A Hybrid Intrusion Detection System for Botnet attack with Data Technique

Neelu Singh*, Dr. Varsha Jotwani 2

Abstract:

A difficult problem in the realm of intrusion detection systems (IDS) is estimating the progress made in the identification of malicious code. Machine learning IDS training is dependent on the datasets provided, but gathering a valid dataset for comparison is difficult. To begin with, it is difficult to compare datasets since there is no standard approach for doing so, and also because there aren't any ground-truth labels or publicly available or real-world environment traffic, among other things [2]. Furthermore, only a few statistics reflect the current state of network traffic, which is almost exclusively encrypted for the sake of communication security and privacy. In the proposed system, a dataset is employed that satisfies both the content and the process requirements. The hybrid system for intrusion detection using data approach was introduced in the suggested study. Cybercrime is committed by a malicious node that can be identified by these tools. The goal of this research is to identify the most relevant and useful attributes for inclusion in a new IDS dataset. An approach for producing optimal ensemble IDS is devised in order to meet the goal. Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty SU, Relief-F (R-F), One-R (OR) and Chi Squared are utilised and compared (CS). Techniques that use feature selection produce a list of the features that have been prioritised. For each of the four classification methods, we trained three other models on three different datasets for scanning and DDoS attacks and compared their performance with the proposed approach. In comparison to other trained models, the results of the experiments show that the proposed approach is more effective in preventing and detecting botnet attacks.

Keywords: Intrusion Detection System(IDS)

DOINumber:10.14704/nq.2022.20.8.NQ44733 **NeuroQuantology2022;20(8):7093-7101**

I.Introduction

We now live in a world without boundaries, where nothing is out of reach. Computerization's period has been exposed to new risks and vulnerabilities as a result of rapid technological advancement. According to a recent study, the increasing demand for online services necessitates an increase in cybercrime. In the late 1970s and early 1980s, audit logs were used to monitor user activity for suspicious or malicious conduct, but this has since changed dramatically [1]. It's becoming increasingly difficult to deal with the increasing number of threats and attacks. Because new attack methods are being developed on a daily basis, some system must be in place to keep tabs on them. This new

threat necessitates the development of new forms of defence.

Every month, security professionals uncover more and more security vulnerabilities in the global internet infrastructures due to the ever-increasing interconnectedness of these infrastructures [1]. Vulnerabilities in these systems' security make them attractive targets for hostile actors looking to conduct their activities online [2]. There is a need to build various intrusion detection systems to protect computers and networks from hackers, who may attack the network system and steal or damage financial, medical, or

other vital information from databases [2].The traditional Intrusion Detection System (IDS) usually utilizes precise description such as rules orsignatures tracking of vehicle movement [3] is an important function. The likelihood of a false positive is low when using a signature-based method. Experts must constantly update the database of rules and signatures since new infiltration strategies are being developed and invented every day. Developing adequate signatures for more sophisticated assaults that grow from past attempts can be difficult at times.

Whether it's education, home, transportation, or healthcare, the Internet of Things (IoT) is thriving and becoming a part of our daily life. IoT technology faces numerous issues as the number of connected devices rises, including heterogeneity, scalability, quality of service, and security needs, to name just a few. Because of their size, weight, and expense, security management in the Internet of Things (IoT) is often relegated to the background. The lack of security makes people wary of utilising Internet of Things (IoT) devices, which puts them at risk. There are a number of reasons for this, but the most important is that it renders the Internet of Things (IoT) more vulnerable to cyberattacks [3]. For this reason, it is critical to assess current security risks and consider potential future threats in order to be well prepared to deal with them. Several layers of IoT security have been examined, including the perception, network, support, and application layers; a particular emphasis was placed on Distributed Denial of Service (DDoS) attacks. Since they have the capacity to bring down their victims, DDoS attacks pose a serious threat to cyberspace. An in-depth look at DDoS attack types, IoT device DDoS attacks, the effects of DDoS attacks, and mitigation

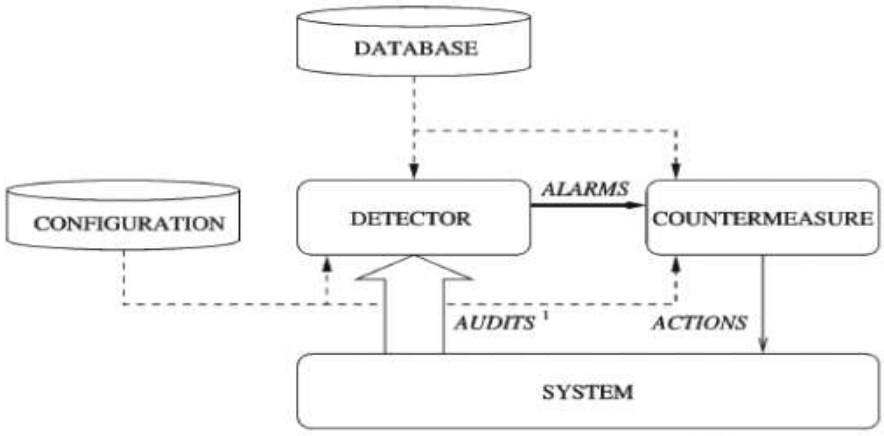


Figure 1 : A simple Intrusion Detection System

strategies is provided. Intrusion Detection and Prevention (IDP) models for mitigating DDoS attacks are the focus of the review work described here. Also presented were classifications for IDS, various anomaly detection methods, various IDS models built on datasets, and various machine learning and deep learning algorithms for data pre-processing and malware detection.. Finally, a wider perspective was envisioned while examining research obstacles, potential solutions and the future of the field.

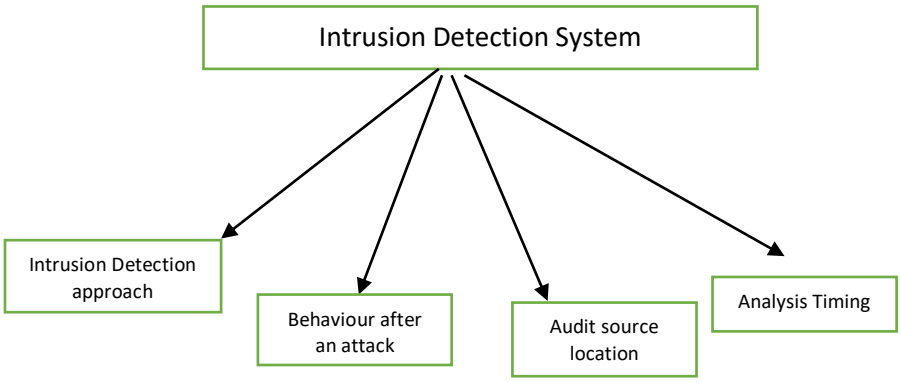


Figure 2: Characterises of IDS

II Literature Survey

An Intrusion Detection System (IDS) based on graph theory is proposed in the publication [4]. EEA (Energy Exhaustion Attack) Resistance and CAMD (Centralized and Active Malicious Node Detection) make up the hybrid IDS (DPER). The genetic algorithm-based data collection technique incorporates CAMD. CAMD identifies malicious nodes exploited by cyber criminals and gives digital data for forensics purposes.. In order to lessen the impact of EEA assaults, a set of communication protocols includes DPER. To test this hybrid IDS' detection and tracing accuracy without negatively impacting energy efficiency, the author used the NS-3 platform for simulation. In addition, the impact of EEA assaults carried on by cyber thieves was effectively diminished.

In article [5] aims to identify the most important aspects that can be employed in the creation of a new IDS dataset by analysing relevant elements. An approach for producing optimal ensemble IDS is devised in order to meet the goal. Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty SU, Relief-F (R-F), One-R (OR) and Chi Squared are utilised and compared (CS). Techniques that use feature selection produce a list of the features that have been prioritised. Four classification methods, namely Bayesian Network (BN), Nave Bayesian (NB), Decision Tree: J48 and SOM, will be used to categorise attacks based on the best selected number of features from the feature ranking phase for each feature selection methodology,

Including To build ensemble IDSs, the best features from each feature selection method are mixed with each classifier method. As a last step, we test the ensemble IDS with validation methods such as K-fold hold-up and F-measure, as well as statistical validation methods. Using Weka's tools on the ITD-UTM dataset, the best ensemble IDSs using (SU and BN), using (CS and BN) or (CS and SOM) or (IG and NB), and using (OR and BN) with respective ten, four, and seven best selected features achieved 81.0316 percent, 85.2593 percent, and 80.8625 percent of accuracy, respectively. Using ten and six best features, respectively, ensemble IDSs based on (SU and BN) or (OR and J48) with a 0.853 and 0.830 F-measure value perform the best, respectively. We'll look at indirect comparisons with other ensemble IDS on various datasets in this article.

An in-depth assessment and classification of deep learning-based intrusion detection systems is presented in paper [6]. IDS architecture and several deep learning algorithms are first introduced in this section. According on the sort of deep learning

approach used, these schemes are categorised. Deep learning networks are used to accurately identify intrusions in the intrusion detection process. A comprehensive review of the IDS frameworks explored is presented, as well as concluding observations and future directions.

III Proposed Work:

This paper designs an information and energy-related IDS with hybrid mechanism for networkapplications with a path-constrained mobile sink. The hybrid IDS provides a trace-back mechanism for network forensics and enhances the network safety.

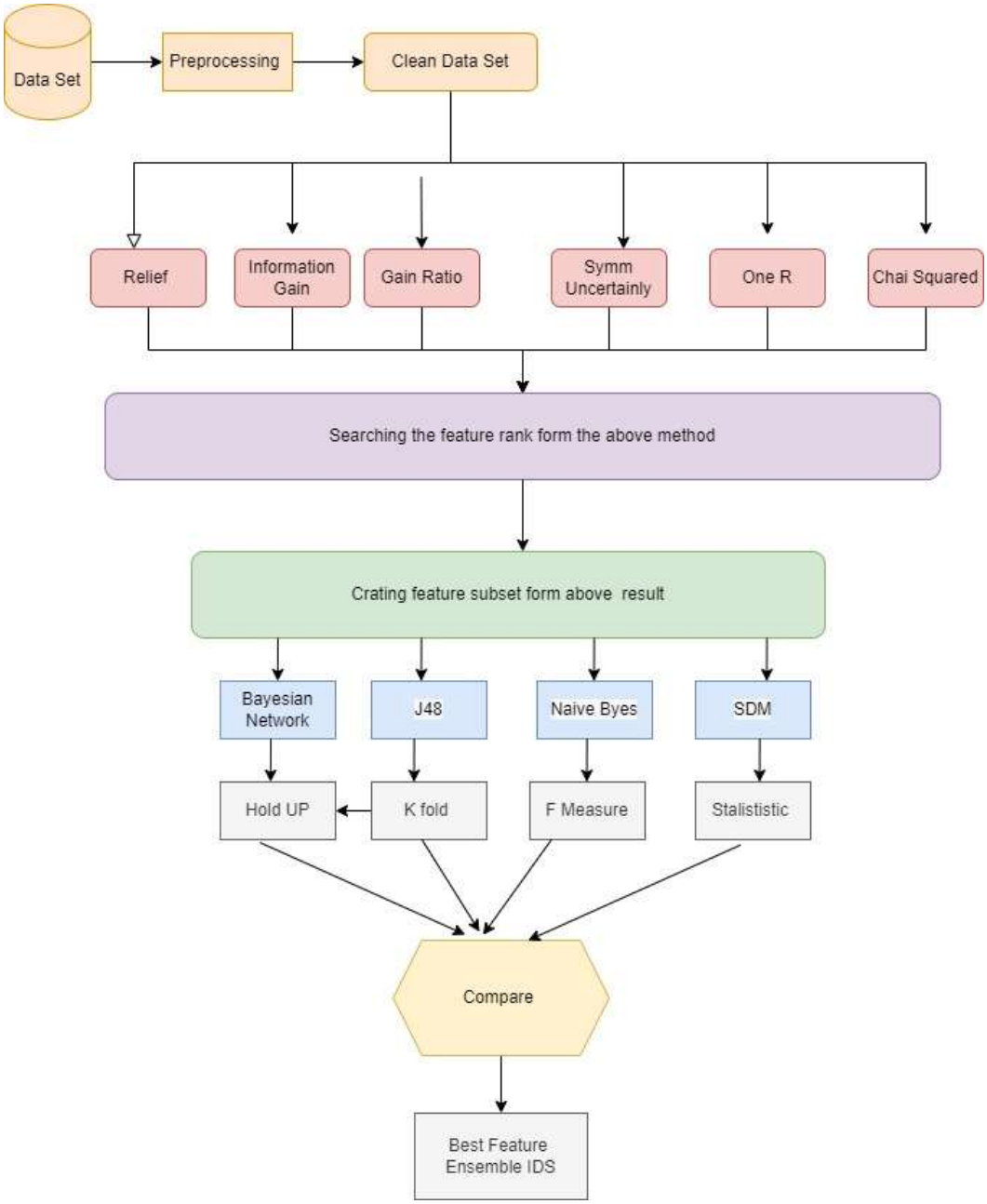


Figure 3: Flow Diagram of proposed Architecture

A. Scanning Attack Detection

To prevent and identify botnet assaults in networks, we developed a revolutionary two-fold machine learning approach. We used a state-of-the-art deep learning model, ResNet-18 [8], to train a pre-attack scanning model to safeguard the network against botnet attacks in our proposed strategy. After that, we trained another ResNet-18 [8] model to detect DDoS attacks perpetrated by intruders who have gained access to the networks protected by inadequate security measures.

1) Data Collection
2) Data Pre-processing
3) Feature Selection
4) Training Model
5) Attack Detection

- **DATA COLLECTION**

The collecting of data is the first step in the suggested scanning attack detection methodology. We began by analysing some of the most prevalent strategies and approaches used by attackers to get access to the IoT network and devices.

- **DATA PRE-PROCESSING**

Pre-processing is necessary after obtaining the scanner traffic. In the first phase, we used the CIC Flowmeter Tool to extract features from the captured.pcap _les of the Scan Lab dataset. More than 60 flow features are extracted for each flow based on a positive ve-tuple of source IP, destination IP, source port, destination port, and protocol in the CIC Flowmeter tool. By using CICFlowmeter, we were able to get the specifics of these properties. It generates a.csv file containing the flow data from a given.pcap file using the CICFlowmeter software package. Unlabeled data was generated as a result. For this reason, we labelled the IP addresses that were used for scanning with the suffix "Scan" whereas all other network traffic was labelled "normal."

- **Feature Selection**

It's now time to narrow down the important features collected from all. pcap les to those that can enable a machine learning model discriminate between normal and scan data. We chose the LR method for feature selection since previous studies have shown it to be more effective [6], [8]. Our initial step was to use LR algorithm to identify the 20 most important characteristics from each dataset, including ScanLab and three others. Afterwards, we carried out a frequency analysis, similar to the one described in [6], on the ScanLab dataset and all three selected datasets using the features picked by the LR method. We identified 15 of the most commonly picked features by the LR algorithm and designated them as features set 1 (FS-1) based on a frequency analysis (Figure 2 and Table 3) of the features selected by the algorithm. To aid in scanning attack detection, the 15 features indicated in Table 3 were chosen from each dataset.

- **TRAINING ML MODEL FOR SCAN DETECTION**

Following the selection of the most useful characteristics, we divided each dataset into three subsets: train, validation, and test. In order to minimise overfitting, we randomly picked 60% of the data for training, 20% for validation, and 20% for testing, so that the ML model could be trained effectively. During the training phase, both the training and validation sets are utilised. The machine learning model is trained on the training data. After each epoch, we validate the trained model on the validation set, and the optimizer method adjusts the weights of the ML model depending on the results. It's finally time to put the ML model through its paces and see how well it performs on an entirely new dataset, or test set. The ResNet-18 [23] model was first trained on the Scan Lab dataset train set, as previously indicated.

- **ATTACK Detection**

Each dataset was broken into three subsets: training, validation, and testing. The ML model could be trained more successfully since we randomly selected 60% of the data for training, 20% for validation, and 20% for testing. Both the training and validation sets

are used in the training phase. The training data is used to train the machine learning model. Iteratively, the weights of ML models are adjusted using the optimizer method based on the results of each epoch's validation set. Once the ML model has been trained, it's time to put it to the test on a completely fresh dataset, or test set. As previously stated, the ResNet-18 [23] model was first trained using the ScanLab dataset.

B. Performance Measure Indices

Some of the predominantly used performance metrics for Intrusion Detection Systems are discussed below.

- **CONFUSION MATRIX**

In and of itself, the Confusion Matrix (CM) isn't a performance metric in the traditional sense. Even so, it's one of the most intuitive metrics for determining the quality of a classification model. CM parameters are used to calculate the vast majority of performance indicators. There are two strategies to reduce errors in the Confusion Matrix described in table 1: lowering false negatives and minimising false positives. There is no one-size-fits-all answer to this question, and it varies depending on the situation. False Positives and False Negatives should be minimised in email spam classification and cancer patient classification, respectively.

Table 1: confusion Matrix

	FALSE	TRUE
FALSE	TRUE NEGATIVE(TN)	FALSE POSITIVE(FP)
TRUE	FALSE NEGATIVE(FN)	TRUE POSITIVE(TP)

- **ACCURACY**

Accuracy is defined as the number of correct predictions over total predictions. This metric is ideal for use in the case of a balanced dataset. When there is a majority class in a dataset, the results provided by this metric may not reflect the model's actual performance.

Accuracy = TP + TN /TP + TN + FN + FP . . . (1)

- **PRECISION**

Precision is a measure to calculate the Machine Learning Model's accuracy in finding the number of actual positives out of total predicted positives. This metric is useful when False Positive is of high cost for Model quality, for example, email Spam Detection Model.

IV Simulation and Result

We utilised a Jupiter notebook for both the simulation and the final product. The dataset is simulated on Jupiter notebook using python and associated libraries. As previously stated,

Precision = TP/ TP + FP . . . (2)

- **RECALL/SENSITIVITY**

Recall is a measure to calculate the Machine Learning Model's accuracy in finding the number of positives out of total actual positives. This metrics is useful when False Negative is of high cost for Model quality, for example, Fraud Detection Model. Recall = TP /TP + FN . . . (3)

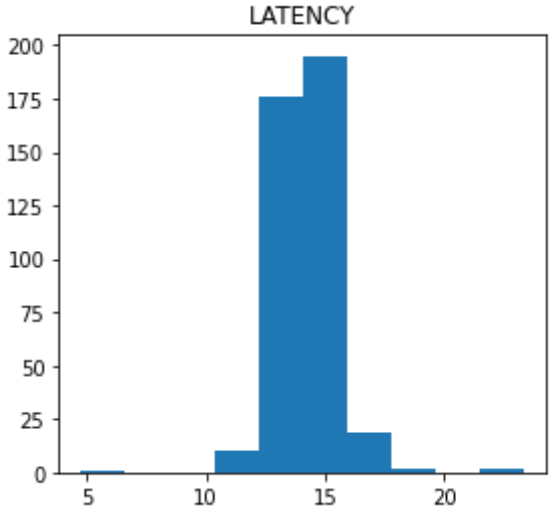
- **F-1 SCORE**

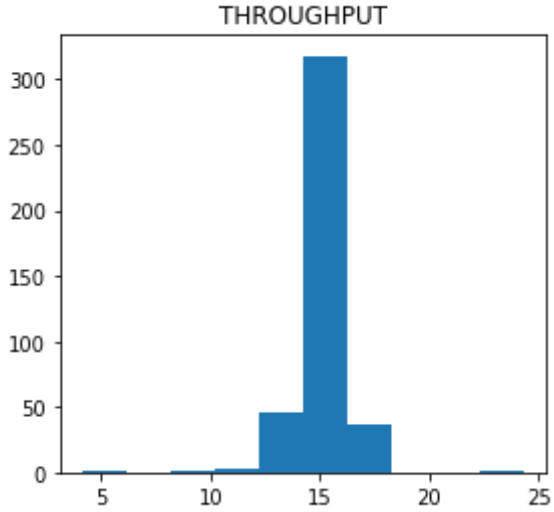
It is calculated as a Harmonic Mean of precision and recall metrics to better evaluate model performance. This is a metric of importance for an imbalanced dataset as in this; equal importance is given to both Precision and Recall. F-1 Score = 2 × Precision × Recall Precision + Recall . . . (4)

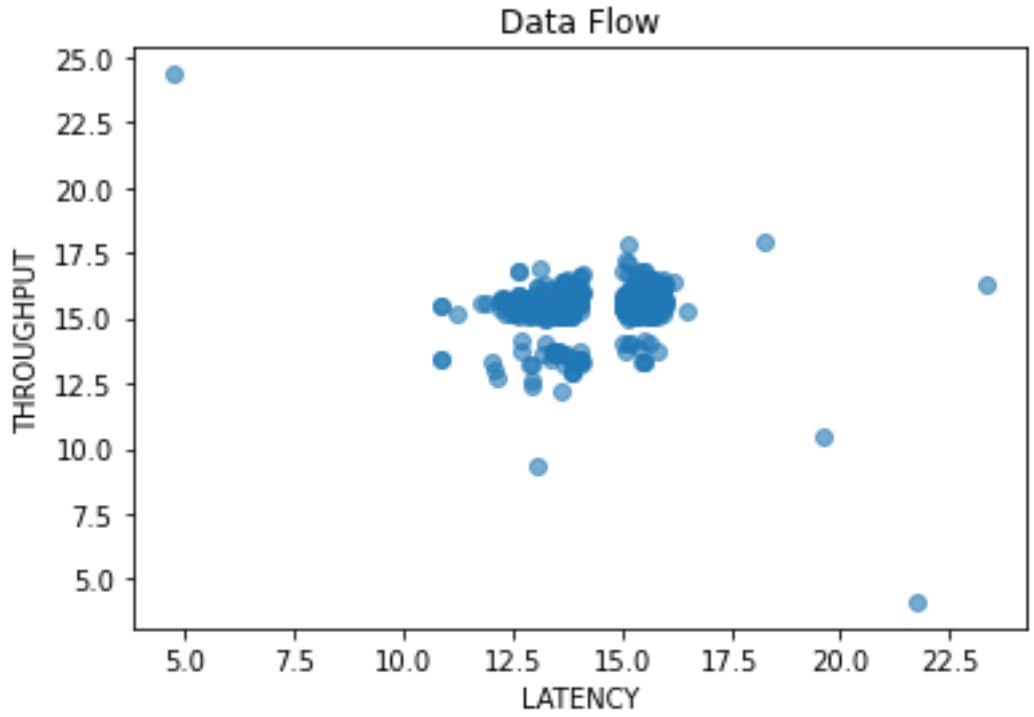
the scanning and DDoS assault methodologies were initially analysed in this study. Using three distinct network traffic generating programmes, namely Nmap, Hping3 and

Dmitry, we created 33 different forms of scanning assaults traffic and 60 different types of DDoS attacks traffic based on our findings. On a Core i7 PC that had 8 GB of RAM and Ubuntu-18 operating system loaded, these utilities were fully set up and running. Wireshark was used to capture the produced network traffic in pcap format. After that, we extracted features from the.pcap files and performed labelling based on the IP addresses of the machines that participated in the experiment. Step 2 of our proposed methodology involved applying feature selection algorithms and splitting the dataset into a train and test set.

Model Name	Trained and Test Over	Accuracy Proposed Model	Hussain Etal	Precision		Recall		F1 Score	
MReNetaScan-1	ScanLAb	99.30	99.20	99.45	99.39	99.30	99.05	99.25	99.22
MResNetScan2	CICDS-19	99.92	99.91	99	100	99.80	99.83	99.90	99.22
MResnetScan3	CICDS-17	99.80	99.79	99.90	99.95	99.80	99.67	98.90	99.81
MResnetScan4	Bot	99.90	98.85	98.90	98.95	98.74	98.66	98.90	98.81







V Conclusion

To prevent and identify botnet assaults, we developed a hybrid machine learning technique. An advanced deep learning model called ResNet-18 was used to train the Modified ResNetScan-1 scanning attack detection model in the Proposed study. ResNetScan and ResNetDDoS models were trained using publically available datasets, and the results of these experiments were used to verify the validity of the Modified ResNet-18 model's effectiveness in detecting DDoS attacks and detecting scans, respectively. ResNetScan and ResNetDDoS models were then evaluated against the test set of other datasets that they had not been trained on. Except for the suggested ResNetScan-1 and ResNetDDoS-1 models, all ResNetScan and ResNetDDoS models performed much worse on datasets on which they had not been trained, according to the findings of the experiments. For scanning and DDoS attacks, the experimental findings showed that all other models were outperformed by the proposed Modified ResNetScan-1 and ResNetDDoS-1 models..

Reference

[1] Mohamed, A.B., Idris, N.B., Shanmugum, B. (2012). A Brief Introduction to Intrusion Detection System. In: Ponnambalam, S.G., Parkkinen, J., Ramanathan, K.C. (eds) Trends in Intelligent Robotics, Automation, and Manufacturing. IRAM 2012. Communications in Computer and Information Science, vol 330. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35197-6_29

[2] J. Lansky et al., "Deep Learning-Based Intrusion Detection Systems: A Systematic Review," in IEEE Access, vol. 9, pp. 101574-101599, 2021, doi: 10.1109/ACCESS.2021.3097247.

[3] N. Mishra and S. Pandya, "Internet of Things Applications, Security Challenges, Attacks, Intrusion Detection, and Future Visions: A Systematic Review," in IEEE Access, vol. 9, pp. 59353-59377,2021,doi: 10.1109/ACCESS.2021.3073408.

[4] Wu, Chao,et al. "A Hybrid Intrusion Detection System for IoT Applications with Constrained Resources." IJDCF vol.12, no.1 2020: pp.109-130. http://doi.org/10.4018/IJDCF.2020010106

[5]D.Stiawan et al., "An Approach for Optimizing Ensemble Intrusion Detection Systems," in IEEE Access, vol. 9, pp. 6930-6947, 2021, doi: 10.1109/ACCESS.2020.3046246.

[6] J. Lansky et al., "Deep Learning-Based Intrusion Detection Systems: A Systematic Review," in IEEE Access, vol. 9, pp. 101574-101599,2021,doi: 10.1109/ACCESS.2021.3097247.

[7] Adel Binbusayyis, ThavavelVaiyapuri, “Comprehensive analysis and recommendation of feature evaluation measures for intrusion detection”, Heliyon, Volume 6, Issue 7, 2020, e04262,ISSN-2405-8440, https://doi.org/10.1016/j.heliyon.2020.e04262

[8]F. Hussain et al., "A Two-Fold Machine Learning Approach to Prevent and Detect IoT Botnet Attacks," in IEEE Access, vol. 9, pp. 163412-163430, 2021