EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Comparison of Classification Model for the Detection of Cyber-attack using Ensemble Learning Models

Muhammad Shoaib Akhtar, Tao Feng*

School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

13.cs.194@gmail.com fengt@lut.com

Abstract

Incorporating digital technologies into security systems is a positive development. It's time for the digital system to be appropriately protected from potential threats and attacks. An intrusion detection system can identify both external and internal anomalies in the network. There are a variety of threats out there, both active and passive. If these dangers aren't addressed, attacks and data theft could occur from the point of origin all the way to the point of destination. Machine learning is still in its infancy, despite its wide range of applications. It is possible to predict the future by using machine learning. A cyber-attack detection system is depicted in this study using machine learning models. Machine learning algorithms were trained to predict cyber-attack scores using data from prior cyber-attacks on an open source website. In order to detect an attack at its earliest possible stage, this research also examined multiple linear machine learning algorithm-based categorization models. Classifiers' accuracy is also compared in the presentation, as is the presentation itself. Balance procedures were followed. Radio Frequency and GBC have the best accuracy, at 87.93%, followed by ABC at 86.11%, BT at 81.03%, ET at 70.31%, and DT at 70.31 percent (84.48 percent).

Keywords: cyberattack, machine learning, ensemble learning.

Received on 05 August 2021, accepted on 11 January 2022, published on 01 February 2022

doi: 10.4108/eai.1-2-2022.173293

*Corresponding author. Email: fengt@lut.com

1. Introduction

Recently, it has become increasingly popular, with numerous applications in industries like fault tolerance, financial and economic crisis detection, diagnostics of diseases and conditions in the earth sciences and meteorology, and the detection of unusual celestial objects in astronomy or astroparticles in cyber-security and so on. [1], [2] Algorithms for anomaly detection are utilised when it is difficult to discover patterns that deviate from the norm. [6] Almost all methods can be categorised by their concept of normality or their method for identifying deviations in the body of knowledge in the literature. It is proposed in this study that anomaly detection be investigated in depth before a more recent comparison of the two methods [7].

The difficulty in generalising traditional machine learning algorithms to new settings causes deep learning algorithms to take longer to develop and may become obsolete in the blink of an eye if they are not developed quickly. Even though deep neural networks and technology can deal with time-consuming and unpredictable attacks in the future, there is still an issue with deep-school gradient disappearance or explosion.

As a result of the difficulty in generalising standard machine learning algorithms to new contexts, deep learning algorithms take longer to develop. They may become obsolete in a flash of invention [8]. A problem with deep-school gradient disappearance or explosion persists even though deep neural network [9] and [10] technology can deal with time-consuming and unpredictable attacks in the future [11]. A deep residual neural network [12–14]. The following are the study's primary goals:

A strategy based on five ensemble learning models will be devised and put to the test to identify cyber-attacks.

Motivation behind this research is the cyber-attacks that harm the digital networks. An algorithm is used to assign a

cyber-attack score to each model, then used for detection and intelligent decision-making in the following stages. The reliability of Internet of Things devices is assessed using various criteria derived from the cyber-attack score obtained in the previous stage. This study's key contribution is that we assist ensemble learning methods for the first time in cybersecurity attack classification. Contributions of this study are:

a) Usage of Feature scaling methods which help machine learning model to classify more accurately than without feature engineering.
b) Implementation and comparative analysis of all tree based algorithms and hyper tuning of all models to optimize the model.

2. Literature Review

Detection system strategies are discussed in this section of the document. There are now supervised and unsupervised learning options available.

A rise in the sophistication and severity of cyber-attacks has prompted security experts to adopt a wide range of machine learning techniques to protect the data and reputation of their clients. IDS or intrusion detection systems are increasingly using deep learning techniques to improve their ability to protect computer networks and hosts. Deep learning-based intrusion detection systems are examined in detail in this review article [8] and a categorization is made. An overview of IDS architecture and many deep learning techniques is provided first. It then categorises these schemes based on the type of deep learning algorithms used in each. Deep learning networks are used to accurately identify intrusions in the intrusion detection process. Conclusions and future directions are made after a detailed examination of the frameworks studied in this study [13].

To avoid the model becoming sensitive to big samples while remaining insensitive to small samples, Hu et al. [15] utilise the ADASYN technique. An improved CNN has been created using the split convolution module (SPC-CNN), which increases feature variety while simultaneously decreasing the impact of redundant interchange information on model training. Finally, an AS-CNN model that combines ADASYN and SPC-CNN is applied for intrusion detection. The AS-CNN algorithm is tested on the NSL-KDD data set as a final step. There is a 4.60 percent and 2.79 percent increase in the DR (Detection Rate) compared to the standard CNN and RNN models, respectively, according to the simulation results. FAR was also lowered by 15.58 percent and 14.57 percent compared to the two other models considered for this study [16].

For this purpose, Yang et al. [6] conducted a small-scale attack simulation to evaluate how the network would respond to an intruder. When compared to existing methods, our paper's simulations show that the suggested method has a greater detection accuracy and true positive rate and lower false positive rates. LeNet-5 and DBN have detection accuracy of 8.82 percent and 0.51 percent, respectively, and recall rates of 4.24 percent and 1.16 percent, respectively, when compared to the traditional models, but the false positive rate is lower than the other three types of models (LeNet-5, RNN, and RNN).

Network attacks against the train ECN, such as IP Scan, Port Scan, Denial of Service (DoS), and Man in the Middle (MITM) attacks are defended using a novel ensemble intrusion detection system developed in this research [7]. (MITM). Thirty-four separate protocol contents are extracted from our ECN testbed's raw data, then aggregated into a specific dataset. A data imaging strategy and a temporal sequence generation method will improve the dataset. There are six base classifiers based on several typical convolutional neural networks and recurrent neural networks: LENET- 5, the VGGNet (also known as SimpleRNN), LSTM and GRU (also known as GRU-R). A dynamic weight matrix voting strategy is offered for incorporating all of the primary classifiers. Using data gathered by the authors, this strategy is evaluated. This technique has an excellent capacity to combine the advantages of the base classifiers and achieve better detection performance with an accuracy of 0.975 in the experiments.

Among them was Kim et al. Installation and operation of an Intrusion Detection System powered by Artificial Intelligence (AI-IDS). It is possible to extract the features of real-time HTTP traffic without encrypting, calculating entropy or compressing the data in any way using an ideal convolutional neural network and long-term memory network (CNN-LSTM) model and normalized UTF-8-character encoding. The authors proved the system's superiority by repeating trials on two publicly available datasets (CSIC-2010 and CICIDS2017) and fixed real-time data. By training payloads that examine true or false positives using a labelling tool and then comparing the findings, AI-IDS identifies sophisticated assaults from harmless traffic, such as unknown patterns, encoded or obfuscated attacks. User-defined functions are divided into independent images in a flexible and scalable architecture that uses Docker images. Snort rules for signature-based intrusion detection systems can also benefit from new patterns identified. As a result of continuous training, it is feasible to assess unknown web-attacks more precisely.

This paper uses hybrid sampling and deep hierarchical networks to detect network intrusions. After reducing the number of noise samples in the majority category using OSS, Jiang et al. [17] then use Synthetic Minority Over-sampling Technique to increase the number of minority samples in their analysis (SMOTE). In this method, a balanced dataset can be produced, allowing the model to completely acquire the features of minority samples while considerably lowering the model training time. Convolution neural networks (CNNs) and bidirectional long-term memory (BiLSTMs) extract spatial and temporal data, respectively. It produces a deep hierarchical network model. According to the NSL-KDD and UNSW-NB15 datasets, the proposed network intrusion detection approach has a classification accuracy of 83.58 percent and 77.16 percent, respectively.

An efficient strategy for discriminating between abandoned items, stolen items, and ghost regions in surveillance camera footage is described in this study by Park et al. [18]. Both a dual background model and object segmentation using mask

areas and CNN features (called Mask R-CNN) are used to extract candidate stationary objects from the background model, then used to generate the object mask information. When given a candidate stationary item from a backdrop model, it is checked to verify if an identical segmented object exists in the current video frame or the previous background frame to consider both the present and previous conditions. A comparative analysis technique provided in this paper is used to consider several situations and then apply the results to decide the final state of the candidate stationary object. Results from a qualitative evaluation of a proposed solution to address the discriminating problem have proven positive. Due to the difficulty of installing standard intrusion detection-based security systems in open spaces like convention centres or parks, this technology is expected to be widely employed for applications like automatically detecting stolen or abandoned things.

This study [9] provides an Intrusion Detection System (IDS) for the Internet that is based on Convolutional Neural Networks (CNN). The proposed intrusion detection system (IDS) is intended to identify network intrusions by categorizing all packet traffic into benign and harmful classes. Each packet traffic classification is assigned a numerical value. The dataset CICIDS2017 (Canadian Institute for Cybersecurity Intrusion Detection System) was used to train and test the suggested model, and the results were published in this paper. The model has assessed several parameters, including overall accuracy, detection rate, false alarm rate, and training overhead. In all of these areas, the model is correct by the data. There will be a comparison between the performance of the suggested model and the performance of nine other widely used classification models in this study.

An intrusion detection system is built around a quantitative model of how ports interact, proposed [19]. (IDS). By taking the arrival time distribution of traffic into account, the model provides a quantitative expression of Port Interaction Mode in Data Link Layer (PIMDL). The model's goal is to improve the accuracy and efficiency of intrusion detection by including this information into the model. The model's applicability is demonstrated by using phase space reconstruction and visualization techniques. To mine the differences between normal and abnormal models while considering the characteristics of long and short sessions, an artificial neural network based on CNN and LSTM is being developed. It has resulted in developing a more effective intrusion detection system based on a multi-model scoring mechanism, classifying sessions in model space depending on the information obtained. As a result of these findings, the quantitative model and the upgraded algorithm developed may be used to successfully prevent the hiding of identity information while simultaneously enhancing computing efficiency and the accuracy of small sample anomaly detection (as demonstrated by the experiments). It's time to come up with a new approach that allows the hypervisor to first build believable trust relationships among virtual machines by taking into account the purpose and the personal trust sources and using vectors to aggregate them. It's also worth mentioning the EFPSO (Enhanced Fuzzy Particle Swarm Optimization), an algorithm that uses real-time load distribution among VMs to help the hypervisor better identify DDoS attacks. A virtual machine is assigned to each new request using the EFPSO algorithm, which prioritises incoming client requests based on their workload. In order to detect DDoS attacks as effectively as possible with the limited resources available, the EFPSO algorithm is proposed by [1] equips the hypervisor with the most effective detection load distribution technique across VMs. Finally, a classifier known as a Convex Support Vector Machine (CSVM) is used to prevent further damage. Detection of DDoS attacks, false positives and negatives, and CPU and memory usage are all factors considered.

A low-cost IoT-based DDoS attack is the subject of [2]study, which looks into the possibilities of launching one. New DDoS attack architecture is proposed in the beginning. This design is ideal for resource-constrained DDoS attackers due to its negligible management costs, high undetectability, and excellent robustness. In this architecture, the optimal design of an attack strategy is simplified to a variational problem, where the objective functional stands for the projected expected impact of a DDoS attack associated with a DDoS attack strategy's DDoS assault. Finally, the variational problem for three DDoS defence techniques has been solved. DDoS assaults based on IoT are now more understood thanks to this study.

Blacklisting is a key component of the classical IDS. Time-consuming and ineffective for new invasions, these procedures are not. Today, machine learning and deep learning models help automate and programme IDS to be dynamic. Data utilised to train these models has a significant impact on the model's overall performance. For the most part, IDS research is conducted using datasets like KDD 99 and NSL KDD, which are incompatible and out-of-date. DDoS attack dataset research is extremely scarce. The purpose of [3]is to look at the impact of already existing IoT datasets. For DNS amplification attacks in IoT, author then propose a real-time data gathering mechanism Port mirroring captures DDoS attack-generated network traffic.

Aski et al [4]purposed a new method to developing a security framework for connected healthcare systems based on edge computing is introduced. The heart of the proposed system is an effective multi-factor access control and ownership transfer mechanism for future healthcare applications based on edge computing. Clustering techniques that evaluate and aggregate large volumes of data from heterogeneous devices independently before they are sent to the cloud are used to achieve data scalability. It is a first-of-its-kind data/device ownership transfer arrangement, as well. Patients' medical records and medical device ownership rights can be transferred from one registered user to another during this phase. There are many common IoT attacks like as insider attacks, distributed denial of service (DDoS) attacks, and traceability attacks that can be avoided by performing a formal and informal security analysis.

The Internet of Things would dramatically speed up diagnosing and monitoring patients, as small IP-based wireless sensors on the patient's body can monitor his physiological data, such as blood pressure and heart rate,

remotely and continually. Patients' medical records must be kept private at all times in this scenario. Only caregivers and authorised individuals should have access to these records. Throughout the healthcare application scenario, security must be assured. A security model is proposed by [5] as a solution to the potential security issues in this application. Symmetric cryptography is used in author's model, which includes a proposed key management system and a technique for authenticating network nodes.

Insider-assisted DDoS assaults can be mitigated using an EDIP (Early Detection and Isolation Policy). When an insider is found among all valid customers, EDIP migrates it to an attack proxy so that other legitimate clients can no longer access the system. This is further supported by [6]by developing an effective algorithm that aims to maximise attack isolation while reducing disturbance to non-malicious clients. Using load balancing also helps to keep proxies from becoming overburdened.

For healthcare monitoring systems using IP VPNs over optical transport networks, an attack detection model is being developed by [7]. Here, author describes the vulnerabilities in healthcare monitoring systems networks that use VPNs over optical transport layer architecture. The IP and optical layers are further integrated into multi-layer network architecture, and a DoS attack detection programme is introduced. For remote healthcare control devices that have limited processing and memory capabilities, the proposed application is a lightweight .implementation. Toward this end, a tutorial-oriented method to attack detection is provided, which can also be used by nonprofessionals for practical and educational purposes.

Using key elements that aid in understanding how these attacks can be carried out, [8]aims to study the complete spectrum of application layer DDoS attacks. There is also a discussion of defence systems against different types of attacks, with special emphasis on aspects that aid in detecting distinct types of attacks. Discussions such as these can assist researchers understand why a certain set of attributes is effective in identifying certain threats.

Kamble et al [9]provides an in-depth look at recent developments in Reinforcement Learning Algorithms and Machine Learning, both of which can be used to improve IoT security. Wearable sensors and LoRaWAN are part of the proposed system, which collects user data and connects to the Internet. These monitor the patient's heart rate and body temperature, which are essential indicators of health. Raw sensor signals can be processed by a local server, which can also display health information and send out alerts in the event of an emergency being discovered. IoT cloud server deployment includes features such as web monitoring and mobile apps. The design makes use of data categorization and key management services for security purposes. Authentication, detection and mitigation of DDoS and Man-in-the-Middle attacks, as well as intrusion detection and malware analysis were all taken into account when implementing encryption and reinforcing learning for IoT security and privacy. Author plan to investigate a wide range of security threats and solutions in the future.

According to[10],DDoS attacks can be identified using two different tactics. Consider the severity of the attack when determining whether or not you are being targeted by a DDoS attack. Second, to identify the DDoS attack, an improved KNN algorithm based on Machine Learning (ML) is utilised in conjunction with ML. According to theoretical and empirical evidence, DDoS attacks can be detected more effectively using author's proposed methods than utilising other methods now available.

Deep transfer learning is used in this study to develop a small-sample DDoS assault detection system. The first step is to employ deep learning techniques to train many neural networks that can be deployed in DDoS assaults with enough samples. Transferability metric is then developed to allow us to compare the transfer performance of various networks. This measure can be used to identify the best network out of the four available. For a limited sample of DDoS attacks, this study found that the detection performance dropped from 99.28% to 67% when using the deep learning detection technique. In the end, deep transfer of the 8LANN network to the target domain improved detection performance by 20.8%. As demonstrated by [11], deep transfer learning can significantly increase the effectiveness of deep learning techniques for small sample DDoS attack detection, as demonstrated in the experiment.

As part of this study's approach for visualising network traffic, an image-based CNN model, referred to as ResNet, was trained on the translated traffic data and then tested. Specifically, in the case of binary classification, the proposed methodology demonstrated 99.999 percent accuracy in detecting DoS and DDoS attacks. [12]proposed methodology was successful in identifying a total of eleven separate DoS and DDoS attack patterns, which is 9 percent more accurate than the existing state of the art.

For the classification of network traffic into benign and DDoS assault traffic, [13]provides an innovative architectural design that incorporates an AutoEncoder (AE) with a Deep Neural Network (DNN) that is stacked well for feature learning. In order to detect DDoS assaults, the parameters of AE and DNN are fine-tuned using specially devised techniques. Reconstruction errors are reduced, gradients do not explode or vanish, the network is smaller, and overfitting is avoided by the changes proposed in this article. Using performance criteria such as detection accuracy, precision, recall, and F1-Score, the suggested approach was compared to ten other leading-edge approaches. Data from the CICIDS2017 and NSL-KDD standards were used in the validation process. The proposed method outperforms current approaches on the NSL-KDD dataset and produces competitive results on the CICIDS2017 dataset.

The ability to continuously monitor a patient's health is what makes smart health such an important and current topic for researchers and practitioners alike. The goal of smart health is to give patients with access to medical care at any time and from any location. Because most smart health monitoring systems rely on wireless networks, they are at risk of attack from outside sources. Health monitoring applications and systems are vulnerable to a variety of assaults. DoS assaults, Fingerprint and Timing-based snooping, Router Attacks, Select and Forwarding Attacks, Sensor Attacks and Replay Attacks are a few of the many types of cyber-attacks that can be used. Here, [14]examine how these attacks affect health monitoring systems, and

author offer some recommendations based on author's research.

Divergence measures such as the LeCam divergence measure introduced in this research can be used to detect various sorts of distributed denial-of-service (DDoS) attacks. The technique suggested by [15] is demonstrated on the DDoSTB, MIT Lincoln, and CAIDA datasets, all of which are widely used in industry. The novel LeCam Divergence metric outperforms the more traditional Kullbeck-Leibler, Bhattacharyya, and Pearson Divergence metrics, as well as the more recent Kullbeck-Leibler and Pearson Divergence metrics.

Various studies have been carried out to date in order to identify and determine the most appropriate requirements for a defensive solution that helps safeguard online applications against HTTP-based DoS and DDoS attacks. There are still some gaps in author's understanding of what makes a protective solution effective, and so [16]aims to fill in those gaps and identify the missing specifications. In order to determine and define the ideal specifications for a protective framework against HTTP-based DoS and DDoS assaults, this article conducts a detailed survey of all types of HTTP-based DoS and DDoS attacks.

The Internet of Things (IoT) is one of today's most prominent developing technologies, and it has been employed in a variety of ways to make everyday tasks easier for humans. Internet of Things (IoT) devices is multiplying at a rate that no one could have predicted. Many security experts are concerned about the enormous number of hazardous vulnerabilities in these devices. IoT devices being used to launch DDoS assaults is one of these issues. On the basis of projections for the amount of IoT devices in use by 2020, [17]examines the size of DDoS assaults involving IoT devices and some of its effects.

Ukraine will gradually implement an extensive eHealth system over the following few years. There are many exciting prospects for the future because to this system's efficiency. The deployment of such systems, however, presents a number of challenges. The cybersecurity provision is one of the most common sources of contention. Security in cyberspace is one of the most pressing issues in today's modern society. The ability to quickly identify and respond to computer network threats is critical to the successful operation of a wide range of industries. Using the Neyman-Pearson criterion and a fixed sample size, a detection strategy for distributed denial-of-service (DDoS) assaults is synthesised by [18]. The experimental investigation of traffic usage in the absence and presence of a DDoS assault was a necessity for the development of such a process. Experiment and computer simulation both support the validity of the proposed technique.

There have been severe security issues with computer networks due to the expanding usage of technology, and the Internet of Things (IoT) is no exception. E-Health services, for example, are vulnerable to the same vulnerabilities as other IoT-based services. Attacks on E-Health servers in the Internet of Things, such as Denial of Service (DoS) and Distributed Denial of Service (DDoS), would put patients' real-time monitoring and the entire reliability of E-Health services at risk. It has been discussed by [19]how to protect the servers against DoS/DDoS assaults in IoT, and a solid solution has been offered.

Exponential Smoothing is used to forecast network traffic in the future, and the time series of prediction error is then generated based on the difference between the projected and actual network traffic. Attack traffic causes this time series to become chaotic, as demonstrated by [20]. Lyapunov exponent analysis on the projected time series is used to detect the DDoS attack and predict it using a Recurrent Neural Echo State Network (SCESN). The Darpa98 dataset, a typical dataset for intrusion detection system evaluation, is used to test LEAESN's technique. LEAESN's capacity to forecast DDoS attacks is adequate.

IoT-Flock, a new open-source IoT data generator, is part of the proposed architecture. Researchers can utilise the IoT-Flock tool to create an IoT use-case that includes both legitimate and malicious IoT devices and generate traffic as a result of their work. In addition, an open-source tool is included in the proposed framework for turning the IoT-Flock traffic recorded into an IoT dataset. First, [21]created an IoT healthcare dataset that includes both normal traffic and IoT attack traffic using the approach author provided in this study. Using the created dataset, author then utilised various machine learning techniques in order to detect cyber-attacks and safeguard the healthcare system against them. Author hope author's framework will aid in the development of context-aware IoT security solutions for sensitive use cases, such as healthcare.

Meng et al [22]propose a security enforcement architecture based on SDN for smart healthcare data sharing platforms. It is possible to offer group data services to authorised service consumers or IoT devices in author's architecture, where each virtual machine provides a dedicated virtual machine for each patient. Additional protection is provided by the SDN-based gateway, which provides a firewall mechanism and ensures that only authorised entities can access the patient's virtual machine. Author's platform can verify resource-constrained IoT devices and combat issues caused by identity theft because each object has a unique MAC address. Using POX controller and Mininet emulator, author builds an experimental system to verify the framework's effectiveness and viability. Tests carried out in various environments have shown that author's framework works. The framework can still work well and perform adequately if the information flow model is expanded to a larger size.

DDoS assaults on numerous networks are detected more accurately using a new collaborative source-side DDoS attack detection method that considers the detecting performance in different time zones. Based on the detection rate and false positive rate of each network, the results of each attack detection are weighted. One way to tell if DDoS attacks have occurred is to gather the weighted detection findings. The strategy proposed by [23]decreases false positives by 35% while retaining a high detection rate after thorough testing with real network traffic data.

More than a billion people have been infected by the COVID-19 pandemic this year alone. If everyone keep their distance from each other, they should all be safe at this time.

As a result, online technologies have seen an increase in utilisation, but so have the risks associated with them, such as cyber attacks. A DDoS attack, the most common and deadliest of them all, disables a website's users' access to it. This syudy outlines a filtering strategy that is effective in detecting a DDoS assault in the COVID-19 scenario. DDoS attack traffic can be detected using statistics such as packet score and entropy variation, on which author's approach is based. Using Omnet?, author have tested the efficiency of author's suggested solution using a variety of test cases. 96 percent of the time, author were able to accurately recognise DDoS assault traffic and separate it from the rest of the flash crowd using the approach [24] have provided.
In most cases, a distributed denial-of-service (DDoS) assault will cause a targeted machine, service, or network to become unusable. DDoS attacks are still a severe danger to the security of cyberspace, despite massive efforts to counteract them in the previous decade. Detection and defence against DDoS assaults are the focus of [25]in this study. The following three important DDoS attack-related topics will be the focus of this presentation. Attack tracing, detecting low-rate DDoS attacks, and telling DDoS attacks apart from flash crowds are all aspects of DDoS defence.

3. Methodology

This section explains the research methods used in this project. [17] [18] [20] Data acquisition, raw data processing, data cleaning stages, data pre-processing (controlling imbalance in the dataset and handling the outliers), feature engineering, model development (machine learning algorithm based), and performance evaluation of produced models are all part of this process. [14] [19] See Figure below for a visual representation of how this research was conducted.

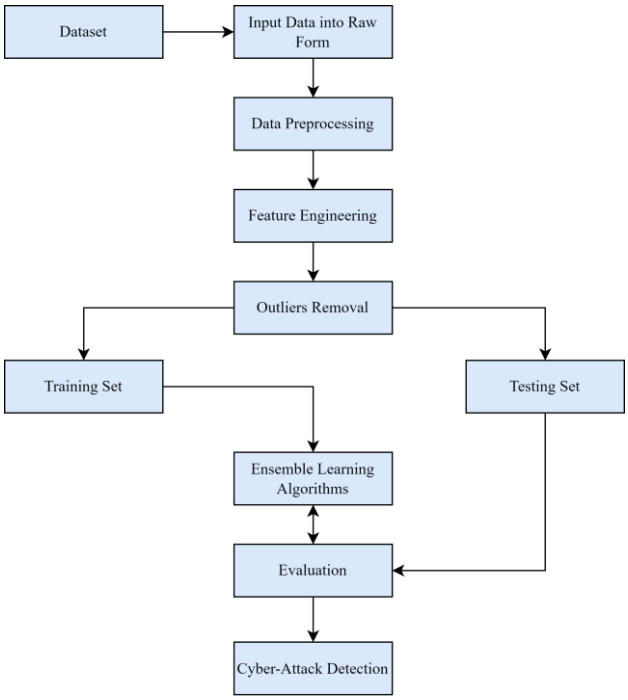

Figure 1. Proposed Flow Work

A. Dataset

Kaggle is the source of this information (an online data source). There are several independent variables in the dataset and a single dependent one (Outcome).

Table 1. Dataset information

Attributes	Values	Description
IDs	Any integer number	IoT cyber-attack detection relies on the use of IDs, which are the numbers of devices connected to IoT.
Flags	Any integer number	A flag is a signal sent to an analogue device to detect suspicious activity in the system.
Number of Packets	Any integer number	In computing, a packet is the amount of data that is transmitted or received.
Sources	Any integer number	If the data supplied and received are not the same, these sources will be able to tell you exactly what happened.
Destinations	Any integer number	The devices that collect and receive data are referred to as "destinations."
Protocols	Any integer number	Transmission and reception protocols are shown in the form of a series of integers.
Attacks Type	DDOS = 1 No = 0	In the dataset, DDoS assaults have been identified. Flag 1 is activated when an attack is imminent, whereas Flag 0 is activated when there is no attack.

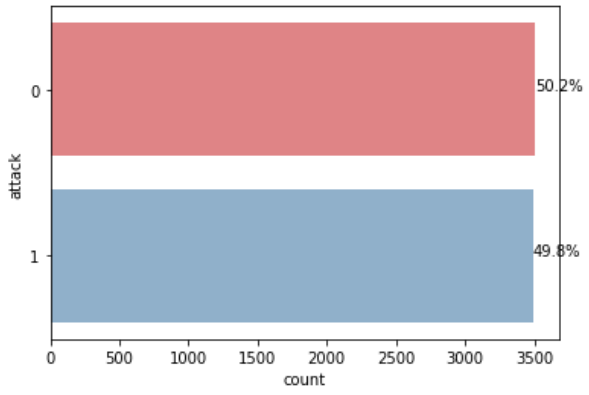

Figure 2. Relation between count and outcomes

The graph displays the total number of targets (0 or 1). As long as an assault has not been identified, the counter is set to zero. You've got the raw information. After that, the data was cleaned up using additional methods, such as deleting duplicates or null values. In data mining, this technique is used to turn raw data into a format that can be analysed. There are times when data from the real world is

incongruent, inconsistent, or missing. The following are a few examples of pre-processing methods: Skewed classification is a major problem for predictive modelling. Typically, the number of samples provided for each class is the same in classification machine learning approaches. For minorities, this leads to erroneous models. As a result, minorities are more vulnerable to classification errors than the majority. [20] This is a concern. In order to ensure that this investigation is correctly balanced, the outliers have been eliminated from the data set since this study was conducted, resampling methods have improved dramatically. [18] [14] Extracting data from each cluster and under-sampling is an example of how to preserve information. By oversampling rather than making exact replicas of minority class data, we can generate more varied synthetic samples.

B. Feature Engineering

Data from a certain domain is utilized to construct learning machines' functions. Raw data is transformed into machine-learning forms by removing non-essential features. This study makes use of a correlation matrix to figure out how different variables are related to one another. A covariance matrix is a correlation matrix. Linear association's power is summarized by correlation. Correlation is a way to summarize the relationship between two quantitative variables in a straight line. Values r denotes the range of input values: -1 to +1.

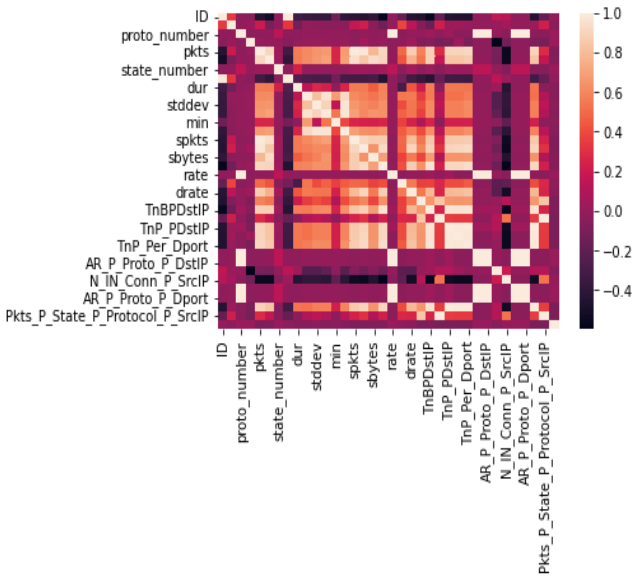

Figure 3. Correlation Matrices IoT based dataset

C. Ensemble Learning Algorithms

An ensemble learning model is a hyper-tuned tree model with optimized hyper parameters and many trees that perform better than simple trees. It takes a lot of trial and error to build a machine learning model. In addition to improving performance, Hyper tuning reduces the complexity of the model simultaneously.

D. Gradient Boosting Classifier

Gradient boosting classifiers are a group of machine learning approaches that integrate weak models into a powerful prediction model. Decision trees are widely used in gradient boosting.

E. Random Forests

A machine learning technique known as a random forest can tackle regression and classification problems. Ensemble learning is a technique for combining many classifiers to solve complex problems. A random forest method employs a large number of decision trees.

F. Adaboost Classifier

AdaBoost classifiers are meta-estimators that begin by fitting a classifier on the original dataset and then fitting consecutive copies of the classifier using a weight modification for poorly categorized instances. Future classifiers focus more on challenging scenarios.

G. Bagging Trees

Combining predictions from various base classifiers (either by voting or average) on random subset sets of the original dataset is known as a Bagging classifier. If no alternative estimator can be used, the default is a Decision Tree Classifier.

H. Extra Trees

This classifier employs the results of numerous de-correlated decision trees collected in a "forest" to classify. This is group learning. Extra Trees uses unpruned decision trees from the training dataset. Forecasts for regression and classification are made by averaging decision tree predictions or by majority voting.

I. Decision Trees

DTs are a non-parametric supervised learning method for classification and regression. Learn to forecast a target variable's value using simple decision rules drawn from the data. A tree is an approximation with piecewise constants. A decision tree can visualize all possible outcomes for a set of criteria. Using a decision tree, we try to establish a condition on the dataset's features to purify all labels or classes at each phase or node.

4. Results

After classification of 0 or 1 outcome, Logistics regression has scored highest accuracy among SVM, MNB, GNB and

KNN models. [17] [120] Following figure shows the confusion matrix, based on true positive, true negative, false positive and false negative values of SVM, LR, MNB, GNB and KNN.

A. Random Forests (RF)

Figure below shows the performance of RF model with accuracy of 87.93%, precision 87.54%, recall 87.78% and F1 score 86.32%

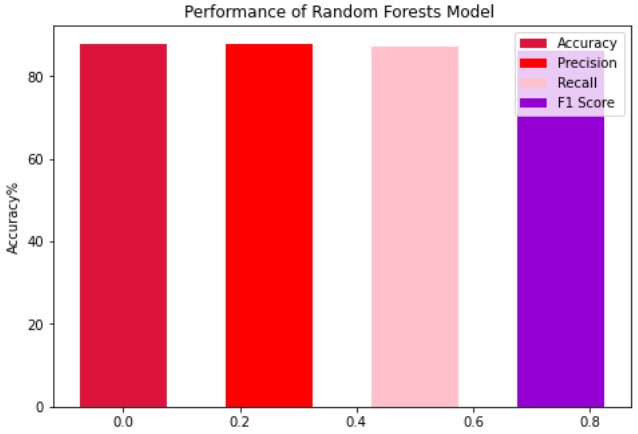

Figure 2. Performance Random Forests

Figure below shows the confusion matrix of RF model with true negative values of 37 and true positive values of 14 having the highest classified values of this research.

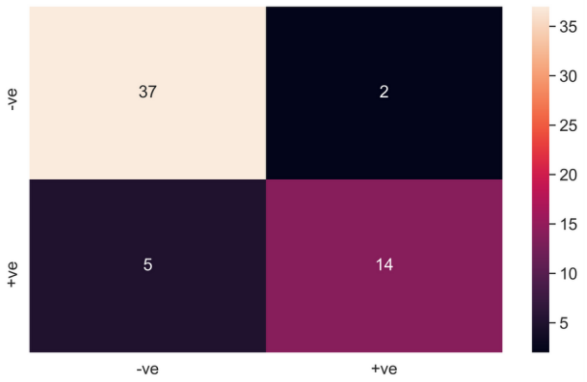

Figure 3. Confusion Matrix of Random Forests

B. Gradient Boosting Classifier (GBC)

Figure below shows the performance of GBC model with accuracy of 87.93%, precision 87.54%, recall 86.78% and F1 score 86.72%.

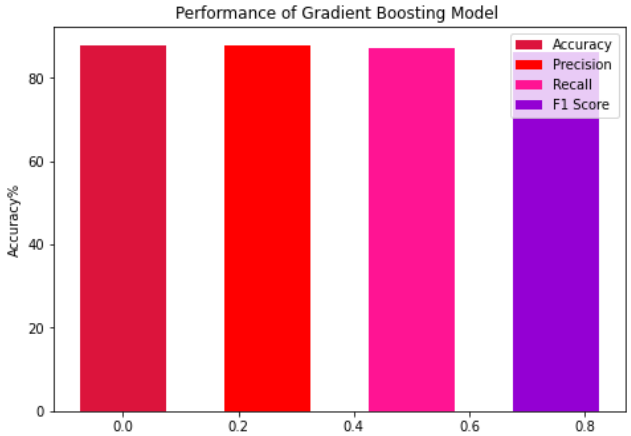

Figure 4. Performance of Gradient Boosting Classifier

Figure below shows the confusion matrix of GBC model with true negative values of 36 and true positive values of only 15 having the highest classified values of this research.

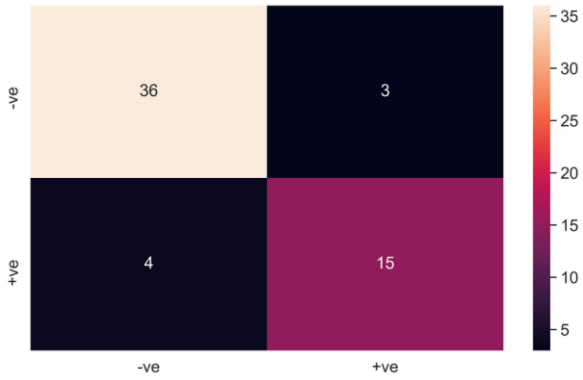

Figure 5. Confusion Matrix of Gradient Boosting Classifier

C. Adaboost Classifier (ABC)

Figure below shows the performance of ABC model with accuracy of 86.21%, precision 86.22%, recall 85.6% and F1 score 88.22%.

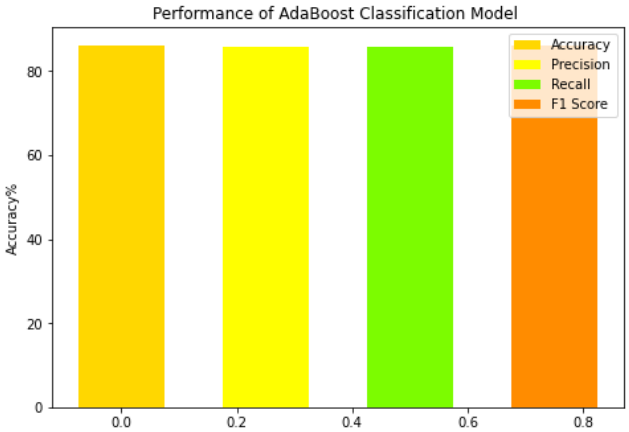

Figure 6. Performance of AdaBoost Classifier

Figure below shows the confusion matrix of ABC model with true negative values of 36 and true positive values of only 15 having the highest classified values of this research.

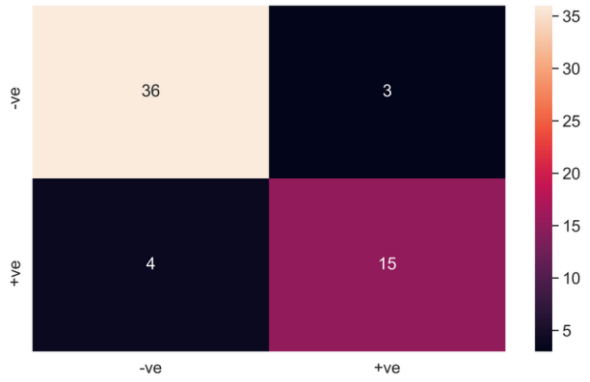

Figure 7. Confusion Matrix of AdaBoost Classifier

D. Bagging Trees (BT)

Figure below shows the performance of BT model with accuracy of 81.03%, precision 80.82%, recall 80.67% and F1 score 81.12%.

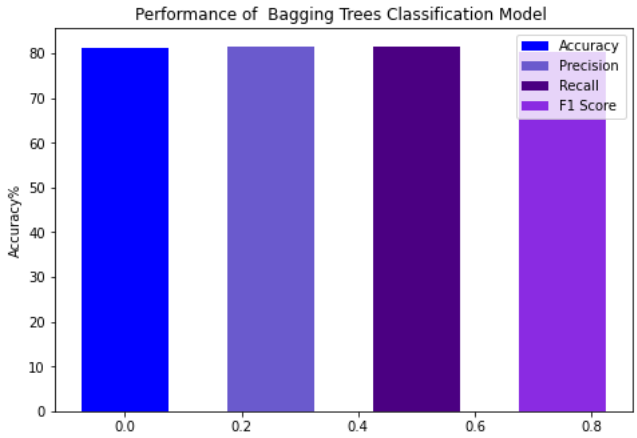

Figure 8. Performance of Bagging Trees

Figure below shows the confusion matrix of BT model with true negative values of 34 and true positive values of only 13 having the highest classified values of this research.

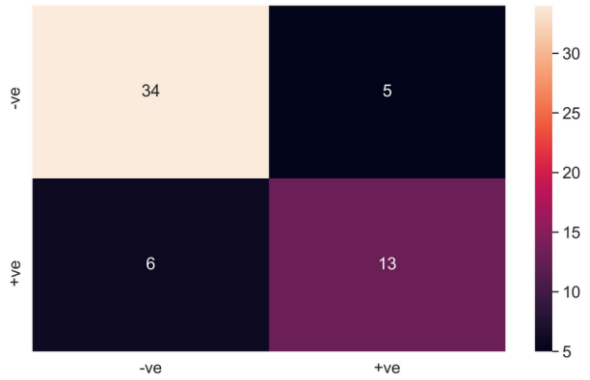

Figure 9. Confusion Matrix of Bagging Trees

E. Extra Trees

Figure below shows the performance of ET model with accuracy of 79.31%, precision 78.82%, recall 76.67% and F1 score 78.22%.

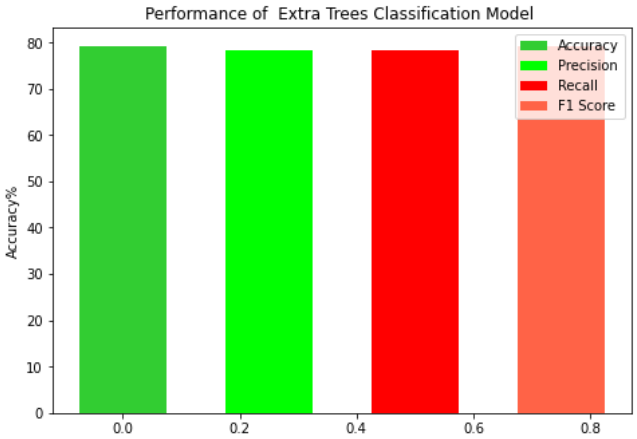

Figure 10. Performance of Extra Trees

Figure below shows the confusion matrix of ET model with true negative values of 36 and true positive values of only 10 having the highest classified values of this research

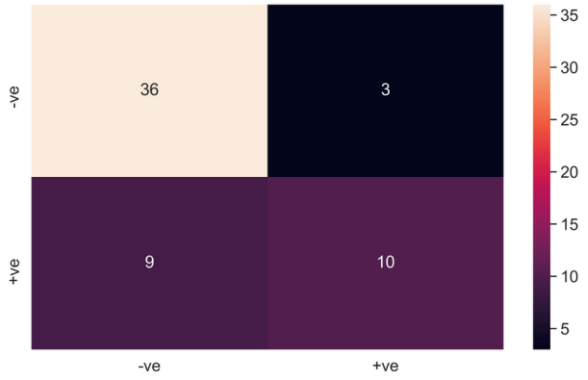

Figure 11. Confusion Matrix of Extra Trees

F. Decision Trees

Figure below shows the performance of DT model with accuracy of 84.40%, precision 84.06%, recall 83.67% and F1 score 84.22%.

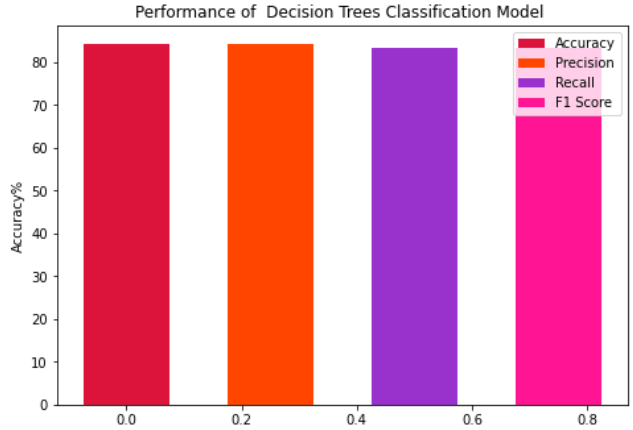

Figure 12. Performance of Bagging Trees

Figure below shows the confusion matrix of DT model with true negative values of 35 and true positive values of only 14 having the highest classified values of this research

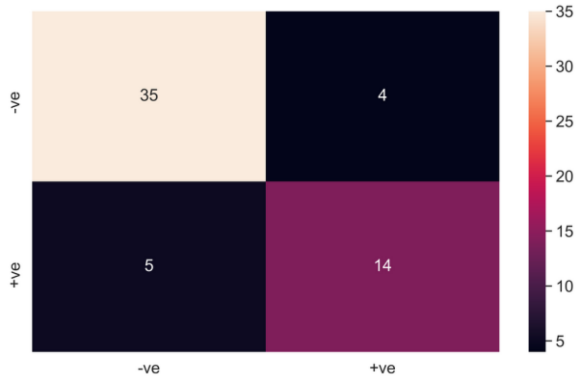

Figure 13. Confusion Matrix of Bagging Trees

G. Comparative Analysis

As can be seen in the image above, RF and GBC has the greatest number of true positive classified values, indicating that it is the best ensemble learning model.

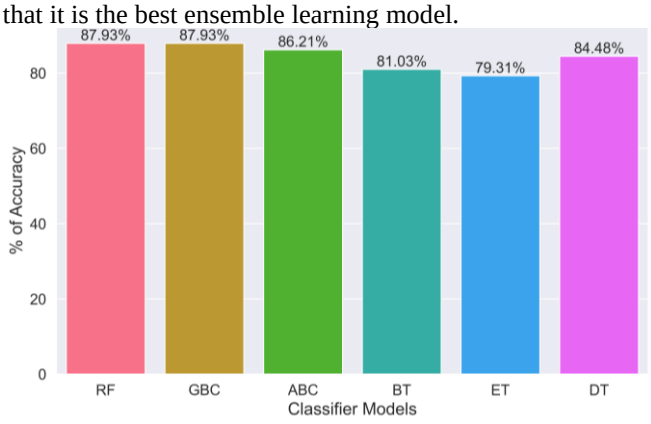

Figure 14. Linear Machine Learning Classification Models Performance

The accuracy of many models is depicted in the image above. In terms of accuracy, RF and GBC have the greatest score of 87.93 percent, followed by ABC, with 86.21 percent, BT, with 81.03 percent, ET, with 79.31 percent, and DT, with 84.48 percent.

5. Conclusions

A beneficial development is the creation of new security measures based on modern digital technologies. Strengthening the digital system's security is long overdue. It is possible to detect both internal and external intrusions with intrusion detection systems. In and around the neighbourhood, there are a variety of active and passive risks. Data can be stolen and moved throughout the system undetected if an attacker uses these flaws. The cutting-edge science of applied machine learning has numerous applications. In recent years, machine learning has become increasingly popular for predicting the future and categorizing data. This paper describes the development of a supervised learning-based machine learning system to detect cyber-attacks. Machine learning algorithms were trained using attacks on an open source website. The chart above shows the degree of precision that can be achieved with various types of modelling systems. RF and GBC, both with 87.93 percent, are the most accurate. Second place goes to ABC, third to BT, fourth to ET, and fifth to DT; all are in the top five in percentage..

References

[1] M. Stoyanova, Y. Nikoloudakis, S. Panagiotakis, E. Pallis, and E. K. Markakis, “A Survey on the Internet of Things (IoT) Forensics: Challenges, Approaches, and Open Issues,” *IEEE Commun. Surv. Tutorials*, vol. 22, no. 2, pp. 1191–1221, 2020, doi: 10.1109/COMST.2019.2962586.

[2] S. Khare and M. Totaro, “Ensemble Learning for Detecting Attacks and Anomalies in IoT Smart Home,” *Proc. - 2020 3rd Int. Conf. Data Intell. Secur. ICDIS 2020*, pp. 56–63, 2020, doi: 10.1109/ICDIS50059.2020.00014.

[3] D. Gibert, C. Mateu, and J. Planes, “The rise of machine learning for detection and classification of malware: Research developments, trends and challenges,” *J. Netw. Comput. Appl.*, vol. 153, no. March, p. 102526, 2020, doi: 10.1016/j.jnca.2019.102526.

[4] C. Report and B. Manjit, ? “Generic Datasets , Beamforming Vectors Prediction of 5G Celleular Networks,” pp. 1–22.

[5] A. Makkar, S. Garg, N. Kumar, M. S. Hossain, A. Ghoneim, and M. Alrashoud, “An Efficient Spam Detection Technique for IoT Devices Using Machine Learning,” *IEEE Trans. Ind. Informatics*, vol. 17, no. 2, pp. 903–912, 2021, doi: 10.1109/TII.2020.2968927.

[6] H. Yang and F. Wang, “Wireless network intrusion detection based on improved convolutional neural network,” *IEEE Access*, vol. 7, pp. 64366–64374, 2019, doi: 10.1109/ACCESS.2019.2917299.

[7] C. Yue, L. Wang, D. Wang, R. Duo, and X. Nie, “An Ensemble Intrusion Detection Method for Train Ethernet Consist Network Based on CNN and RNN,” *IEEE Access*, vol. 9, pp. 59527–59539, 2021, doi: 10.1109/ACCESS.2021.3073413.

[8] Muhammad Shoaib Akhtar, Tao Feng, "Deep Learning-Based Framework for the Detection of Cyberattack Using Feature Engineering", Security and Communication Networks, vol. 2021, Article ID 6129210, 12 pages, 2021. https://doi.org/10.1155/2021/6129210.

[9] S. Ho, S. Al Jufout, K. Dajani, and M. Mozumdar, “A Novel Intrusion Detection Model for Detecting Known and Innovative Cyberattacks Using Convolutional Neural Network,” *IEEE Open J. Comput. Soc.*, vol. 2, no. October 2020, pp. 14–25, 2021, doi: 10.1109/ojcs.2021.3050917.

[10] Zhang Fuyong; Wang, Yi; Liu, Shigang; Wang, Hua, Decision-based evasion attacks on tree ensemble classifiers World Wide Web; New York Vol. 23, Iss. 5,

(Sep 2020): 2957-2977. DOI:10.1007/s11280-020-00813-y

[11] Z. Tsiatsikas, G. Kambourakis, D. Geneiatakis and H. Wang, "The Devil is in the Detail: SDP-Driven Malformed Message Attacks and Mitigation in SIP Ecosystems," in IEEE Access, vol. 7, pp. 2401-2417, 2019, doi: 10.1109/ACCESS.2018.2886356.

[12] R. U. Rasool, U. Ashraf, K. Ahmed, H. Wang, W. Rafique and Z. Anwar, "Cyberpulse: A Machine Learning Based Link Flooding Attack Mitigation System for Software Defined Networks," in IEEE Access, vol. 7, pp. 34885-34899, 2019, doi: 10.1109/ACCESS.2019.2904236.

[13] Hua Wang, Lili Sun, Elisa Bertino, Building access control policy model for privacy preserving and testing policy conflicting problems, Journal of Computer and System Sciences, Volume 80, Issue 8, 2014, Pages 1493-1503, ISSN 0022-0000, https://doi.org/10.1016/j.jcss.2014.04.017.

[14] Vimalachandran P, Liu H, Lin Y, Ji K, Wang H, Zhang Y. Improving accessibility of the Australian My Health Records while preserving privacy and security of the system. Health Inf Sci Syst. 2020 Oct 8;8(1):31. doi: 10.1007/s13755-020-00126-4. PMID: 33088487; PMCID: PMC7544771.

[15] Hu, L. Wang, L. Qi, Y. Li, and W. Yang, "A novel wireless network intrusion detection method based on adaptive synthetic sampling and an improved convolutional neural network," IEEE Access, vol. 8, pp. 195741–195751, 2020, doi: 10.1109/ACCESS.2020.3034015.

[16] A. Kim, M. Park, and D. H. Lee, "AI-IDS: Application of Deep Learning to Real-Time Web Intrusion Detection," *IEEE Access*, vol. 8, pp. 70245–70261, 2020, doi: 10.1109/ACCESS.2020.2986882.

[17] K. Jiang, W. Wang, A. Wang, and H. Wu, "Network Intrusion Detection Combined Hybrid Sampling with Deep Hierarchical Network," *IEEE Access*, vol. 8, no. 3, pp. 32464–32476, 2020, doi: 10.1109/ACCESS.2020.2973730.

[18] H. Park, S. Park, and Y. Joo, "Detection of Abandoned and Stolen Objects Based on Dual Background Model and Mask R-CNN," *IEEE Access*, vol. 8, pp. 80010–80019, 2020, doi: 10.1109/ACCESS.2020.2990618.

[19] A. Liu and B. Sun, "An Intrusion Detection System Based on a Quantitative Model of Interaction Mode between Ports," *IEEE Access*, vol. 7, pp. 161725–161740, 2019, doi: 10.1109/ACCESS.2019.2951839.

[20] Wenqi Wang, Lina Wang, Run Wang, Aoshuang Ye, Jianpeng Ke, Better constraints of imperceptibility, better adversarial examples in the text, International Journal of Intelligent Systems, 10.1002/int.22696