Journal of Computing, Science &Technology
https://focjournal.unidel.edu.ng/
editorjcst@unidel.edu.ng

ISSN: 3043-6818 Print https://focjournal.unidel.edu.ng

Anomaly-based Intrusion Detection Techniques in Internet of Things Ecosystem-A Systematic Review

Ede Ewariezi Hyacinth [1], Edje E. Abel[2], Omede U. Edith[3], Atonuje Ephriam[4], Ogeh Clement[5], Akazue I. Maureen[6]
[1-6]Department of Computer Science, Faculty of Science, Delta State University, Abraka, Delta State
ede123@gmail.com[1], edjeabel@delsu.edu.ng[2], edithomde@delsu.edu.ng[3], atonujeehpriam@delsu.edu.ng [4], ogehclement@delsu.edu.ng [5], akazue@delsu.edu.ng [6]

Corresponding Author's Email: edjeabel@delsu.edu.ng

ABSTRACT

Article Info

Date Received: 14-01-2024
Date Accepted: 12-04-2024

Keywords:
Anomaly, Intrusion Detection, Internet of Things, Cybersecurity, Detection Techniques

With a vast array of smart and connected devices and applications available in many areas, including green IoT-based agriculture, smart farming, smart homes, smart transportation, smart health, smart grid, smart cities, and smart environment, the Internet of Things (IoT) technology has emerged to enhance people's lives. IoT devices are susceptible to cyberattacks. Though, researchers have sufficiently embraced the use of diverse techniques and algorithms as a means of securing data and information generated and transmitted in the Internet of Things ecosystem. Additionally, these techniques have been effectively applied in a number of domains, demonstrating its superiority in tackling intrusion detection attacks. The anomaly-based Intrusion Detection System (IDS) has an edge in identifying zero-day attacks because signature-based detection is limited when it comes to unknown threats. Therefore, this paper explicitly and systematically analyzed current techniques deployed in IoT ecosystem for the detection of anomaly-based intrusion attacks. Also, the processes and functionalities adopted by the techniques to predict the abnormality-based intrusion attacks, development and simulation tools adopted to implement and evaluate the effectiveness and performance of the techniques are highlight and discussed extensively. Finally, a summary of challenges and weaknesses of the techniques are briefly discussed, for onward investigation in future researches.

1.0 INTRODUCTION

The term "internet of things" (IoT) describes a new paradigm for communication in which devices are equipped with sensors and actuators to detect their environment, connect with each other, and exchange data via the internet [1]. All of the IoT's apps, goods, and services must be connected to a platform in order to gather, exchange, store, access, and share/transmit data from the outside world. Currently, there are around 50 billion internet-connected devices, and over the coming years, this number is predicted to increase significantly [2]; [3]. These enormous quantities of gadgets generate a massive amount of of information that numerous programs can utilize. Food, agriculture, smart farming, demotics, assisted living, e-health, and improved learning are just a few examples of the many scenarios where IoT technologies could be used. By the end of 2025, for example, there will be 15.3 billion IoT devices for smart agriculture. For real-time monitoring and environment assessment in many industrial sectors, a vast array of sensors and actuators is required in order to produce actionable insights and enable prompt decision-making [4]. Nonetheless, numerous obstacles impede the complete integration of IoT in both academia and business. These difficulties encompass a wide range of issues, including as mobility, scalability, dependability, security, and trust [5]. IoT devices are vulnerable to numerous possible security risks due to their immature and brittle communication protocols and applications that connect them to the worldwide internet [6]; [7]. The IoT ecosystems face major problems from the rising threat of cyberattacks. Additionally, in order to enhance their connectedness, IoT devices employ a variety of platforms and a blend of network connection protocols, including Ethernet, Wi-Fi, ZigBee, and wire-based technologies. To reduce security threats, these standards and protocols must be coordinated. In addition to the variety of technologies employed by the IoT sector, IoT applications are becoming more heterogeneous and scattered. In order to create a cyber-physical environment where everything can be located, operated, inspected, and upgraded, the IoT model was developed. The likelihood of network attacks rises due to its connectedness. Security issues arise because various IoT architecture levels may be impacted by hostile incidents and assaults. Depending on the malware's structure. In a similar vein, intrusion detection systems in IoT were further described by [8]. Intrusion detection systems, or IDS, are used to identify unwanted intrusions into computers and networks [9]. When these devices detect an intrusion, they are known to sound alarms. In order to facilitate secure online communication, numerous IDS have been introduced. It keeps a close eye out for malicious behavior on the network and notifies the system administrator when threats are discovered. IoT devices are compact and simple to set up in isolated locations, however, because of their small size and small battery capacity, the computing power is rather poor [10]. Additionally, they communicate via lightweight protocols.

These factors dictate that attack detection algorithms should be energy-efficient and lightweight. To shield IoT devices from cybercriminals, numerous intrusion detection system (IDS) solutions have been put forth. Proactive and reactive security solutions are separated into two categories. The Internet of Things can be effectively protected from outside attacks by taking preemptive precautions. However, because the Internet of Things is connected to the worldwide web, there is a considerable risk of intrusion by malicious actors who can defeat preventive measures. As a second line of defense, intrusion detection systems (IDSs) can stop a lot of cyberattacks. Researchers and industries working in the IoT space have given IDS solutions a lot of attention, and numerous IDS solutions have been put out [11.12]. IDS solutions can be divided into three groups based on the detection method: hybrid IDS model, anomaly, and signature. Generally speaking, known assaults respond better to the signature-based strategy, but unknown attacks respond better to the anomaly-based method. However, the signature-based approach is inefficient and ineffective for IoT due to the heterogeneity, dynamicity, and complexity of the IoT network because it necessitates ongoing human intervention and knowledge expertise to extract attack patterns and signatures in order to update the IDS model [13]. Anomaly-based intrusion detection system (IDS) can identify zero-day threats and requires less human interaction in the IoT ecosystem [13]. Both anomaly-based and signature-based techniques are used in the hybrid method. However, the application of signature-based intrusion detection systems (IDS) in IoT networks is restricted due to the impracticality of relying on pre-defined attack patterns [12]. Therefore, anomalous intrusion detection systems are essential for detecting intrusions in Internet of Things environments. An anomaly-based intrusion detection system, is an intrusion monitoring system for discovering both network and computer invasions and exploitation by identifying unusual activity in the system and classifying it as either normal or abnormal [14]. Various researchers have conducted literature survey on Intrusion based detection in IoT ecosystem. For example, [14], conducted a survey on intrusion detection models, compared with prevention models to mitigate DDoS attacks. Additionally, various anomaly detection methods, classifications of intrusion detection systems, and models of intrusion detection systems based on datasets were discussed. Consequently, [15] specifically performs analysis from the following angles: datasets, coauthor relationships, evaluation metrics, application areas, data preprocessing, and intrusion attack-detection approaches. Furthermore, [16] presented a survey on benchmark detection rates and key performance parameters, as well as the necessary effectiveness of the different approaches. For the purpose of classification, the following four machine learning techniques were assessed: Artificial Neural Network (ANN), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR). In [17], IoT cybersecurity threats as investigated, relevant suggestions for a real-time intrusion detection system, and a real-time dataset for assessing security systems that protect against various cyber-attacks. Consequently, [18] categorizes IDS approaches in IoT as specification-based, hybrid, anomaly-based, signature-based, and based on the detection technique. Additionally, the authors provide a parametric comparison for the IDS techniques. Next, the advantages and disadvantages of the selected mechanisms are discussed. At some point, an examination of unresolved issues and prospective trends are presented. [19], investigated current research on machine learning algorithms in the literature pertaining to IoT intrusion detection. Classifying them into three groups: Security Problem, Detection Method, and Agent Placement Strategy. Highlighting some possible open questions and future research directions. Previous studies have significantly analyzed and show the implications and prospects of applying various machine and deep learning algorithms for detection of anomaly-based intrusion in IoT ecosystem, but they are unable to explicitly and systematically analyzed the processes or functionalities of these techniques. As a result, this research tends to investigate, explicitly and systematically analyzed current techniques and algorithms deployed from 2017 to 2023, for detecting anomaly-based intrusion in IoT paradigm. The contributions of this research are listed as follows;

- To explicitly and systematically analyzed current techniques deployed for anomaly-based intrusion detection in IoT ecosystem.
- Processes and functionalities used by the techniques to predict abnormality-based intrusion IoT infrastructure.
- The programming languages and simulation environment utilized to implement and evaluate the effectiveness and performance of the current techniques for anomaly-based intrusion detection in IoT.
- Challenges and weaknesses of the current techniques that may lead to future research exploration are also presented.

The remainder of this paper comprise the research methodology used to achieve the research contributions, the research findings that explicitly and systematically analyzed the algorithms deployed, the processes and functionalities that the algorithms adopted, the simulation environment and programming languages that were used to develop and assess the algorithms' performance as well as their challenges. concluding with a discussion section that highlights the advantages and constraints of this research.

1.1 Materials and Methods

Through a thorough analysis of the predefined objectives, the research methodology aims to assist in understanding the overall impact of using various algorithms for the prediction of anomaly-based intrusion in IoT infrastructure. The main Research Questions (RQ) that are developed to investigate the current study are listed below.

- What features and attributes do the algorithms have?
- What are the advantages and disadvantages of existing algorithms for detecting anomaly-based intrusion in IoT infrastructure?
- What processes or procedures are employed by the algorithms to detect anomaly-based intrusion in IoT infrastructure?
- What about the programming languages and simulation environments used for developing and assessment of the algorithms' performance?

A search was conducted for relevant papers that would support the study in five major electronic research repositories: IEEE Xplore, Elsevier, Springer, Wiley Online Library, and Science Direct. However, a few articles from MDPI and Hindawi that are slightly related to the subject are also included in this analysis. To do a comprehensive automated text search using both manual and search engine screening, we have established the following keywords for the search procedure: "anomaly-based intrusion detection," "Internet of Things," and "anomaly-based intrusion detection in Internet of Things" in the context of the research domain. Boolean operators with the predefined keywords and within the scope of the formulated research questions were used in the study to classify relevant articles. Additionally, the following inclusion criteria were used to screen, filter and retrieved the relevant articles:

- The relevance of the paper to the application of anomaly-based intrusion detection techniques in IoT infrastructure.
- The predominant language for articles published between 2017 and 2023 should be English.
- Selecting exclusively primary studies from related researches.
- The document ought to provide a comprehensive comprehension grounded in the formulated research questions.

A time limit was set for the search in compliance with the inclusion criteria, ensuring that all pertinent papers were located and gathered with a focus on the pre-specified keywords. To reduce their length and make them easier to read, the paragraphs of the research articles that were found were further screened using the keywords. During the first phase, an estimated 188 articles were collected for the years 2017–2023. Using screening based on keywords and titles, a total of 90 articles were filtered out. The final stage involved filtering the remaining articles based on the abstract using the predefined searching terms and the Boolean AND operator. The authors selected the final 20 papers based on the inclusion criteria for further research and analysis, considering each of the research questions that had been specified. Figure 1 illustrates the entire screening process.

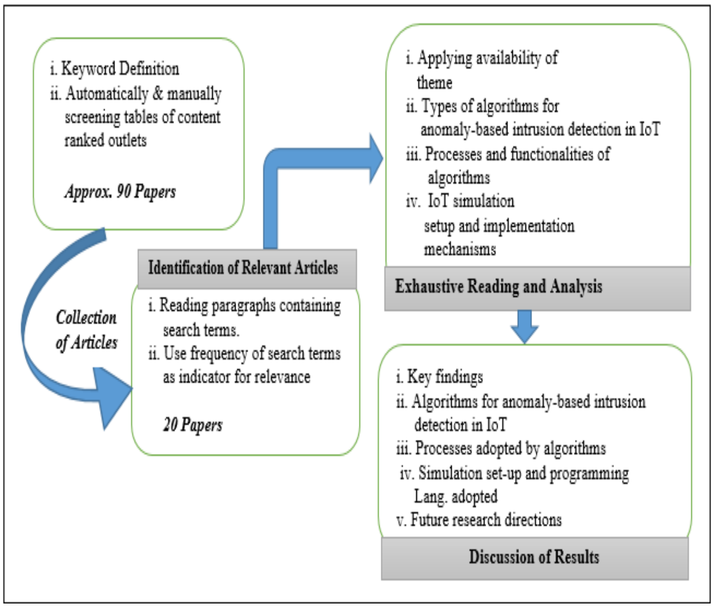
i. Keyword Definition
ii. Automatically & manually screening tables of content ranked outlets
Approx. 90 Papers
Collection of Articles
Identification of Relevant Articles
i. Reading paragraphs containing search terms.
ii. Use frequency of search terms as indicator for relevance
20 Papers
i. Applying availability of theme
ii. Types of algorithms for anomaly-based intrusion detection in IoT
iii. Processes and functionalities of algorithms
iv. IoT simulation setup and implementation mechanisms
Exhaustive Reading and Analysis
i. Key findings
ii. Algorithms for anomaly-based intrusion detection in IoT
iii. Processes adopted by algorithms
iv. Simulation set-up and programming Lang. adopted
v. Future research directions
Discussion of Results

Fig.1: Research Methodology Structure

1.2 Results

An improved Naïve Bayes (INB)-enabled Principal Component Analysis techniques was developed for Network-based Intrusion Detection System (NIDS) anomaly detection [20]. The PCA technique is utilized to extract relevant features from the dataset by computing the correlation between sample dataset mean and standard deviation. Thereafter, the correlation between data sample are computed for the removal redundant data records to obtain the actual datasets. Then, the INB technique is deployed to classify the actual datasets by applying the conditional probability on data sample training. It computes the probability of every sample features to retrieve the anomaly data records. In [21] proposed an Average One Dependence Estimator (AODE) technique for the detection of network-based intrusion anomalies on the cloud. The AODE technique is a supervised learning algorithm that initially normalizes the data samples to remove excessive noise. Then perform classification operation on the data sample based on their similarity differences to discover potential anomalies data records.

The utilization of various machine learning and artificial neural network techniques for the detection of network attack and anomaly in IoT Sensors were proposed by [22]. Machine learning algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT) and Random forest were used to perform classification processes on the benchmark dataset in other to predict possible attack or anomalies. The performance of the machine learning algorithms was compared to that of ANN and was discovered that the Random forest performs better. In [15], a Modified Artificial Neural Network (MANN)-enabled Gray Woolf Optimization (GWO) technique was developed for the prediction of anomaly-based network intrusion attack on IoT sensor-cloud. The datasets are transferred into the ANN as input to be trained using the back propagation algorithm to reduce the error between the actual and desired output datasets. The output dataset is further processed by sorting it as either normal or anomalies as captured in Figure 2 Consequently, the GWO technique is applied to speed up the entire process by minimizing training error.

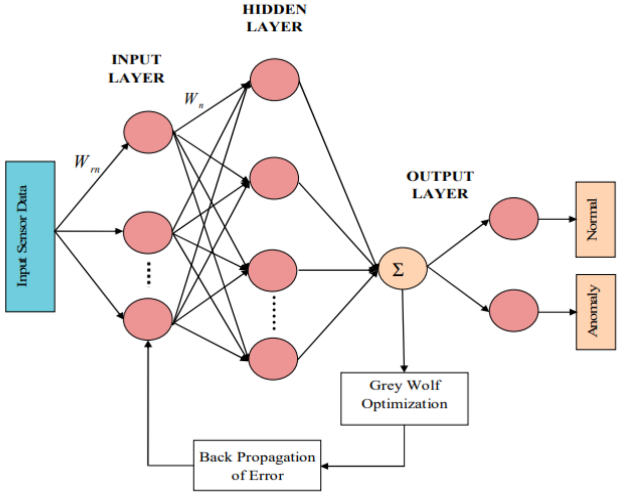

Fig.2: MANN-based GWO Technique

An Interquartile Range technique was proposed for the detection of outliers from Intrusion datasets on IoT-cloud environments [23]. The uninterrupted range of data input was divided into quartiles which are further analyzed to identify the range of outliers. Thereafter, the retrieved outliers are discarded using a filter method called Remove with Value (RWV). Consequently, [24] developed an ensemble machine learning technique to improve the accuracy of anomaly detection in intrusion environments. It uses the Neural Network Architecture Search (NAS) technique to train, test and validate the dataset with the support of Bayesian optimization (BO) method to determine the deep learning architecture of the entire datasets, which optimizes accuracy. Thereafter, the Kalman Filter model is applied to predict the anomalies data records which are regarded as potential intrusion attack.

Semi-supervised graph-based clustering technique is proposed for the detection of anomaly-based intrusion detection attack on dynamic data streams generated from IoT cloud platform [25]. It consists of K-means and Local Density Score (LDS) techniques. The K-means is used to select similar features from the datasets and combined them together as a local cluster. Then, the LDS technique is applied on each local cluster to discover potential anomaly data records with the support of Divide-and-Conquer method. Thereafter, the anomalies detected in all the local clusters are recomputed on the entire dataset to avoid bias and obtain desired anomalies. Conversely, [26] developed a hybrid machine learning technique to improve the detection of anomaly-based intrusion attacks on IoT. It uses the Feature Importance Decision Tree (FIDT) technique to select relevant features from the entire datasets. The FIDT deploys the filter-based method to compute data records of each feature and determine the optimal subset combination of the dataset features. Thereafter, the embedded-based method is used to compute the probability number of records for each feature in other to retrieve the relevant features from the entire datasets. Then, feature reduction process is performed on the selected features, by discovering and eliminating the desired anomalies using the Local Outlier Factor (LOF) technique.

A Hybridized Data Optimization-enabled machine learning algorithm is proposed for intrusion network attack-based anomaly detection [27]. At the initial stage, it uses the isolation forest (IF) technique for the selection of significant feature subsets from the datasets with the support of genetic algorithm (GA), which speeds up the feature selection process to optimize the data sampling ratio in each feature subset. After which the random forest technique is deployed to test/trained the feature subsets of the whole datasets, to generate potential anomalies. In [28], evaluate the performance of various machine learning techniques for improving the detection of intrusion network attack-bases anomaly detection. The machine learning techniques comprise of Support Vector Machine, Naive Bayes, K-nearest Neighbor, Decision Tree and Random Forest. The interquartile method was used to split the dataset into three subsets before they were trained by the algorithms for the discovery of potential anomalies. Experimental result shows that the K-nearest neighbor algorithm outperform the other techniques.

Unsupervised machine learning technique is proposed for the detection of anomaly in network traffic [29]. At the initial stage, the dataset is preprocessed to minimize redundancy and imbalance data records in feature classes. The standard scalar method was used to normalized the feature datasets in other to reduce the presence of imbalance issue and principal component analysis technique is applied to eliminate redundant data records from the feature datasets. Thereafter, the isolation forest technique is applied to train the preprocessed dataset for the discovery of potential anomaly data records that poses as intrusion attack to the network. Consequently, [30] adopted various machine learning classification techniques to analyze the detection of anomaly-based intrusion attack on network traffic in the IoT-Cloud platform. The datasets were preprocessed using statistical methods to remove irrelevant data records that always increases false alarm. Then, the classifier techniques which comprise K-nearest neighbor, support vector machine, decision trees. Logistic regression and random forest where used to test/trained the preprocessed dataset to be classified as either normal or anomaly-based intrusion data.

A statistical-enabled Optimized Deep Learning approach is developed for the discovery of intrusion attack on IoT Cloud infrastructure [31]. Firstly, the big dataset is preprocessed to eliminates irrelevant data records and converts features into one-hot-encoded vector with the support of Absolute Deviation (MAD) Estimator method. Followed by extracting and enhanced huge amount of correlated features using big visualization and statistical analysis methods. The feature extraction methods remove features with null values higher than 80% and retains the most relevant features as input for the deep auto-encoder. Then, deep Auto-encoder (AE) is applied to train the input features in a greedy-knowledgeable pattern for the detection of possible threats. Consequently, [32] proposed a Difficult Set Sampling Technique (DSST) algorithm for the detection of anomaly-based intrusion imbalanced network. At the initial stage, it uses the Edited nearest Neighbor (ENN) algorithm to split the imbalanced training dataset into the complex and simple datasets. Followed by compressing the majority dataset in the complex set category to minimize the majority using K-means techniques. Furthermore, the minority data records features are zoom in and out continuously in the complex set, generating new datasets to increase the minority number. Therefore, the simple set, the compressed set of the majority in the complex and the minority in the complex set are joined with its augmentation datasets to produce a new training set. After which different classifier algorithms such as Random Forest, Support Vector Machine, XGBoost, Long and Short Time memory (LSTM), AlexNet and Mini-VGGNet were used to train the new training set for the discovery of either normal or intrusion attack.

An Extreme Value Machine (EVM) technique is proposed to predict intrusion network attack on IoT-Cloud platform [33]. The EVM resolves the issue of generating new dataset classes from the entire original dataset during feature selection process. Thereafter, the Extreme Value Theory method is used to train the entire classes of the dataset to detect intrusion attack. More so, [33] introduced a deep neural network model for the prediction of intrusion-based anomies in a network system. First, the dataset is preprocessed using Correlation Feature Selection (CFS) method to eliminate redundant features. The CFS calculate the correlation value for every feature vector to every other feature to ascertain their similarity difference. Therefore, features with high correlation value that is greater than 0.8 are discarded from the entire data samples. The leftover feature data samples are further deduced using a random forest classifier technique. Then, a Shallow Neural Network (S-NN) Model is applied to train the deduced feature data sample for predicting possible intrusion attack. The S-NN model comprise of one input layer with 100 neurons and one output layer with five neurons. The deduced data samples are fed into the input layer which is activated using the Rectified Linear Unit (ReLU) method and processes the data samples via its neurons before forwarding to the output layer. The output layer uses SoftMax activation function to covert data samples into probabilities that compute the entire samples. After which the Deep-optimized neural network model (D-ONN) is used to categorize the attacks by optimizing the different hyper-parameters.

An Autoencoder-enabled Network model is developed to improve the performance of anomaly-based intrusion attack on IoT cloud platform [34]. First, the dataset is preprocessed to eliminate irrelevant data samples using the 95th Percentile Rule method, after which the relevant features are normalized using Min-Max function and characterized in the latent space. Furthermore, the latent space is utilized for the reconstruction of output. Then, the similarity difference between the output and the original dataset is determined and a reconstruction error is calculated. The optimal value of all reconstruction errors is envisage as the threshold to predict anomalies. Conversely, [35] proposed an intrusion attack detection through an optimized feature selection model. It utilizes the k-means clustering technique for the selection of relevant features while discarding irrelevant ones. The relevant feature data samples were trained using the Decision Trees (DT) that is based on RJ48 to detect potential intrusion attack.

A hybridized machine learning technique is proposed for the prediction of intrusion-based anomaly on IoT-Cloud infrastructure [36]. First, the dataset is preprocessed using Naïve-Bayes algorithm to eliminate redundant data records from the data samples. Then, relevant data features are selected using Optimized Support Vector Machine (OSVM) Algorithm, while discarding irrelevant ones. Thereafter, the relevant data is trained with the support of Prioritized K-Nearest Neighbor (KNN) algorithm to detect attacks. Consequently, [37] developed a two-phased machine learning technique for the prediction of anomaly-based intrusion attack on IoT Cloud infrastructure. In the first phase, the datasets are sub-divided into four groups in accordance to the data types such as integer, binary, floating and nominal. Thereafter, they were classified using Naïve-Bayes algorithm, followed by selecting the relevant data samples after the classification process using the Majority Voting method. The relevant data samples are passed to the second phase for further classification using Elliptic Envelope Method for the prediction of potential attack, as denoted in Figure 3.

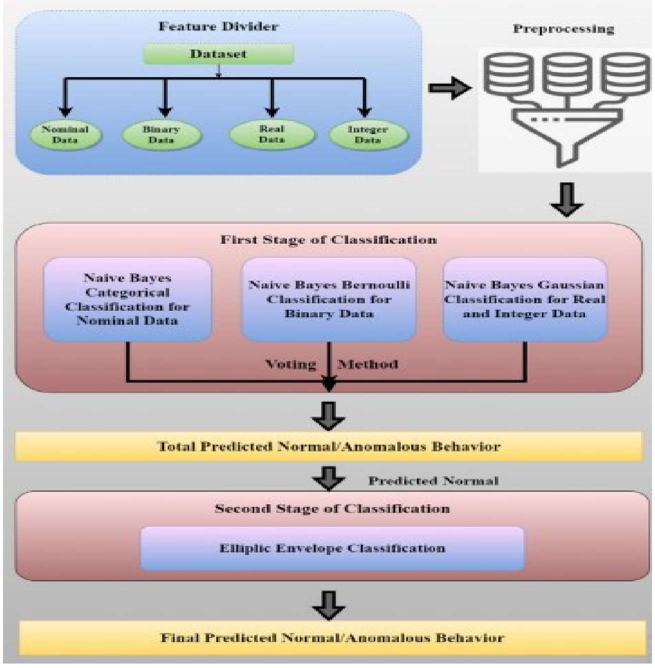
Feature Divider
Dataset
Nominal Data
Binary Data
Real Data
Integer Data
Preprocessing
First Stage of Classification
Naive Bayes Categorical Classification for Nominal Data
Naive Bayes Bernoulli Classification for Binary Data
Naive Bayes Gaussian Classification for Real and Integer Data
Voting Method
Total Predicted Normal/Anomalous Behavior
Predicted Normal
Second Stage of Classification
Elliptic Envelope Classification
Final Predicted Normal/Anomalous Behavior

Fig.3: A Two-phased Machine Learning Technique

Table.1: The Properties and Features of the Techniques

Article Title	Algorithm	Processes	Challenges Resolved	Outcome	Simulation Package	Benchmark	Metrics	Weakness
IoT-Fog-Cloud model for anomaly detection using improved Naïve Bayes and principal component analysis, Manimuranga [20]	Improved Naïve-Bayes (INB) And Principal Component Analysis (PCA) techniques	Classification and Clustering	Redundant data records as possible cyber attack	Improved outlier detection on network intrusion dataset based on Accuracy Precision	Tempdump Software and Cyber Range Lab	Random Forest (RF), K-nearest Neighbor(KNN) and Support Vector Machine (SVM)	Accuracy, Recall, FI-score, Precision	Limited in false positive rate issues
Effective and efficient network anomaly detection system using machine learning algorithm, Nawir [21]	Average One Dependence Estimator (AODE) technique	Classification	Fatal destruction of Centralized network system coursed by malicious code.	Improved anomaly detection based on accuracy and execution time	Java programming language, Eclipse and WEKA	Naïve-Bayes(NB), Multi-Layer Perceptron (MLP), Radial Basis Function Network (RBFN) and J48 trees	Accuracy and Execution Time	Time Complexity
Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches [22]	Machine Learning (ML) and Artificial Neural Network (ANN) techniques	Classification and Deep learning	Network attacks such as denial of service and anomalies	Improved detection of network attack and anomaly based on accuracy	Python	Logistic Regression, Random Forest, Artificial Neural Network,	Accuracy, Recall, Precision and F1-score	Limited in detecting attacks and anomalies in big datasets.
Privacy Preservation in Edge Consumer Electronics by Combining Anomaly Detection with Dynamic Attribute-Based Re-Encryption, Yang [15]	Artificial Neural Network (ANN) and Gray Wolf Optimization (GWO)	Deep Leering and Optimization	Issues of network hackers modifying private data.	Improved anomaly detection and intrusion attack	Java JDK 1.8	Artificial Neural Network (ANN)	Accuracy, Detection Rate, false Alarm Rate and execution Time	Local search minima and time complexity.
Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset [23]	Interquartile Range Technique	Statistical	Excessive False Alarm	Improved detection of outliers by minimizing false alarm.	MATLAB and Java	NS	Accuracy, False Positive and Recall	NS

Table.1: The Properties and Features of the Techniques (Continue)

Article Title	Algorithm	Processes	Challenges Resolved	Outcome	Simulation Package	Benchmark	Metrics	Weakness
An Ensemble of Prediction and Learning Mechanism for Improving Accuracy of Anomaly Detection in Network Intrusion Environments [24]	Neural Architecture Search and Kalman Filter Techniques	Deep learning And Statistical	Malicious attacks on network system	Improved anomaly detection rate	Python	Dynamic Neural Network	Detection Rate, False Alarm Rate, Accuracy, Recall and Precision	NS
Multistage System-Based Machine Learning Techniques for Intrusion Detection in WiFi Network [25]	K-means and Local Density Score (LDS)	Clustering	Compromised data injected into the Wi-Fi network channel	Improved anomaly intrusion detection rate	Python	Local Area Factor (LOF)	False Alarm Rate, Accuracy, Half Total Rate, False Negative Rate and Execution Time	NS
A hybrid machine learning method for increasing the performance of network intrusion detection systems [26]	Feature Importance Decision Tree (FIDT) and Local Outlier Factor (LOF)	Classification and Clustering	Unauthorized access to network system	Improved anomaly-based intrusion detection rate attack.	Jupyter Notebook with Python 3.7.7	NS	Accuracy, Sensitivity, Specificity and false alarm rate.	Imbalanced data records for each feature and limitation of the LOF cluster size.
Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms [27]	Isolation Forest (IF), Genetic Algorithm (GA) and Random Forest (RF).	Classification and Optimization	Low detection rate and High false alarm	Improved intrusion-based anomaly detection rate with minimum false alarm	Python using Pycharm 2017.	AdaBoost, RUSBoost, Support Vector Machine	Accuracy, False Alarm Rate, F1-score and Precision	Time complexity
Feature Classification and Outlier Detection to Increased Accuracy in Intrusion Detection System [28]	Support Vector Machine (SVM), Naive Bayes (NB), K-nearest Neighbor (KNN)	Classification	Unauthorized access to network	Improved anomaly detection rates	WEKA	NS	Accuracy, Execution Time	Global Optima entrapment

Table.1: The Properties and Features of the Techniques (Continue)

Article Title	Algorithm	Processes	Challenges Resolved	Outcome	Simulation Package	Benchmark	Metrics	Weakness
Anomaly detection in Network Traffic Using Unsupervised Machine Learning Approach [28]	Principal Component Analysis (PCA) and Isolation Forest (IF) techniques	Classification and Clustering	Imbalance and redundant data values	Improved anomaly detection rate with minimum false alarm rate.	Python		False Positive Rate, False Negative Rate, Accuracy and Execution Time	NS
Machine Learning for Classification analysis of Intrusion -KDD Dataset [29]	K-nearest Neighbor, Support vector machine, Decision Trees, Logistic Regression and Random Forest technics	Classification	Minimize false alarm rate and dimension of datasets	Improved anomaly-based intrusion detection rate with minimum false alarm	Python	NS	Accuracy, F1-score, Precision and Recall	NS
Statistical Analysis Driven Optimized Deep Learning System for Intrusion Detection [30]	Absolute Deviation (MAD) Estimator and Deep Auto-encoder (AE) techniques	Statistical and Clustering.	Issue of intrusion attack on big datasets	Improved the discovery of potential intrusion threats.	Python	Shallow MLP Classifier	Precision, Recall, F-measure and Accuracy	Limited in detecting intrusion attack on real data stream.
Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning [31]	Difficult Set Sampling Technique (DSSTE) algorithm, Edited Nearest Neighbor (ENN) and KMeans technique.	Deep learning and clustering	Imbalanced network traffic and malicious cyber-attacks	Improved intrusion detection rate	Python	Smote Algorithm	Accuracy, Precision, Recall and F1-score	NS
Incremental Open Set Intrusion [32]	Extreme Value Theory and Extreme Value Machine techniques	Clustering	Open set incremental learning problem	Improved intrusion detection rate based on accuracy	Python	Weibull-Support Vector Machine (W-SVM)	Accuracy	Imbalanced issue

Table.1: The Properties and Features of the Techniques (Continue)

Article Title	Algorithm	Processes	Challenges Resolved	Outcome	Simulation Package	Benchmark	Metrics	Weakness
An intrusion detection system using optimized deep neural network architecture [33]	Correlation Feature Selection (CFS) model, Shallow Neural Network and Deep Neural Network	Deep learning	Inefficient detection of intrusion attack	Improved intrusion-based anomaly detection rate.	Python	Support Vector Machine, Naïve-Bayes, Random Forest	Accuracy, Precision, Recall and F1-score	NS
Data Reduction for Optimizing Feature Selection in Modeling Intrusion Detection System [34]	K-means and Decision Trees-based RJ48 techniques	Clustering and Classification	Improper detection of intrusion attack due to irrelevant data samples	Improved intrusion detection rate with minimum false alarm	Python	Genetic-based Logistic Regression and Multimodal Fusion algorithms	Accuracy, Detection Rate and Precision	The presence of irrelevant data
Improving Performance of Autoencoder-Based Network Anomaly Detection on NSL-KDD Dataset [35]	Auto-encoder (AE) Model	Clustering/Deep learning	Imbalance in dataset resulting to overfitting	Improved Intrusion detection attack based on accuracy	Python	AE-Support Vector Machine	Accuracy, Precision, Recall and F1-score	NS
A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers [36]	Optimized Support Vector Machine, Naive-Bayes and Prioritized K-Nearest Neighbor techniques	Clustering and Classification	Missing data and Overfitting	Improved intrusion detection rate with minimum false alarm	Python	Principal Component Analysis	Detection Rate, Root Mean Squared Error and Execution Test Time	Global optima search space challenge
A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection [37]	Naïve-Bayes, Majority Voting and Elliptic Envelope Techniques	Classification and Statistical	Imbalanced data records	Improved intrusion detection rate	Python	K-nearest neighbor, Logistic Regression, Decision Trees, Linear Discriminant Analysis and Gradient Boosting	Accuracy, Precision, Recall and F1-score	NS

1.3 Processes Adopted by existing Algorithms

The literature review's findings show that most approaches rely on classification and clustering techniques. The classification process is largely used to identify anomalies in a given dataset, whereas the clustering process is mostly used to select and extract relevant characteristics from the dataset. The classification process is used in supervised learning techniques, which train models on labeled data. Determining the mapping function between the input and output variables is a task that every model must accomplish. In unsupervised learning, patterns deduced from unlabeled input data are employed. It aims to create patterns and order out of the incoming data. Since unsupervised learning makes use of the clustering process, supervision is not necessary. Rather, it autonomously groups the data into patterns or clusters.

Optimization process is also adopted by few algorithms to improve the processing speed or minimizes execution time it takes to detect any irregularities based on intrusion on the IoT ecosystem. Examples of algorithms that utilizes optimization process include bio-inspired algorithms such as Ant Colony, Bee Colony, Simulated Annealing, Particle and Glowworm Swarm algorithms. These algorithms have been tried and tested with regards to speed up anomaly-based intrusion detection processes in the IoT Ecosystem with an improved optimal performance.

1.4 Simulation and Programming languages adopted

It was also discovered that different types of programming languages and simulation tools are adopted in the implementation and simulation of the existing techniques. Simulation technologies like Weka, OmNet++, Contiki Cooja, Visual Basic.Net, and CloudSim were used to obtain verified performance evaluations of the models that were already in place. On the other hand, the present models were implemented mostly using programming languages including Python, Java, MATLAB, C, and R-Studio.

In addition to handling common data mining tasks like feature selection, data preparation, clustering, regression, and classification, Weka also includes a variety of visualization tools, algorithms, and graphical user interfaces for functions related to data analysis and predictive modeling. The Attribute-Relational File Format (ARFF) is the format that Weka expects its input to be named with.

The Objective Modular Network Testbed in C++, or OmNet++, is primarily used to build network simulators. It has also been used recently to simulate data mining procedures. OMNeT++ is freely available for use in non-commercial simulations, such as those run by academic institutions and in educational contexts.

In COOJA, a simulated Contiki is a real, compiled, and operational Contiki system. COOJA oversees and manages the system. Different Contiki libraries can be built and loaded to simulate different kinds of sensor nodes (heterogeneous networks) within the same COOJA simulation. COOJA uses a few functions to operate and evaluate a Contiki system. For instance, the simulator accesses all of the Contiki system's memory for analysis or provides instructions to the system on how to react to an event.

The.NET framework is required for Visual Basic.NET to function, and the language produces highly scalable and reliable programs. With VB.NET, you may create fully object-oriented programs that are equivalent to those created in other languages like C++, Java, or C#. Programs made with VB.NET can also be used with applications developed in Visual C++, Visual C#, and Visual J#. In VB.NET, everything is handled as an object.

The infrastructure and services of cloud computing are modeled using an open-source framework called CloudSim. It was made by the CLOUDS Lab team and is entirely written in Java. It is used to model and simulate a cloud computing system in order to test a hypothesis prior to developing software and reproduce tests and outcomes.

Python is a programming language that may be used to develop software, generate websites, automate tasks, and analyze and visualize data. The models that are now in use can also be implemented using the Java programming language. Without having to write in numerical codes, programmers can construct computer instructions with Java by employing commands that are based in English. A programming language called R-studio is used for statistical analysis and data visualization. It has been embraced by the fields of data mining, bioinformatics, and data analysis. The R language comes with a ton of extension packages that include reusable code, example data, and documentation.

The abbreviation MATLAB stands for "Matrix Laboratory." It is a fourth-generation programming language. MATLAB is multi-paradigm. As such, it is compatible with several programming paradigms, such as object-oriented, functional, and visual.

1.5 Challenges and Future Research Direction

The challenges of the existing models include data over-fitting and under-fitting, missing data issues, and computational complexity (memory usage and untimely execution). **Overfitting** occurs when the model tries to cover more data points in the dataset than is necessary; this leads to the model caching noise and incorrect values found in the dataset, all of which reduce the accuracy and efficiency of the model. Despite the improvements in performance, these challenges may still arise and influence future research directions.

Under-fitting occurs when our machine learning model is unable to recognize the underlying trend in the data. To avoid overfitting, which could lead to the model learning insufficiently from the training data, the training data stream can be stopped early. As a result, it might not be able to assess how well the data matches the dominant pattern.

The uncertainty that **missing data** introduces into datasets is a concern. When any of the observations in a data set are blank, this is known as missing data (often called missing values). Furthermore, if an observation contains missing data for a variable, it is deemed odd. Therefore, any research that assumes the missing value fits nicely into the rest of the data is faulty.

The two basic goals of **computational complexity** are to ascertain the possibilities and limitations of algorithmic efficiency as well as the computing resources required to solve an issue requiring time, space, or communication. One of the central concepts of computational complexity theory is the P versus NP dilemma, which poses the question of whether every problem that can be verified in polynomial time can also be solved in polynomial time. The problems in class P can be solved in polynomial time, while the problems in class NP have solutions that can be verified in polynomial time.

1.6 Discussion

According to the study, machine learning and deep learning algorithms are mostly used to identify abnormality-based intrusion activities in the IoT Network Infrastructure. In terms of anomaly-based intrusion, are the predictions of impending security threats to IoT network device components and data/information generated and processed to aid decision making. With minimum execution time, the majority of the algorithms adopted for this study produced notable performance outcomes based on accuracy, precision, recall, and specificity. this suggests that regardless of the size (big data) of datasets in question, machine learning algorithms are dependable and effective for processing IoT generated and transmission of processed datasets. To select relevant features that are further classed to detect actual anomaly-based intrusion threats, the majority of these algorithms uses the clustering process. Programming languages are used mostly in simulation environments to conduct the experiment. This may allow for more study in the future by carrying out the experiment in a real-time setting. With the use of the algorithms, faulty IoT devices (e.g sensors) that generates error data can easily be detected to minimize inaccurate information that could lead to disaster. Also, the actual intrusion threat events will be detected at ease within minimal time frame. For example, detecting unauthorized access event into IoT-enabled network system, thereby compromising the integrity of data generated or during transmission process to other infrastructure such as the Cloud. Moreso, it will help to mitigate data traffic congestion during transmission process from a device to other devices within the IoT network ecosystem.

In conclusion, this research presented an explicitly and systematic review on various techniques deployed for the detection of anomaly-based intrusion attacks in IoT environments. It further highlights and discusses the process and functionalities, development and simulation tools used for implementing and evaluating the effectiveness and performance of the techniques. The specified inclusion and quality criteria resulted in the inclusion of a total of 20 current researches. Based on the findings of a study done on anomaly intrusion detection in IoT using various techniques, an extensive taxonomy was offered. Using diverse approaches, this work shed light on the characteristics and state of the art for anomalous intrusion detection in Internet of Things environments. Finally, the paper covered the challenges and weaknesses in using the prevailing current techniques for anomalous intrusion detection in the Internet of Things. Future research will explore the development of a suitable unsupervised learning algorithm for predicting anomaly and signature based intrusion attacks in IoT ecosystem.

REFERENCES

[1] Edje E. Abel and Muhammad Shafie Abd Latiff (2021), The utilization of algorithms for cloud internet of things application domains: a review. (Springer), 15(3), 1-27

[2] Almiani, M., AbuGhazleh A., Al-Rahayfeh A., Atiewi S. and Razaque A. (2020), Deep recurrent neural network for IoT intrusion detection system. Simulation Model Practice and Theory (Elsevier), 101, 10-31

[3] Nugent, C.D.; Zhang, S.; Cleland, I. (2020) IoT reliability: A review leading to 5 key research directions. CCF Trans. Pervasive Computing Interaction (Springer*)*, 2, Pages 147–163.

[4] Ruan, J.; Wang, Y.; Chan, F.T.S.; Hu, X.; Zhao, M.; Zhu, F.; Shi, B.; Shi, Y.; Lin, F. (2019), A Life Cycle Framework of Green IoT-Based Agriculture and Its Finance, Operation, and Management Issues. IEEE Communication Magazine, 57, Pages 90-96.

[5] Pal, S.; Hitchens, M.; Rabehaja, T.; Mukhopadhyay, S. (2020), Security Requirements for the Internet of Things: A Systematic Approach. MDPI Sensors, 20(20), 1-35.

[6] Da Xu L., He W., and Li, S. (2014), Internet of things in industries: A survey, IEEE Transaction on Industrial Informatics. 10, 2233–2243.

[7] Lin J., Yu W., Zhang N., Yang X., Zhang H. and Zhao W (2017), A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. IEEE Internet Things, 4, 1125–1142.

[8] N. K. Suryadevara and G. R. Biswal (2019), Smart plugs: Paradigms and applications in the smart city-and-smart grid. MDIP Energies, 12(10), 1–20.

[9] Muaadh A. Alsoufi, Shukor Razak, Maheyzah Md Siraj, Ibtehal Nafea, Fuad A. Ghaleb, Faisal Saeed and Maged Nasser (2021), Anomaly-based intrusion detection systems in iot using deep learning: A systematic literature review. MDPI Applied Sciences, 11, 2-24

[10] A. E. Edje, Shaffie Muhammad Abd Latiff, and Howe Weng Chan (2021). Enhanced Non-parametric Sequence-based Learning Algorithm for Outlier Detection in the Internet of Things. Neural Processing Letters (Springer), 54, 1889-1919.

[11] Shi, W.-C. and Sun, H. (2020), DeepBot: A time-based botnet detection with deep learning. Soft Computing (Springer), 24, 16605–16616

[12] Shone N., Ngoc T.N., Phai V.D., Shi Q. (2018), A Deep Learning Approach to Network Intrusion Detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 2, 41–50.

[13] Fahim M. and Sillitti (2019), A. Anomaly Detection, Analysis and Prediction Techniques in IoT Environment: A Systematic Literature Review. IEEE Access, 7, 81664–81681.

[14] Nivedita Mishra and Sharnil Pandya (2021), Internet of Things Applications, Security Challenges, Attacks,

Intrusion Detection, and Future Visions: A Systematic Review. *IEEE Access*, 9, 59353-59377.

[15] Eunmok Yang, Velmurugan Subbiah Parvathy, P. Pandi Selvi, K. Shankar, Changho Seo, Gyanendra Prasad Joshi and Okyeon Yi (2020), Privacy Preservation in Edge Consumer Electronics by Combining Anomaly Detection with Dynamic Attribute-Based Re-Encryption. Mathematics (MDPI), 08, 1-13.

[16] Khalid Albulayhi, Abdallah A. Smadi, Frederick T. Sheldon and Robert K. Abercrombie (2021), IoT Intrusion Detection Taxonomy, Reference Architecture, and Analyses. Sensors (MDPI), 21(19), 1-30.

[17] Ines Martins, Joao S. Resende, Patricia R. Sousa, Simao Silva and Luis Antunes (2022), Host-based IDS: A review and open issues of an anomaly detection system in IoT. Future generation Computer Systems (Elsevier), 133, 95-113.

[18] Arash Heidari and Mohammad Ali Jabraeil Jamali (2023), Internet of Things intrusion detection systems: a comprehensive review and future directions. Cluster Computing (Springer), 26, 3753–3780

[19] Mudhafar Nuaimi, Lamia Chaari Fourati and Bassem Ben Hamed (2023), Intelligent approaches toward intrusion detection systems for Industrial Internet of Things: A systematic comprehensive review. Journal of Network and Computer Applications (Elsevier), 215, 222-234

[20] Manimurugan S. (2021), IoT-Fog-Cloud Model for Anomaly Detection using Improved Naïve Bayes and Principal Component Analysis. Journal of Ambient Intelligence and Hummanized Computing (Springer), 1-10.

[21] Mukrimah Nawir, Amiza Amir, Naimah Yaakob, Ong Bi Lynn (2019), Effective and efficient network anomaly detection system using machine learning algorithm. Bulletin of Electrical Engineering and Informatics, 8(1), 46-51.

[22] Mahmudul Hasan, Md. Milon Islam, Md Ishrak Islam Zarif and M.M.A. Hashem (2019), Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things (Elsevier), 07, 1-14.

[23] Vinutha H. P., Poornima B. and Sagar B. M. (2018). Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset, Advances in Intelligent Systems and Computing (Springer), 511-518

[24] Imran, Faisal Jamil and Dohyeun Kim (2021), An Ensemble of Prediction and Learning Mechanism for Improving Accuracy of Anomaly Detection in Network Intrusion Environments. Sustainability (MDPI), 13, 1-22.

[25] Vu Viet Thang and F. F. Pashchenko (2019), Multistage System-Based Machine Learning Techniques for Intrusion Detection in WiFi Network. Journal of Computer Networks and Communications (Hindawi), Vol (2019), 1-13.

[26] Achmad Akbar Megantara and Tohari Ahmad (2021), A hybrid machine learning method for increasing the performance of network intrusion detection systems. Journal of Big Data, 8(142), 1-19.

[27] Diadong Ren, Jiawei Guo, Wang Qian, Huang Yuan, Xiaobing Hao and Hu Jingjing (2019), Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms. Security and Communication Networks (Hindawi), Vol (2019), 1-11.

[28] Nachiket Sainis, Durgesh Srivastava and Rajeshwar Singh (2018), Feature Classification and Outlier Detection to Increased Accuracy in Intrusion Detection System. International Journal of Applied Research, 13(10), 4249-7255.

[29] Aditya Vikram Mohana (2020), Anomaly detection in Network Traffic Using Unsupervised Machine Learning Approach. IEEE Conference on Communication and Electronics Systems, 476-479.

[30] Faheem Masoodi, Alwi M. Bamhdi and Tawseef A. Teli (2021), Machine Learning for Classification analysis of Intrusion Detection on NSL-KDD Dataset. Turkish Journal of Computer and Mathematics Education, 12(10), 2286-2293.

[31] Cosimo Ieracitano, Ahsan Adeel, Mandar Gogate, Kia Dashtipour, Francesco Carlo Morabito, Hadi Larijani, Ali Raza, and Amir Hussain (2018), Statistical Analysis Driven Optimized Deep Learning System for Intrusion Detection. International Conference on Brain Inspired Cognitive Systems (Springer), 759-769.

[32] Lan Liu, Pengcheng Wang, Jun Lin and Langzhhou Liu (2021), Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning, IEEE Access, 09, 7550-7563.

[33] James Henrydoss, Steve Cruz, Ethan M. Rudd, Manuel Gunther, and Terrance E. Boult (2017), Incremental Open Set Intrusion Recognition Using Extreme Value Machine. IEEE International Conference on Machine Learning and Applications, 1089-1093.

[34] Mangayarkarasi Ramaiah, Vanmathi Chandrasekaran, Vinayakumar Ravi and Neeraj Kumar (2021), An intrusion detection system using optimized deep neural network architecture. Transactions on Emerging Telecommunications Technologies, 32(4), 1-17.

[35] Wen Xu, Julian Jang-Jaccard, Ampardeep Singh, Yuanyuan Wei and Fariza Sabrina (2021). Improving Performance of Autoencoder-Based Network Anomaly Detection on NSL-KDD Dataset. IEEE Access, 09, 140136-140146

[36] Ahmed I. Saleh, Fatma M. Talaat, Labib M. Labib (2019), A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers. Artificial Intelligence Review (Springer), 51, 403-443

[37] Monika Vishwakarma, Nishtha Kesswani (2023), A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection. Decision Analytics Journal (Elsevier), 07,1-8.