International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Machine Learning-Based Security Enhancement in Heterogeneous Networks Using an Effective Pattern Mining Framework

Manohar Srinivasan[1], Senthilkumar N. C.*[2]

Submitted: 30/06/2023 **Revised**: 07/08/2023 **Accepted**: 30/08/2023

Abstract: The world becomes an internet-based system as a result of the Internet of Things. IoT benefits people greatly in many different ways. The world is growing more technologically adept, yet a lot of problems are also concurrently emerging. Security is a primary priority considering that everything is connected. Occasionally, everyone hopes that their data will be transmitted safely over the internet. Currently, the rise of security concerns is proportionate to the development of technology. Due to the large number of vulnerable devices, IoT technologies are insecure and unreliable. IoT follows unique regulations and procedures, so conventional security measures cannot be applied. IoT security concerns need to be addressed on numerous fronts. Internet seclusion, wireless security, and privacy protection are only a few of them. Network intrusion and detection is a key topic of study in addition to the problems mentioned above. Due to the tens of thousands of networked devices in the IoT, it is extremely difficult to identify unusual access, unanticipated attacks, or odd device activity. In order to reduce security risks and detect intrusions in the Internet of Things, a number of tactics and algorithms have been proposed. Machine learning-based intrusion detection systems have demonstrated excellent accuracy and efficiency in recent years at spotting intrusions. This research offers a novel approach for boosting data transmission security, identifying attack-capable devices, and identifying devices that have been infiltrated by an intruder. The suggested approach uses Autoencoder (AE), a technique for machine learning that uses feature extraction along with principal component analysis. The website CloudStor is where you can find the intrusion and detection dataset. One unsupervised machine learning approach that effectively trains data is the autoencoder. The result shows that, when compared to other machine learning techniques, the suggested strategy produces better results.

Keywords: *Autoencoder, Intrusion, and Detection System, Internet Of Things*.

1. Introduction

Internet of Things plays a crucial role in the rapid growth of information and communication technologies. This sequence of events is referred to as the Internet of Things. It consists of a collection of devices or sensor nodes that are interconnected via a network and capable of producing and transferring data over the internet (IoT). In IoT, human involvement is not necessary. When devices are tagged, systems can monitor, inventory, and manage the devices. Later, it was developed into technologies such as barcodes, QR codes, etc. 2013 marks the deployment of the Internet of Things into the systems utilizing a variety of technologies. Simply said, the Internet of Things consists of all internet-connected gadgets. For instance, a Blood Pressure Monitor implanted in a person may transfer data to the doctor through the internet or inform the individual when their blood pressure is high or low. The expansion of the Internet of Things increases annually. The technology of the Internet of Things may differ from device to device and depend on the applications the IoT is designed to execute. The Internet of Things allows individuals to control smart devices and automates the whole house. Internet of Things is not limited to residential applications. IoT applications may be categorized as consumer, commercial, industrial, and infrastructure. IoT has aided individuals in several ways. IOT in organizational applications include IoMT (Internet for Medical and health utilization), transportation, vehicle to everything communications (V2X), and IOT in industrial applications are also known as IIoT (Industrial Internet of Things) as well as Manufacturing, Agriculture. IoT infrastructure applications include large-scale installations in metropolitan areas, energy management, environmental monitoring, and living labs. IoT military applications are known as IoMT (Internet of Military Things) and are utilized for surveillance and reconnaissance. IoBT (Internet of Battlefield Things) is a component of IoMT, Ocean of Things, and product digitalization.

The architecture of the Internet of Things varies depending on the application, however the fundamental architecture of IoT is outlined in Fig. 1.

[1]*Research Scholar, School of Information Technology & Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India.*
ORCID ID: 0000-0003-1943-3503
[2]*School of Information Technology & Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India*
ORCID ID: 0000-0002-2050-1297
** Corresponding Author email: ncsenthilkumar@vit.ac.in*

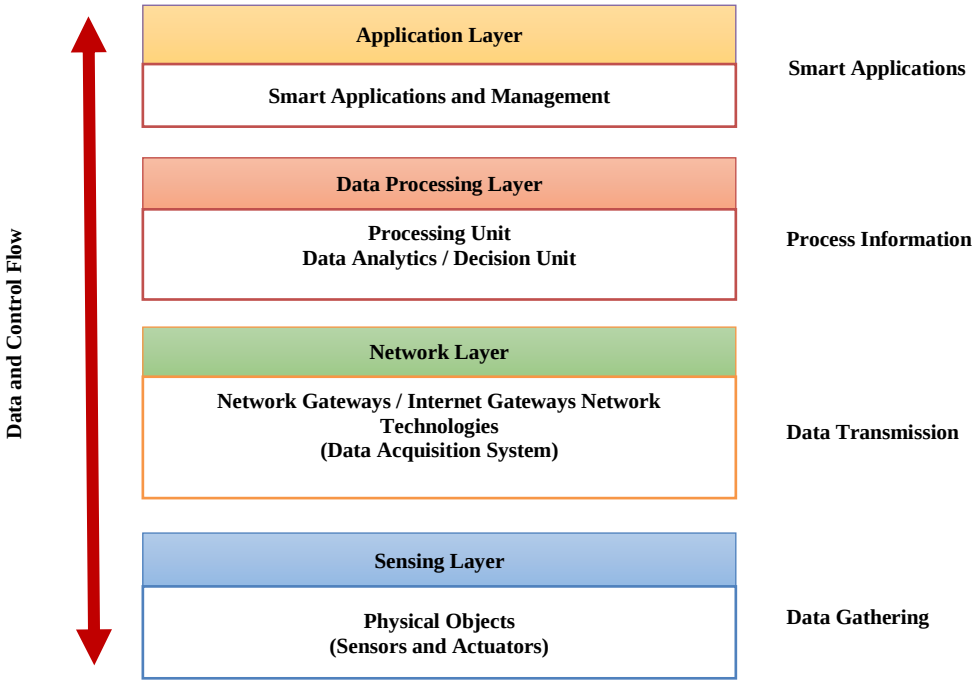

Fig. 1 Basic Architecture of IoT

Network Layer – This layer is responsible for aggregation and digitalization of information. Data Acquisition System performs both aggregations of data and conversion of analog data to digital data. In this layer, a huge amount of data or information is gathered from the previous layer and compress to the ideal size for the upcoming analysis. Advanced Gateways performs the basic functionalities such as filtering the data depend upon making the decision, protecting the malware etc.,

Data Processing Layer – Data Analyzation and Pre Processing the data is done in this layer using Edge IT or Edge Analytics. After the completion of data analysis and Pre Processing, the data moved to the data center, Datacenter – The data is monitored and regulated here at the same time, the data can be accessed by the applications.

In this last layer the data is to be undergone on analysis and managing the data and how to store the data. Data are collected and available in the data centers, where the data is managed and will be used by the end-users.

2. Challenges in IoT

The Internet of Things has been dealing with many sectors such as Medical, Information Technology, Industry, Commercial, Agriculture, etc., IoT has to cope up with the various challenges in many areas. Some of the key challenges of IoT are:

a) Scalability – Tons of devices are connected via the internet. Every day large volumes of data are produced. All the data needs to be stored, analyzed, and processed. So, IoT needs cloud storage to store the data.

b) Lack of Regulation- Due to the drastic evolution of technology, IoT is still at security risk because in IoT all the devices are connected to the Internet, for example, In-home all the devices such as smartphones, medical devices, toys, car, Air cooler are connected to the internet. If they are any security risk happens to that, it will lead to the serious cause to them.

c) Bandwidth – As the size of the IoT applications increases, there is also a demand for bandwidth. The current IoT Client-server model doesn't have much space to allocate the bandwidth. Here connectivity comes into a challenge.

d) Security – Security is a key concern in IoT applications. Because of the millions of nodes connected to the huge network, information security comes into the problem. It's hard to find system vulnerabilities is more and the huge information is shared between the devices this may lead to the hacker hacking the confidential information. So, Confidentiality, Integrity, and the Availability of the data should be ensured.

e) Challenge based on Design – Less computational power, Lower energy, and Low memory are the outcome when designing the technology.

f) Lack of Interoperability – There is not a common framework or standard in IoT, which makes the device in the lacking of interoperability. Legacy IoT devices affect more because the technology standards are splintered.

This research mainly concentrates on addressing potential privacy and security risks in an IoT-based environment. Initally, a Genetic algorithm is used to choose the most effective dataset characteristics for training the model. In

addition, soft voting classifiers are used to categorize the genetic algorithm-selected features. For this model, four voting classifiers were selected, and the one with the highest accuracy would be used for the following stage, known as Regularization. Finally, contractive Autoencoder is the Regularization approach used to this model to deducts the problem of overfitting to the dataset. The model is trained and evaluated using 10-fold cross-validation for experimental validation.

3. Background Study

Every node in a computer network exchanges data with every other node. So, each node has data, which may pertain to a neighboring node, must be transmitted to another node, and contains information about the node itself. Numerous methods and studies are being used to protect data from hackers. Network security is a crucial component of computer networks. The network is secure if it has the properties of secrecy, integrity, authentication, dependability, interoperability, and data availability. This article addresses the network security technologies. Information security, computer system security, network security, network information security, etc., are but a few of the means by which computer networks safeguard data. Various technologies are used to safeguard the network, and a few of them are described below. 1. Authentication – To determine whether or not the user is the authenticated user while sharing information with others. Authentication is comprised of message authentication, access authorization, and digital signatures. 2. Data Encryption — To achieve data privacy, data encryption is needed. The data have been encrypted using the two following methods: I Symmetric Key Encryption against ii) Asymmetric Key Encryption 3. Firewall - A barrier between an internal network and an external network. It prevents the external user from unofficially accessing the data. Proxy gateway, packet filtering, and application gateway technologies are used by the firewall to safeguard network access. 4. Intrusion Detection System - This system recognises the activity of network nodes and devices. It conducts security audits, attack identification and response, and device or node monitoring. This may be accomplished by the identification of abuse and abnormalities. 5. Antivirus – It protects the system's data from viruses, and users must maintain their antivirus software up-to-date to safeguard disc data. Virtual Private Network (VPN) is the last computer security technology. VPN is a private network used in the public network to safely send data via a virtual channel. [1]

In terms of network technology, security is a crucial element. Due to the widespread use of the internet, security is seen as a key obstacle. Currently, every individual has a mobile phone with a high-speed internet connection. Every day, a vast volume of information is transported globally across all industries. Therefore, security methods are essential to preserve the information's integrity, confidentiality, and accessibility. This article describes the many forms of network assaults that are feasible, as well as the security procedures that are implemented based on the network's design. Active attacks, passive attacks, distributed attacks, insider attacks, close-attacks, spyware attacks, phishing attacks, spoof attacks, hijack attack, and exploit attack are all capable of disrupting a network. Various security measures, such as Cryptographic systems, Firewalls, IDS, Anti Malware Software, SSL (Secure Sockets Layer), Dynamic Endpoint Modeling, and Mobile Biometrics, are used to fight against internet threats. By applying the security mechanism, we may avoid potential network assaults. Internet of Things is a heterogeneous network in which nodes are heavily linked. IoT consists of millions of interconnected nodes or devices for data flow from one device to another. IoT enhances everyone's lives as a result of advancing technology. The fact that millions of devices are networked without the proper standards and procedures raises serious questions about the security of the Internet of Things. IoT is dependent on the technology used, thus its standards and protocols vary. This article highlights the problems associated with threats in the Internet of Things. Four strategies were used to offer security for end-to-end communication in the Internet of Things and Wireless Sensor Network. Centralized Approaches, Alternative Delegation Architectures, Special Purpose Hardware Module Solutions, and Protocol-based Extensions and Optimizations. [3]

Network security is now the most essential need for emerging businesses. In defense-like regions, resource accessibility is crucial. The key components of information security are authentication, integrity, and confidentiality. This document describes the essential factors and actions required to construct a secure network inside a company. Wi-Fi provides intruders access to a wireless network's diverse resources and the ability to connect to a variety of devices, making it the most prevalent network intrusion method. This article examines the measures that must be performed to protect the different networks and give the company with a completely secure environment. The author of this paper describes the categories of attack types: active and passive attacks, provides some key measures to secure the network, tools that are used to secure the network, such as N-map, Nessus, Wire-shark, Snort, Net Cat, Kismet, etc., and follows some security methods such as cryptography, firewalls, and their types. [4]. Recently, significant suppliers have been providing complete network security to industries and organizations. The security measures must be devised and implemented. A business or organization is needed to assess its security requirements at various organizational levels. Design some security rules, which should be updated in the future based

on the circumstances, may be acceptable, and should be easily managed. Security policies should be of the flexible kind, not the rigid type, and they should be applied at several levels so that they meet the needs of an organization and can manage future security risks. [5]

The Intrusion Detection System guaranteed the infrastructure's security. It safeguards the organization's or industry's network. In 1980, Jim Anderson initiated the monitoring of user activity with the use of logs and computer records. Monitoring user activity is used to protect data from unauthorized users and wrongdoers. [6] [7] Intrusion deduction and expert system (IDES) is used to more effectively and efficiently handle assaults and threats. IDES is a real-time, independent IDS that blends semantic based with anomaly detection [8]. The primary objective of both systems is to identify an intruder in the system, although their underlying mechanisms are distinct [9]. Anomaly-based IDS allows the transaction that is regarded as typical behavior. All usual behaviors are preset and built into the system; if any behavior other than a predetermined one occurs, the system will halt immediately. Signature-based IDS systems are preconfigured with a list of actions that are deemed an attack; IDS allow all activities except those on the list of assaults. All known forms of assaults are stored in rules format inside the sensor. It will be useful for network traffic analysis (incoming and outgoing). Biermann et al. [10] claimed that dissimilarities between various systems were studied, and each had its limitations in identifying all types of incursion. Machine Learning techniques are used to improve the performance of intrusion detection systems (IDS) by increasing the precision of detection and decreasing the complexity of computation. Furthermore, compound intellect systems are used in IDS; for instance, the hybrid intelligent system merges decision tree (DT) and support vector machine (SVM) [11]. After detecting an intrusion, it is vital to notify the user. The accuracy of the attentiveness is determined by the parameters for detecting the intruder. IDS generate a large number of notifications. The majority of alarms, however, are false-positives, i.e., typical human behavior that the IDS interprets as an intrusion[12].

An intrusion Detection System offers protection against hostile activities for an organization. It recognizes network intruders and provides the user with information about their activities. There are several Intrusion Detection Systems available on the market. An Intrusion Detection System is a godsend for network managers, allowing them to investigate the passage of large data packets in the network and evaluate any abnormal network activity. However, research is ongoing to identify hackers in many networks with greater precision. This document describes the parameters used to calculate the IDS. IDS measurement variables include FP (False Positive), FN (False Negative), TP (True Positive), and TN (True Negative). Information leak or threat and intrusion are the forms of Intrusion Detection System[13]. The author of this article describes the various intrusion detection system methodologies. The strategies consist of various methods of intrusion detection using Data Mining Techniques [14] for Detecting Intrusion, such as Data Mining Algorithms for designing the Intrusion Detection System [15], IDS using K- means clustering algorithm[16], Improved DBSCAN for detecting the Intrusion[17], IDS based on K-Means Clustering and OneR Classification[18], and clustering algorithm for incongruity detection[19]. To limit network assaults by an intruder and to be able to identify anomalous incursions, the system should be updated with the most recent IDS technology. [20]

One of the programming approaches inspired by Charles Darwin's theories is the Genetic Algorithm. It is via the process of natural selection that the appropriate people are picked for reproduction in order to generate offspring[30]. Genetic Programming and Genetic Algorithms are used to accurately identify the many sorts of network breaches. Utilizing a Genetic Algorithm to determine the categorization criteria. [21][22]. GA is used to determine the essential traits and the best settings. Several IDS-related articles [26][27][28][29] attest to their influence on network security. Genetic Algorithms are used to solve a variety of issues, and the major influence on the algorithm's efficiency is determined by three factors: GA parameters, individual representation, and fitness function. This work describes the detection of infiltration using a genetic algorithm, and the dataset was obtained from KDD99. Intrusion detection datasets [23][24][25]. This system is divided into two main phases: 1. Pre-calculation 2. Identification In Pre-calculation, a set of chromosomes is constructed using training data. In Detection, the population for test data is produced, and then the assessment procedures are executed. The assessment procedure consists of selection, hybridization, and mutation. Predict the kind of test data. The chromosomal set is compared to the population to determine its optimal fitness. In this section, the Detection and False positive rates are computed. The detection rate is measured as the proportion of accurately detected intrusions to the total number of incursions. The false-positive rate is determined by comparing the total number of normal connections to the total number of normal connections that are incorrectly recognized as intrusions. Based on the table value, the detection rate (DR) = 0.9500 and False Positive Rate (FP) = 0.3046 [31]. Overall the performance of the IDS is better when compared to the other methodologies.

In Clustered Wireless Sensor Network, to detect the intrusion the Hybrid intrusion detection system is used. Support Vector Machine (SVM) and Misuse Detection are used in this approach. KDD99 intrusion detection dataset is

used [23][24][25]Hierarchical topology split the wireless sensor network into clusters. Every cluster has a Cluster Head (CH), each IDS node receives the data from the adjacent IDS nodes throughout the multi-hop communication. The distributed learning algorithm is used to train the SVM to detect abnormal and normal behavior with a high accuracy rate of over 98 %. Select the training model into a hybrid intrusion detection module (HIDMs) to acquire a detection system with more accurate. SVM classifier and Signature-based IDSattain the highest detection rate with a low false-positive rate. Based on the experimental results, the routing attacks can be identified with a low false alarm rate[32].

4. Materials and Methods

This segment describes the strategy of Contractive Auto Encoder in detail together with the extensive description of dataset and algorithms to examine the performance of distinctive intrusion detection on the sensor nodes or devices.

4.1 Dataset Description

To manifest the intrusion detection on sensor nodes or devices is kick off with examining the dataset. The dataset is fetched from the journal [32]. The testing environment for IoT Intrusion and Detection is a fusion of IoT systems and interlinking formation. SKT NGU and EZVIZ Wi-Fi camera are the two IoT devices used to generate the dataset. The devices are interconnected to the home Wi-Fi router and the rest of the devices can connect to the router. The target devices for hacking are SKT NGU and EZVIZ and other devices are considered to be assaulting devices. The IoT Intrusion and Detection dataset comprise 80 features of the network and three features for classification. The classification features are binary, class, and subclass. The below Table 1 and 2 have a detailed description of the data group.

Table 1. Features Classification

Binary	Class	Sub-Class
Normal	Normal	Normal
Anomaly	DoS	Syn Flooding
	MITM	ARP Spooling
	Mirai	HTTP Flooding, UDP Flooding, Brute Force
	Scan	Host Port, OS

Table 2. Dataset Description

List of Parameters	Descriptions
Sources	Ref [32]
Datatype	Continuous (Numerical)
Number of Instances	625785
Number of Features	83
Total Classes	5

The preprocessing steps are requisite to the dataset due to the format and data types of few features that are not appropriate for learning algorithms. Here, intelligent retrieval system, learning algorithms and column normalization techniques were used to standardize and examine the dataset. The grouped features may deteriorate the detecting ability of machine learning algorithms. So, the combined characteristics are eradicated in this IoT intrusion detection dataset.

4.2 Feature Extraction using Principal Component Analysis

An algorithm for the unsupervised linear transformation that collects features using statistical methods. The dataset is then projected onto a lower-dimensional space with a predetermined number of dimensions after it has been searched for the eigenvectors in a covariance matrix that have the greatest eigenvalues (features). These extracted characteristics form a collection known as the principle components, which does not include any correlations. PCA is sensitive to outliers and missing data; nonetheless, its goal is to minimize dimensionality in a way that does not sacrifice too much of the information that is either relevant or useful. The principal component analysis (PCA) approach that was constructed in this research makes use of the Singular Value Decomposition (SVD) solver. It is necessary to investigate a variety of dimensions in order to ascertain the impact of making changes to the input dimensions and locate the optimum number of extracted features to put to use.

4.3 Autoencoder

An Autoencoder is a fully connected network. Rather than training the autoencoder as a classifier, to produce the digits (0 to 9), trained the network to generate the pixels of the original data. The core objective of the autoencoder is to determine, how to effectively compress the data and encoded it, learn to reconstruct the data from the compressed encoded representation to the representation which is nearest to the input (original). It diminishes the dimensions of data by training the network to disregard the noise in the input data. In the architecture of autoencoder, it requires the bottleneck that might a symbolic representation of the original input.

The architecture of the autoencoder encloses 4 stages:

1. Encoder Layer – It reduces input data dimensions and compressed it into an encoded form.
2. Bottleneck Layer or Hidden Layer: It holds the compressed depiction of input data. This layer has the feasible dimensions of input data.
3. Decoded Layer: This layer performs the data reconstruction from the encoded form which is to be near to the original input data.
4. Reconstruction Loss: It measures the performance of the decoder and evaluates the output with the original input. Reconstruction loss would be a tedious task when the input data features are independent of each other.

The training of the network is done with the help of the backpropagation method to decrease the reconstruction loss of the network.

4.3.1 Parameters of Autoencoder:

The below parameters of Autoencoder must be set before the training gets initiated:

a) Size of the code - It depicts the total nodes present in the hidden layer. Compression is more when the code size is small.

b) The total number of layers- The layer is depended upon the user.

c) Total number of nodes each layer - The number of nodes at each layer decreases at each succeeding layer in the encoder and the more nodes increases at every succeeding layer in the decoder.

d) Loss function–Two loss function is used is autoencoder such as Binary cross-entropy and mean squared error, either one of them is used. Binary Cross entropy loss or log loss, evaluate the quality service of a classification model and the out value range from 0 and 1. Mean Squared Error is suitable for the regression problems due to high decision boundaries.

Let us take the input data as x, the output of the data as x', the network could be trained by reducing the reconstruction error as L (x, x'). Reconstruction error measures variation between original input data and subsequent reconstruction data. The hidden layer (bottleneck) is an essential element in the autoencoder network. The network can simply memorize the values of the input and transferring those values throughout the network without the presence of a bottleneck. The bottleneck restricts the amount of information traversing throughout the entire work to build the compression of the learned input data. The examples of the autoencoder attain from the following: Optimal feature subset to the inputs is sufficient to form a reconstruction precisely whereas imperfect feature subset to the inputs leads the training data to overfit.

4.3.2. Regularization in Autoencoder

A regularization is an approach by adding a further penalty term in the error function. It is used to tuning the function. It is in the regression from. It regularizes the coefficient evaluation towards 0. This approach prevents learning a highly complex model to avoid the overfitting risk.

4.4 Contractive Autoencoder

Contractive Autoencoder is an enhancement of autoencoder and it belongs to regularized autoencoders. Regularized autoencoder learns the dormant variables by adding some noise or penalty to the loss function. The core objective of Contractive Autoencoder is to have a robust learned representation. It is a low optimal feature subset to slight variations in the input data. By implementing the penalty term to the original cost or loss function, a robust learned representation of the data is made. In the Jacobian matrix, the Frobenius norm acts as a penalty term here. Placed on the input data the Frobenius norm for the middle layer (hidden layer) is evaluated. The resultant data after applying the penalty term provoke mapping which contracting with the data strongly.

Let's take the input image as x examined as a column vector, di is the size of the input layer, d_{hi} represents the total number of neurons present in the hidden layer. To encode the input image to encode into the hidden layer hi (bottleneck layer) use the below activation function.

Hi=sigmoid($Gx + b_v$), where G is the matrix of $d_{hi}*d_x$ where $d_{hi} < d_x$ and $b_v$is the bias vector which is represented as a $d_{hi}*1$ for the hidden layer. $(Gx + b_v)$ is the column vector, so the results of the output is a column vector of size

$d_x > 1$. To decoding the hidden layer to the output layer apply the below formula,

x_r= sigmoid (G^Thi + o); x_r– reconstruction
(1)

The resultant image of the autoencoder is the same as the input and o is the bias vector for the output layer, so o is d_x * 1. By making use of error propagation compute and minimizing the cost of reconstruction about G, b_v, o

$$\text{Loss}_{AE}=\frac{1}{N}\sum_{x\in S}^{n}(x - x_r)^2$$
(2)

Where, S is the compiled set of all input samples, N is the cardinality of set S and reconstruction loss function is a squared error function. The following is the equation of

the Frobenius norm of the Jacobian matrix, which is applied in Autoencoder.

$$\lambda \| J_{hi}(x) \|_F^2 \quad (3)$$

The above equation can be expanded as the below, the value of λ is considered as a 0.1

$$\| J_{hi}(x) \|_F^2 = \sum_{j=1}^{d_{hi}} \left(h_j \left(1 - h_j\right) \right)^2 \sum_{k=1}^{d_i} G_{jk}^2 \quad (4)$$

It will be rewrite as,

$$\| J_{hi}(x) \|_F^2 = \sum_{j=1}^{d_{hi}} \left(\left(1 + h_j\right)\left(1 - h_j\right) \right)^2 \sum_{k=1}^{d_i} G_{jk}^2 \quad (5)$$

Therefore,

$$\text{Loss}_{CAE} = \frac{1}{N} \sum_{x \in S}^{n} (x - x_r)^2 + \lambda \| J_{hi}(x) \|_F^2 \quad (6)$$

Applying 5th equation in 6th equation, to get the minimizing loss of CAE,

$$\text{Loss}_{CAE} = \frac{1}{N} \sum_{x \in S}^{n} (x - x_r)^2 + \sum_{j=1}^{d_{hi}} \left(\left(1 + h_j\right)\left(1 - h_j\right) \right)^2 \sum_{k=1}^{d_i} G_{jk}^2 \quad (7)$$

5. Proposed Method

The recognition of noteworthy features assists the overall performance of the model. To sort out the features, conceivable feature selection techniques are vital. To select the best optimal feature set selection has been done by Genetic Algorithm.A Genetic Algorithm is used to resolve the optimization problems.

Individual (Chromosome) Selection - Depends upon the good fitness value, the individual is selected.

Crossover Operator – Two individuals (chromosomes) are chosen by using a selection operator. The crossover operator selected some genes from the two individuals and produce the third individuals shown in Fig. 2. The crossover operator is of three types: One point, Multi-point, and Uniform crossover.

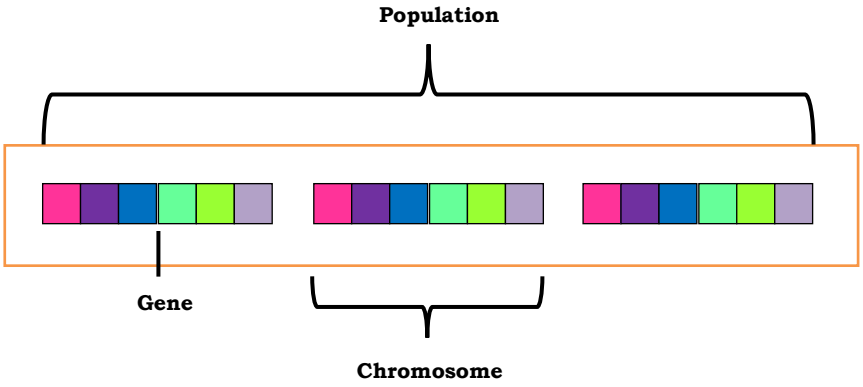

Fig. 2 Gene, Chromosome and Population

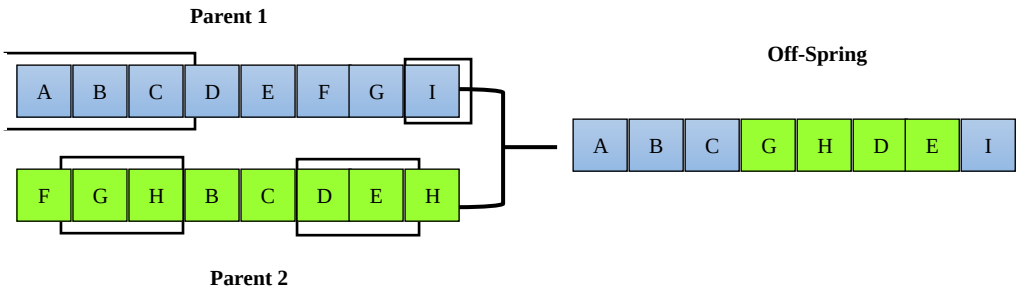

Fig. 3 Crossover Operation

Mutation Operator - Two individuals (chromosomes) are chosen by using a selection operator and perform the gene changes randomly to generate a new chromosome shown in Fig. 3 The mutation operator is of different types such as Bit-Flip, Swap Mutation, Scramble Mutation, and Inverse Mutation.

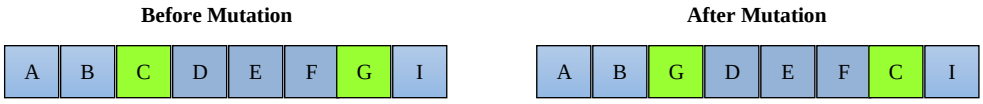

Fig. 4 Mutation Operation

Once, the above stages are done, the final step is to terminate the process. The termination of the genetic algorithm usually arises when the solution of the problem is near to the optimal solution. GA provides better results in every iteration but when reaches the subsequent stages the improvements are comparatively less. Genetic Algorithm terminates the process when it reaches one of the below conditions – 1. No improvements in the population iterations. 2. When it hit an entire no of generations.

5.1 Voting Classifier

Voting Classifier is also called Ensemble Learning. In a single set, various models are working together as a collection is called Ensemble Learning. Instead of using one model, using different models produce a better result. Voting is the easiest way of fusing the predictions from various algorithms of machine learning. Set of different algorithms are need to be trained and evaluated in collateral to perceive the various weirdness of each algorithm. The data has to be trained with the help of various algorithms and orchestra to predict the final result. The final result depends upon the majority of the vote based upon the two distinct strategies: Hard Voting or Majority Voting, Soft Voting. Hard Voting – the class acquires the number of votes from that the maximum number of votes will be chosen. Soft Voting – Each predicted class has the probability vector for all the classifiers. Those vectors are summed up and averaged. The class which attains the highest value will be chosen up. Here, in this approach, the soft voting classifier is used to attain the output from the various classifiers during prediction.

5.2 Contractive Autoencoder

It is used to extract certain features from the predicted output of the voting classifier. It is also used for performing regularization on the dataset.

5.3 K - Fold Cross-Validation

To execute the models of machine learning with the minimum sample of data, cross-validation is used. It holds the single parameter value called ‘k’. k refers to the number of groups, the samples of data has to be split. Whatever the value of k is selected, that validation is called based upon the k value. For example, k =10 changed as 10 fold cross-validation. The below procedure describes the working nature of K – fold cross-validation.

Initially, the dataset has to be randomly shuffled and split up the dataset into k groups. Every group is regarded as a test data set, while the remaining groups are regarded as training data sets. Choose a model, use the training data set to test it, and then assess it on the test data set. Keep the evaluation score and ignore the model. Conclusively, the result is evaluated using performance evaluation. The performance evaluation has been done by the three metrics such as accuracy, precision, recall, and Log Loss. The resultant output of the performance evaluation will be considered as the predicted output and it will be the best solution to the problem.

Algorithm: Proposed Contractive Autoencoder Model

Population Initialization:

```
curr_gen :=1
        for p :=1 to n do
        for q := 1 to M do
                r[p][q] := rand_num(0,1)
```

Population Evaluation:

```
        for p :=1 to n do
            eval(r[i])
```

Selection:

```
        for p :=1 to n do
            parent[p] := tourn(r.tourn_size)
```

Crossover:

```
        for p :=1 to n do step 2 do
            for q := 1 to M do
if  (q <=cross_pt)
offspr[p][q]=parent[p][q]
offspr[p+1][q]=parent[p+1][q]
                else
offspr[p][q]=parent[p+1][q]
offspr[p+1][q]=parent[p][q]
```

Mutation:

for p :=1 to n do

if(rand_number<mutat_rate)

mutat (offspr[p])

Offspring Evaluation:

for p :=1 to n do

eval (offspr[p])

Generate new population:

r = offspr

curr_gen+ = 1

Termination Condition:

if (curr_gen<maximum_gen)

return to step 3

```
Y_train, Y_test, z_train, z_test = train_test_split(Y, Z, test_size = 0.20, rand_state = 42)

est = []

est.append(('LR', LogisticRegression(solver ='lbfgs', multi_class ='multinomial',

max_iter = 200)))

est.append(('SVC', SVC(gamma ='auto', prob = True)))

est.append(('DTC', DecisionTreeClassifier()))


# Voting Classifier

voting_soft = VotingClassifier(ests = est, voting ='soft')

voting_soft.fit(Y_train, z_train)

z_prediction = voting_soft.prediction(Y_test)

# Accuracy_score

score = acc_score(z_test, z_prediction)

Output Score;
```

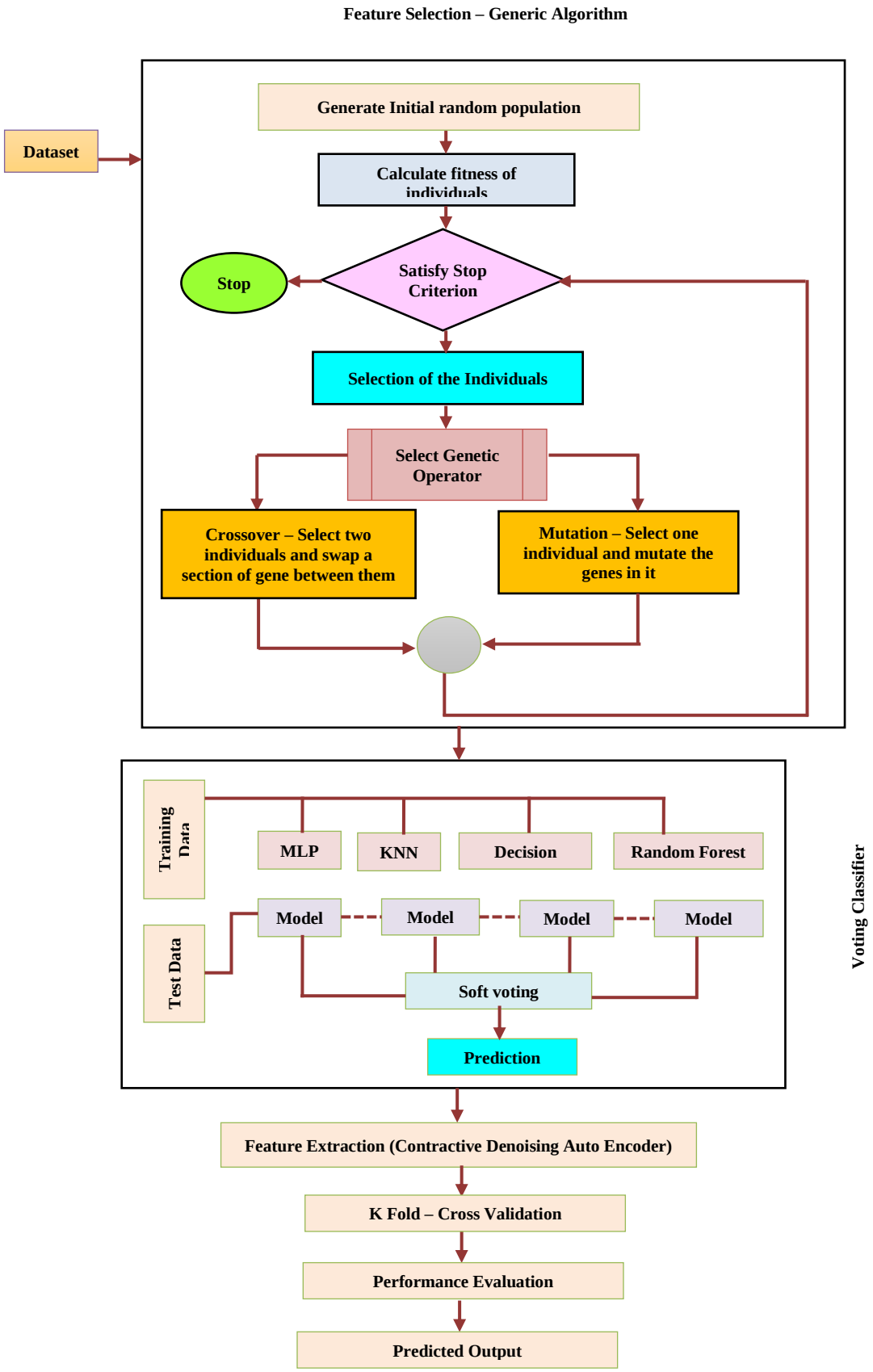

Fig. 4 Block Diagram of Proposed CAE

6. Results and Discussions

To identify the efficacy of the system, evaluating the model is an essential one to validate its noteworthy. This paper is concentrated on identifying the intruders on the sensor network. The simulation's parameter settings are listed below. The Network Simulator 3 (NS3) is used to simulate the projected models. The first step involves measuring mobile duration variations, and the second is when speed alterations are made. The attribute metrics used in the simulation platform are shown in Table 3.

Table 3 Parameter setup

Parameters	Data
Node counted	55, 110, 220
Target area	450*450m
Simulation prior	50-150 s
Pause time	5-20 s
Maximum speed rate	5-25 mps
Transmission radius	30-300 m

This paper is composed of three vital approaches to procure the expected outcome. The previous section describes the entire working mechanism of the system. This segment scrutinizes testing the consequences of the system. It comparing the outcome with the various classifiers such as BayesNet, Naive Bayes, Random Forest, Decision Tree, Logistics Regression, and Support Vector Machines. Five performance metrics were examined to get the overall performance of the system with various classifiers. They are Classification Accuracy, Precision, Recall, F – Score, and ROC (Receiver Operating Characteristics Curve).

Classification Accuracy - Total number of correct predictions from all the predictions.

Precision - Total number of positive occurrences (in terms of %) from the total number of predictions of positive occurrence.

Recall - Total number of positive occurrences (in terms of %) from the total number of actual positive occurrences.

F –measure is used to test the accuracy with the help of precision and recall. The harmonic mean of precision and recall can be called as F1 Score.

ROC – Receiver Operating Characteristic Curve – It is a probability curve (graph), it demonstrates the effectiveness of classification models at several thresholds.

Table 4 Performance Metrics (%) of various classifiers before CAE

Classifier	Precision	Recall	F - Score
Logistic Regression	86.87	78.49	74.71
Naïve Bayes	87.75	79.69	73.65
BayesNet	97.39	97.11	97.11
Random Forest	96.77	96.88	96.87
Decision Tree	97.68	97.39	97.41
Support Vector Machines	98.67	79.49	73.29

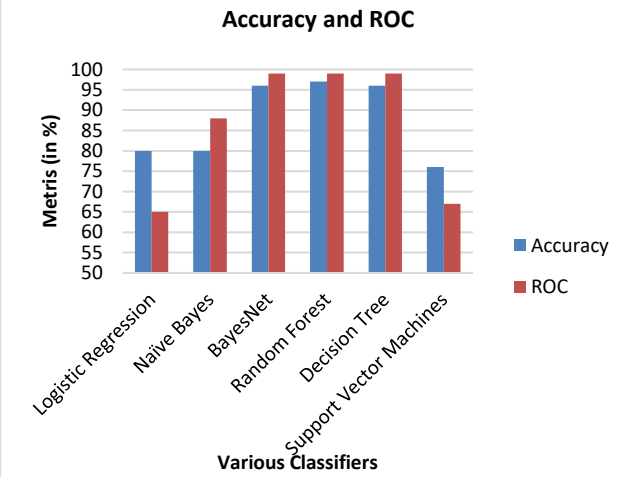

Fig. 5 Comparison of Accuracy and ROC before compression

Table 4 and Fig. 5 illustrate the performance of the model without the implementation of Contractive AutoEncoder. Among all the features from the dataset. Genetic Algorithm selects the appropriate features and to train the dataset, different classifiers were introduced and performed the soft voting to get predicted outcome. Among several features, 9 features are extracted from the dataset. By evaluating the performance model for those 9 features generate the above results. Based on the experimental outcome, the Decision tree classifier performs well in all the categories of the performance evaluation model. It attains an overall 96% in all the evaluation models, through this result, the decision tree works well among all the classifiers.

Table 5 Performance Metrics (%) of various classifiers after CAE

Classification Models	Precision	Recall	F - Score
BayesNet	97.09	97.1	97.1
Naïve Bayes	99.54	82.31	86.74

Logistic Regression	93.02	82.41	87.07
Random Forest	97.21	97.21	97.21
Decision Tree	97.69	97.29	97.39
Support Vector Machines	98.89	82.29	86.79

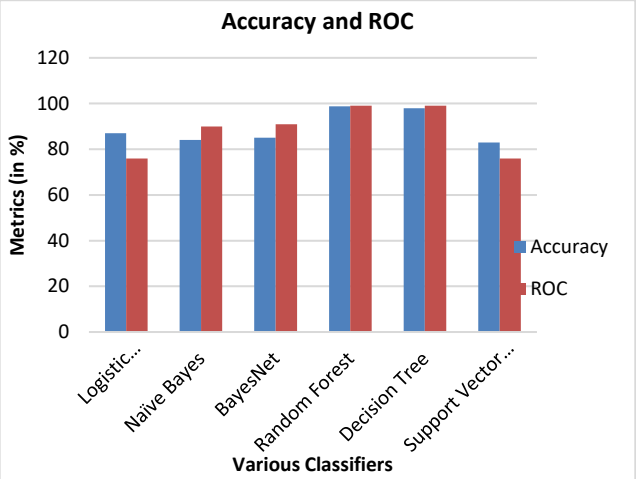

Fig. 6 Comparison of Accuracy and ROC after compression

Table 5 and Fig. 6 depicts the effectiveness of using Contractive AutoEncoder. Contractive AutoEncoder is used to compress the input data and encode it. GA selects the features from the dataset and train the data accordingly, perform the soft voting classifier to predict the final result with 9 features. Contractive AutoEncoder compresses those 9 features and makes them as 4 feature vectors. K – fold cross-validation is done on 4 feature vectors and evaluates the performance model on it. Based on the outcome of the experiment, decision tree classifiers outperformed with 97% when compared to other classifiers. By using Contractive AutoEncoder, the effectiveness of the system increased by 1% when compared to the existing model. In network security, the performance evaluation model is a vital one. Moreover, the performance of the proposed algorithm is benchmarked with the existing swarm algorithms. Based on the results the proposed algorithm performed well in terms of model accuracy. 1 Based on the correct predictions only, the system identifies the intruder on the sensor network. Identifying the intruder on the WSN is a key part because every person's datais stored on the network and communication is also done on the network. So, to safeguard the data, every person needs awareness about hackers and different types of hackings.

7. Conclusion

This research focuses on identifying potential privacy and security risks in an IoT-based environment. Currently, Internet of Things is an unavoidable technological advancement, as its ability to minimize human effort by linking electrical equipment and accessories for ease of use and security makes it a need for living a normal life in this contemporary world. In this research, a genetic algorithm is used to choose the most effective dataset characteristics for training the model. Soft voting classifiers are used to categorize the genetic algorithm-selected features. For this model, four voting classifiers were selected, and the one with the highest accuracy would be used for the following stage, known as Regularization. Contractive Autoencoder is the Regularization approach used to this model detects the issue of overfitting to the training dataset. Additional 10-fold cross-validation is used to train and assess the model. Various strategies achieve high levels of accuracy (97.34%) and precision (97.40%), making them more effective than current models. This model efficiently identifies intruders based on the findings obtained from performance measurements. In the future, it will be necessary to subcategorize the hacking kinds in order to identify the intruder on the wireless sensor network with greater precision and accuracy. It makes the system more secure and safe in terms of privacy because the majority of IoT devices contain information about the individual, such as medical details, home appliances, etc., and if it is hacked by cybercriminals, the data may be modified or deleted, causing serious problems for the individuals. Consequently, conducting the sub-categorization of the numerous hacking categories may make it easier for the system to detect intruders and more effectively identify the dangers that fall under which categories.

References:

[1] Fan Yan1Yang Jian-wen2, Cheng Lin1, "Computer Network Security and Technology Research," 2015.

[2] Monali S. Gaigole, Prof. M. A. Kalyankar, "The Study of Network Security with Its Penetrating Attacks and Possible Security Mechanisms," May 2015.

[3] Muhammad A. Iqbal, OladiranG.Olaleye and Magdy A. Bayoumi, "A Review on Internet of Things (Iot): Security and Privacy Requirements and the Solution Approaches," 2016.

[4] Farrow, R. "Network Security Tools," found at http:// sageweb.sage.org/pubs/whitepapers/ farrow.pdf .

[5] Shailja Pandey, "Modern Network Security: Issues And Challenges," Vol. 3 No. 5 May 2011.

[6] Lunt, T.F., "Automated audit trail analysis and intrusion detection: A survey," In: 11th National Computer Security Conference, 1988.

[7] Dewan, M.F., Mohammad, Z.R., "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm,". Journal of Computers, 2010.

[8] Lunt, T.F., Tamaru, A., Gilham, F., Jagannathan, R., Jalali, C., Neumann, P.G., Javitz, H.S., Valdes, A., Garvey, T.D., "A Real-Time Intrusion-Detection Expert System (IDES). Final Technical Report," SRI International, 1992.

[9] Kabiri, P., Ghorbani, A., "Research on intrusion Detection and Response: A Survey. International Journal of Network Security," 1(2), 84–102, 2005.

[10] Biermanm, E., Cloete, E., Venter, L.M, "A comparison of Intrusion Detection systems," 2001.

[11] Quinlan, J.R, " Induction of Decision Trees", Kluwer Academic Publishers, Boston, 1986.

[12] Ashara Banu Mohamed, NorbikBashah Idris, and Bharanidharan Shanmugum, "A Brief Introduction to Intrusion Detection System," 2012.

[13] Sanoop Mallissery , Jeevan Prabhu,and Raghavendra Ganiga, "Survey on intrusion detection method",2011.

[14] Deepthy K Denatious& 2Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection",2012.

[15] Changxin Song, Ke Ma Institute of Computer Information & Technology of Qinghai Normal University Network Center of Qinghai Normal University Qinghai, China., "Design of Intrusion Detection System Based on Data Mining Algorithm",2009.

[16] Li Tian1, Wang Jianwen1 Department of Computer Science, North China Electric Power University (NCEPU), Baoding 071003, China, "Research on Network Intrusion Detection System Based on Improved K-means Clustering Algorithm",2009

[17] Li Xue-yong, Gao Guo- "A New Intrusion Detection Method Based on Improved DBSCAN",2010.

[18] Z. Muda, W. Yassin, M.N. Sulaiman and N.I.Udzir, "Intrusion Detection based on K-Means Clustering andOneR Classification",2011.

[19] Zhou Mingqiang, HuangHui, WangQian, "A Graph-based Clustering Algorithm for Anomaly Intrusion Detection", 2012

[20] Rachna kulhare, Dr. Divakar Singh, "Survey paper on intrusion detection techniques," 2013.

[21] Chittur, "Model Generation for an Intrusion Detection System Using Genetic Algorithms", January 2005.

[22] W. Li, "Using Genetic Algorithm for Network Intrusion Detection". "A Genetic Algorithm

[23] Approach to Network Intrusion Detection". SANS Institute, USA, 2004.

[24] KDD-CUP-99 Task Description; http://kdd.ics.uci.edu/databases/kddcup99/task.html

[25] KDD Cup 1999: Tasks; http://www.kdd.org/kddcup/index.php?section=1999&method=task

[26] KDD Cup 1999: Data; http://www.kdd.org/kddcup/index.php?section=1999&method=data

[27] H. G. Kayacık, A. N. Zincir-Heywood, M. I. Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets", May 2005.

[28] Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham, "Intrusion detection using an ensemble of intelligent paradigms", Journal of Network and Computer Applications, Volume 28, Issue 2, April 2005, Pages 167-182

[29] S. Peddabachigari, Ajith Abraham, C. Grosan, J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems", Journal of Network and Computer Applications, Volume 30, Issue 1, January 2007, Pages 114–132

[30] M. Saniee Abadeh, J. Habibi, C. Lucas, "Intrusion detection using a fuzzy genetics-based learning algorithm", Journal of Network and Computer Applications, Volume 30, Issue 1, January 2007,Pages 414-428

[31] Tao Peng, C. Leckie, Kotagiri Ramamohanarao, "Information sharing for distributed intrusion detection systems", Journal of Network and Computer Applications, Volume 30, Issue 3, August 2007, Pages 877–899.

[32] V. Bobor, "Efficient Intrusion Detection System Architecture ,Based on Neural Networks and Genetic Algorithms", Department of Computer and Systems Sciences, Stockholm University / Royal Institute of Technology, KTH/DSV, 2006.

[33] Mohammad Sazzadul Hoque, Md. Abdul Mukit, Md. Abu Naser Bikas, "An Implementation Of Intrusion Detection System Using Genetic Algorithm," 2012.

[34] Ullah I., Mahmoud Q.H. (2020) A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks. In: Goutte C., Zhu X. (eds) Advances in Artificial Intelligence. Canadian AI 2020. Lecture Notes in Computer Science, vol 12109. Springer, Cham. https://doi.org/10.1007/978-3-030-47358-7_52.

[35] HichemSedjelmaci, Mohamed Feham, "Novel Hybrid Intrusion Detection System For Clustered Wireless Sensor Network," 2011.

[36] Diniesh, V. C. ., Prasad, L. V. R. C. ., Bharathi , R. J. ., Selvarani, A., Theresa, W. G. ., Sumathi, R. ., & Dhanalakshmi, G. . (2023). Performance Evaluation of Energy Efficient Optimized Routing Protocol for WBANs Using PSO Protocol. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 116–121. https://doi.org/10.17762/ijritcc.v11i4s.6314

[37] Pise, D. P. . (2021). Bot Net Detection for Social Media Using Segmentation with Classification Using

Deep Learning Architecture. Research Journal of Computer Systems and Engineering, 2(1), 11:15. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/13

[38] Kothandaraman, D., Praveena, N., Varadarajkumar, K., Madhav Rao, B., Dhabliya, D., Satla, S., & Abera, W. (2022). Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science and Technology, 2022 doi:10.1155/2022/5086622

Manohar Srinivasan received his Bachelor of Computer Science and Master of Computer Science from University of Madras in 1997 and 2002 respectively and he received his Master of Computer Science and Engineering from Anna University in Chennai 2008. He has worked at various esteemed institutions in India and Abroad, He has 14 years of teaching experience in national and international universities. Now he is currently pursuing Ph.D. in Computer Science and Engineering at VIT, Vellore, India. His current area of interest is network security, WSN and IIOT.

Dr. Senthilkumar N C is working as Associate Professor in School of Information Technology and Engineering Department in VIT University, Vellore. He has 20 years of teaching experience and holds PhD degree in CSE after his ME and BE. His research interest includes web mining, network security and data analytics. His research area include web mining, Data analytics, network security, WSN and IIOT.