate using training method, this prediction module accept the test data and process. Finally, it will produce the accuracy module.

2) The model used to predict the unknown value is known as predictor.

3) The predictor is constructed from a training set and its accuracy refers to how well it can estimate the value of new data.

IV. SYSTEM ARCHITECTURE

A system architecture as shown in Figure 3is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.

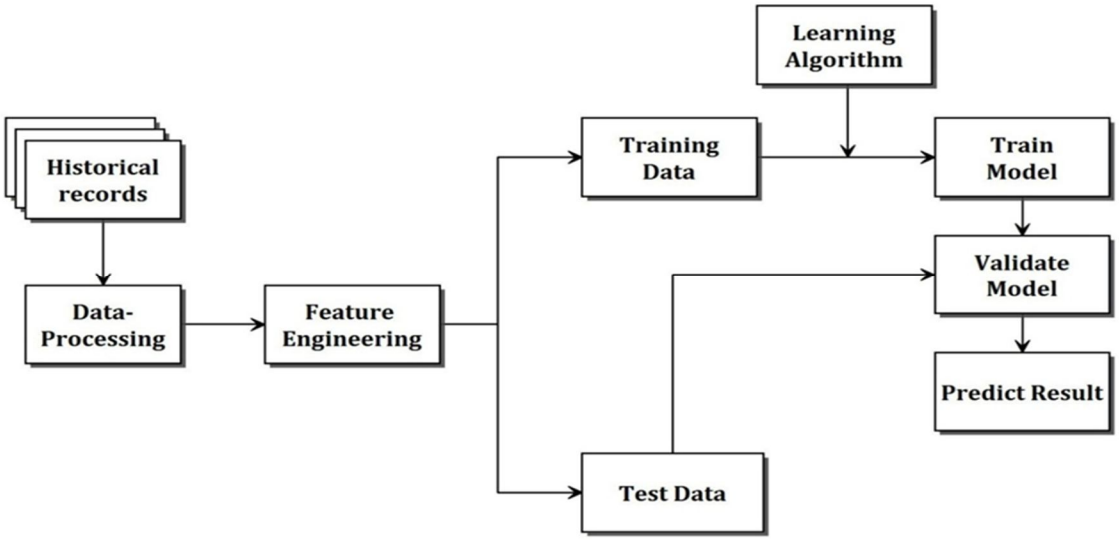


Figure 3: System Architecture

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

V. MODEL EVALUATION

The study used different evaluation metrics namely confusion matrix plot, accuracy, precision and recall training the model after the testing process was complete.

A. Confusion Plot

This is a graph that illustrates the actual and predicted observations in the models. The confusion plot has n rows and n columns with n being the total number of classes in the dataset. The confusion plot guarantees efficiency in the values and can be used to calculate metrics such as precision, recall, and accuracy.

B. Accuracy

The accuracy metric was also used to conduct the study. The metric is commonly used to assess model's performance and can be determined by dividing the number of correct predictions by the total number of observations. Accuracy is expressed as a percentage and the value ranges from 0 to 100. The formula is shown in Figure 4:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}}$$

Figure 4: Accuracy

C. Precision

The precision metric, also known as the positive predictive value (PPV) is expressed as a ratio of the positive classes that were correctly predicted against the total predictions. The formula for precision is shown in Figure 5:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Figure 5: Precision

D. Recall

Recall determines the correctly predicted classes based on the total number of observations. It is also referred to as sensitivity. The formula for recall is as shown in Figure 6:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Figure 6: Recall

VI. RESULT

This section discusses the results of the experiments conducted in this study. The study will also provide a general overview of the performance evaluation based on the algorithms that were earlier identified. The confusion matrix was used to visualise the correctly predicted classifications and falsely predicted classifications across each traffic type. Additionally, each model (experiment) was evaluated based on the precision, recall, f1-score and accuracy. The figures below demonstrate the confusion matrix for running the three experiments. The algorithms used to train the models were the random forest, k nearest neighbour and long short-term memory. From the training and validation loss and accuracy chart, the model does not appear to be over fitting. Each model was then tested with different data (test data) which was not used in the training process. The test data also includes the 4 categories of attacks in the main data.

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

A. Algorithm 1: Random Forrest Classifier

Confusion Matrix for RFC is shown in Figure 7 and the results of the Random Forest Classifier algorithm using the percentage split technique is shown in Figure 8.

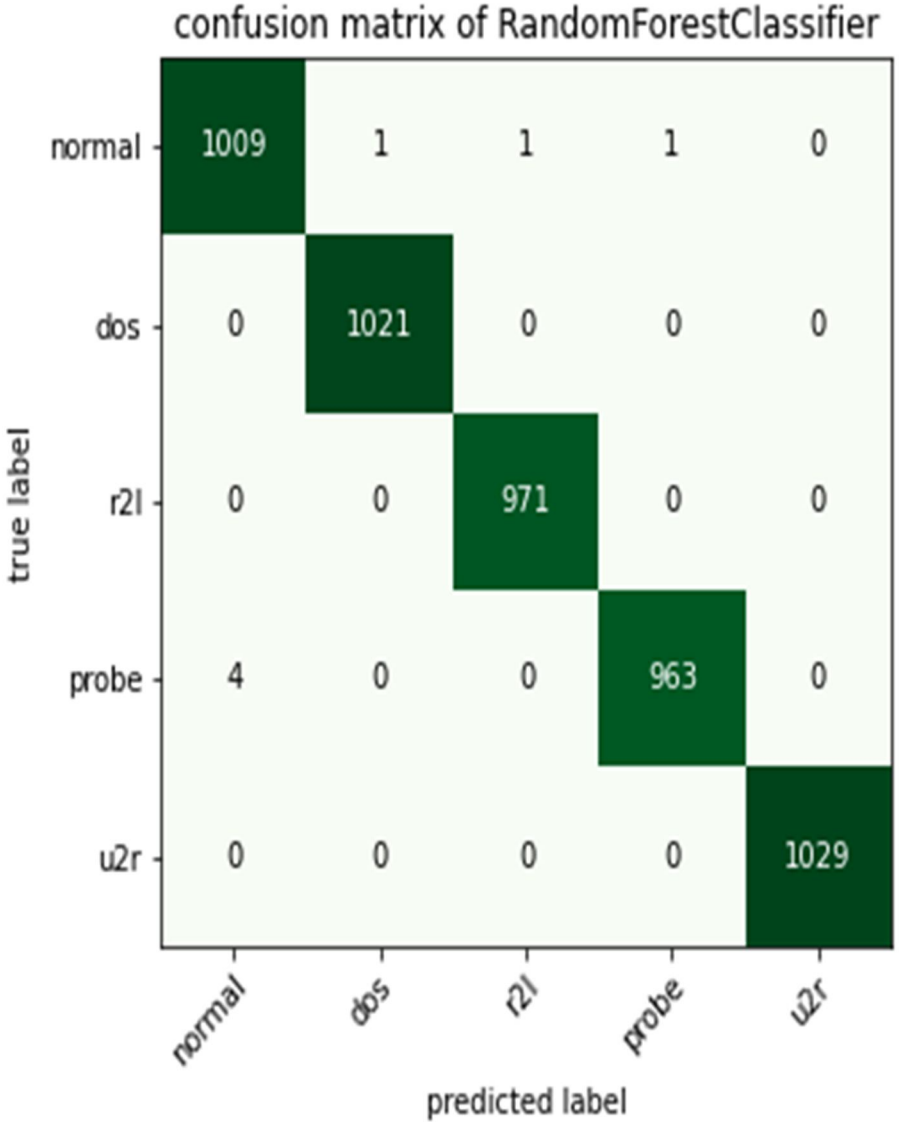


Figure 7: Confusion Matrix for RFC

	precision	recall	f1-score	support
normal	1.00	1.00	1.00	1012
dos	1.00	1.00	1.00	1021
r2l	1.00	1.00	1.00	971
probe	1.00	1.00	1.00	967
u2r	1.00	1.00	1.00	1029
accuracy			1.00	5000
macro avg	1.00	1.00	1.00	5000
weighted avg	1.00	1.00	1.00	5000

Figure 8: Classification Report for RFC

B. Algorithm 2: K Nearest Neighbours

Confusion Matrix for KNN is shown in Figure 9 and the results of the k-nearest neighbours algorithm using the percentage split technique is shown in Figure 10.

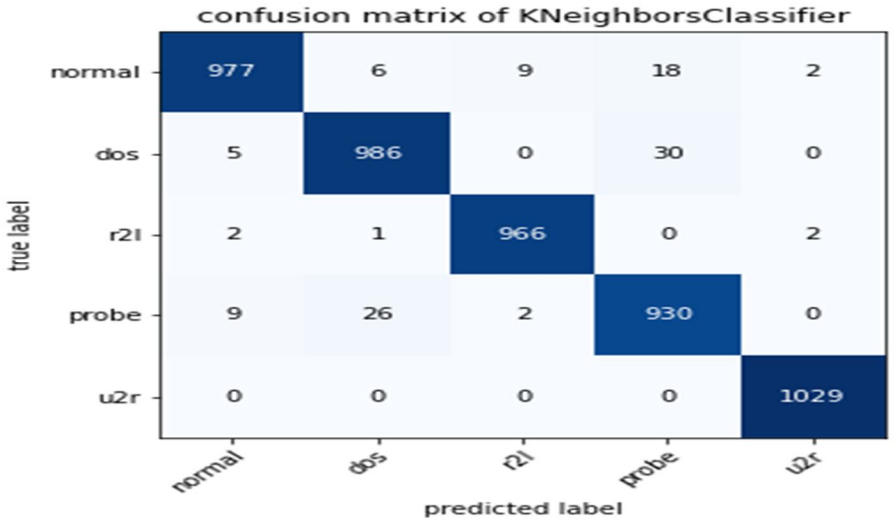


Figure 9: Confusion Matrix for KNN

	precision	recall	f1-score	support
normal	0.98	0.97	0.97	1012
dos	0.97	0.97	0.97	1021
r2l	0.99	0.99	0.99	971
probe	0.95	0.96	0.96	967
u2r	1.00	1.00	1.00	1029
accuracy			0.98	5000
macro avg	0.98	0.98	0.98	5000
weighted avg	0.98	0.98	0.98	5000

Figure 10: Classification Report for KNN

C. Algorithm 3: Long Short Term Memory

Confusion Matrix for LSTM is shown in Figure 11 and the results of the Long short term memory algorithm using the percentage split technique is shown in Figure 12.

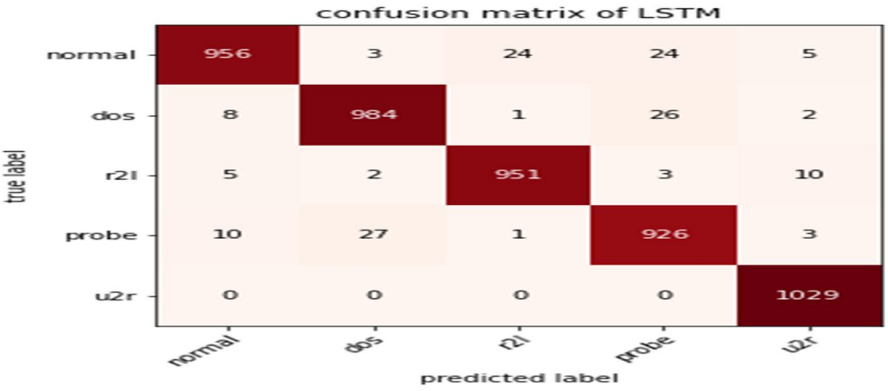


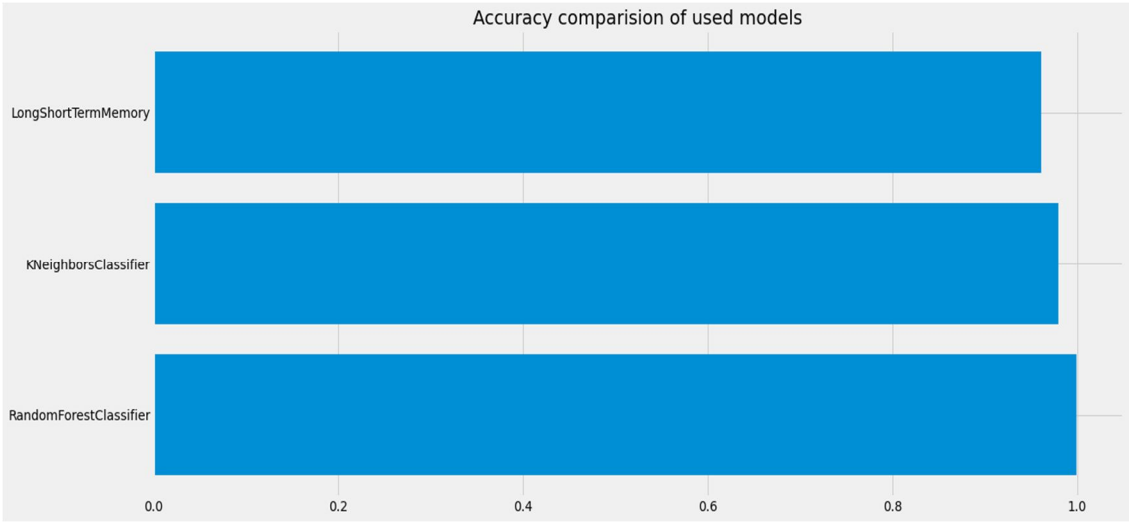
Figure 11: Confusion Matrix for LSTM

	precision	recall	f1-score	support
normal	0.98	0.94	0.96	1012
dos	0.97	0.96	0.97	1021
r2l	0.97	0.98	0.98	971
probe	0.95	0.96	0.95	967
u2r	0.98	1.00	0.99	1029
accuracy			0.97	5000
macro avg	0.97	0.97	0.97	5000
weighted avg	0.97	0.97	0.97	5000

Figure 10: Classification Report for LSTM

D. Performance Analysis

The accuracy comparison of Random Forest Classifier, K Nearest Neighbour Classifier and Long Short Term Memory is shown in Figure 12.



VII. KEY CHALLENGES

1) Uncleaned data might lead the algorithms to make wrong predictions.
2) Data oversampling might lead the prediction get negatively affected.
3) The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time.
4) Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results.
5) Due to regular updates in data, present model might become inaccurate in future. So regular monitoring and maintenance is required to keep the algorithm working.

VIII. CONCLUSION

We have implemented different ML and DL algorithms as an intrusion detection system in this study. As a benchmark dataset, NSL-KDD dataset has been used. The dataset is balanced by re-sampling such that all classes have the same amount of data. Three experiments were performed based on three different algorithms: Random Forest Classifier, K-neighbours classifier, and the Long Short Term Memory algorithm. The first two experiment involving a Machine Learning algorithm and the third involving a Deep Learning algorithm.

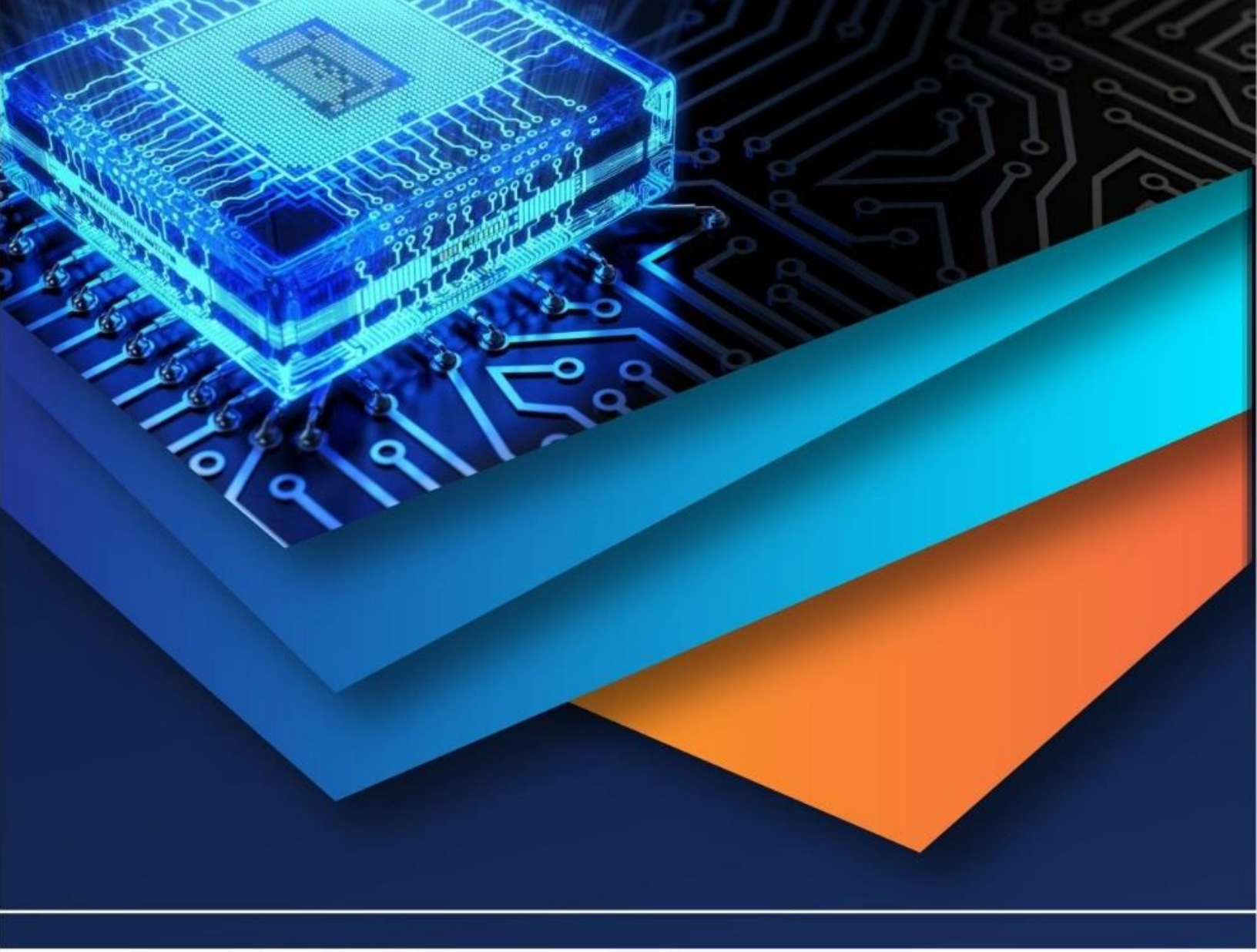
The research emphasises the need to improve information security measures given the high rate of intrusion reported in recent times. Some of the constraints to take into consideration would be factors such as limited power, storage, and processing capabilities before training a dataset to mitigate potential network threats. These constraints also qualify as some of the challenges that were encountered while conducting these experiments.

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

REFERENCES

[1] R. Hattarki, S. Houji and M. Dhage, "Real Time Intrusion Detection System ", 2021 6th International Conference for Convergence in Technology (I2CT), 2021.
[2] A. Kumar and T. Lim, "EDIMA: Early Detection of Malware Network Activity Using Machine Learning Techniques".
[3] Sheikh, N.U., Rahman, H., Vikram, S. and AlQahtani, H., 2018. “A Lightweight Signature-Based”.
[4] M. Eskandari, Z. Janjua, M. Vecchio and F. Antonelli, "Passban IDS: An Intelligent Anomaly-Based Intrusion Detection System”
[5] Q. Ngo, H. Nguyen, V. Le and D. Nguyen, "A survey of malware and detection methods based on static features", ICT Express, vol. 6, no. 4, pp. 280-286, 2020.
[6] Ullah and Q. H. Mahmoud, "A Two-Level Hybrid Model for Anomalous Activity Detection in Networks", Consumer Communication and Networking Conference (CCNC), 2019
[7] H. Song, M. J. Lynch, and J. K. Cochran, “A macro-social exploratory analysis of the rate of interstate cyber-victimization,” American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601, 2016.
[8] P. Alaei and F. Noorbehbahani, “Incremental anomaly-based intrusion detection system using limited labeled data,” in Web Research (ICWR), 2017 3th International Conference on, 2017, pp. 178–184.
[9] M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, “Modeling and implementation approach to evaluate the intrusion detection system,” in International Conference on Networked Systems, 2015, pp. 513–517.
[10] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Toward credible evaluation of anomaly-based intrusion-detection methods,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 516–524, 2010.
[11] A. S. Ashoor and S. Gore, “Importance of intrusion detection system (IDS),” International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4, 2011.
[12] A. S. Ashoor and S. Gore, “Importance of intrusion detection system (IDS),” International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4, 2011.
[13] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, “Anomaly-based network intrusion detection: Techniques, systems and challenges,” computers & security, vol. 28, no. 1–2, pp. 18–28, 2009.

©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)