International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Two-Phase Feature Selection Technique using Information Gain and XGBoost-RFE for NIDS

Mohammed Sayeeduddin Habeeb*[1], Tummala Ranga Babu [2]

Submitted: 29/11/2023 **Revised**: 09/01/2024 **Accepted**: 19/01/2024

Abstract: Many interconnected devices in Internet of Things (IoT) networks result in complicated and high-dimensional data. To protect this high-dimensional data, efficient and effective security is required. Network intrusion detection systems (NIDS) are important in securing IoT networks from unauthorized access, anomalies, and zero-day attacks. However, NIDS has a major issue because of the high dimensional dataset created by IoT devices, analyzing all these features from the dataset results in an increase in system complexity and compromises the detection accuracy, so we need an effective feature reduction technique. This paper addresses this issue by proposing a novel two-phase feature selection technique. In the first phase, the Information Gain (IG) is used to rank the features based on the information contained in each feature of the dataset this results in narrowing the feature space while improving computational complexity. The rest of the feature subset goes through to XGBoost with Recursive Feature Elimination (XGBoost-RFE) in the second phase. The least important features are eliminated in each iteration by XGBoost, a gradient-boosting algorithm to evaluate feature relevance continually. This iteration is continuous until we get optimal features for NIDS. These selected features are given to deep learning specifically to the deep neural network (DNN). Comparative analysis is done with other deep learning approaches using the BOT-IoT 2020 dataset. Experimental results show an improvement in model accuracy of 99.8% and reduced FAR to 0.000 with 16 features selected from the dataset, we compared the results with the well-known DL model to check the effectiveness of our proposed model.

Keywords: *NIDS, Gradient boosting (XGBoost), Recursive Feature Elimination (RFE), IoT, and deep neural network (DNN).*

1. Introduction

The rapid growth of IoT devices has resulted in drastic changes in the way to communicate with our surroundings. IoT is a large number of networks interconnected with different sensors or devices that are capable of collecting and exchanging sensitive data on their own. These IoT devices are used in a wide range of applications and offer to improve quality of life, simplicity, and efficiency in day-to-day life, this IoT is mostly widely used in smart home automation, smart cities, automating industries, agriculture, and smart health monitoring as shown in figure 1 [1]. The interconnected IoT devices have exponentially increased in recent times which results in a wide range of exchange of sensitive data in a network, protecting this sensitive data from any hacker is a major security challenge [2]. Different security risks have arisen due to the increase in several IoT devices and users, each of which has special risks. These challenges include problems like data privacy and the possibility of data manipulation, illegal access, and other types of security risks. Strong security is required to safeguard and protect user privacy and data since IoT networks are complex and devices vary strongly in terms of connectivity [3].

Antivirus, firewalls, and authentication techniques are used as the primary or first line of defense to protect from any intrusion. To add more security intrusion detection system is placed as the second line of defense from any unauthorized activity in the IoT network. Network intrusion detection system (NIDS) continuously monitors the network traffic from different sensor nodes. NIDS detects and stops any unauthorized activity in the network, safeguards the network from any malicious activity, and calcsilicates as normal or attack [4]. To reduce security risks NIDS plays an important role in protecting IoT networks from different and new emerging or mutation of old types of attacks. The increase in the computational complexity and False Alarm Rate (FAR) in identifying zero-day anomalies is the primary issue with the current IDSs. In recent times, researchers have come up with a study to reduce the FAR for NIDS and improve detection accuracy using the deep learning (DL) and machine learning (ML) approaches [5]

[1] Research Scholar, Department of Electronics and Communication Engineering, University College of Engineering, Acharya Nagarjuna University, Andhra Pradesh, India.
ORCID ID: https://orcid.org/0000-0002-4857-8835
[2] Dept. of Electronics & Communication Engineering, R.V.R. & J.C.College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, INDIA
ORCID ID: https://orcid.org/0000-0002-4946-1452
** Corresponding Author Email: msayeeduddinhabeeb@gmail.com*

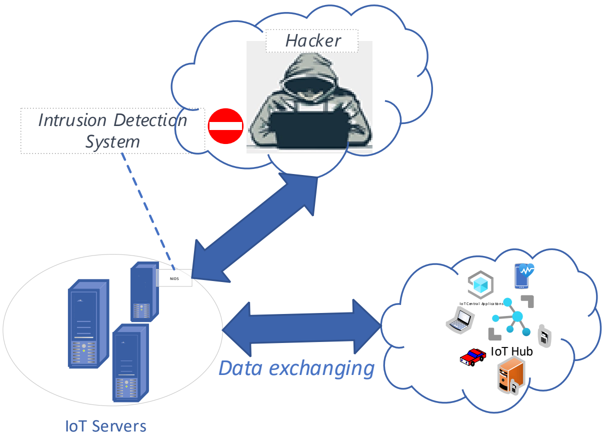

Fig 1 Intrusion detection system in IoT network

In a recent study, researchers showed that both ML and DL techniques can effectively classify the anomaly and benign from the pattern in the network traffic [6]. DL still excels due to its deep architecture and learning pattern without any human involvement, focusing on how crucial it is to use it in NIDS in IoT networks. One of the important DL approaches is Deep neural networks (DNN) have drawn a lot of interest from researchers in network security [7]. DNNs have shown good performance in these domains because they can learn complex patterns due to their multiple layers and can make correct predictions [8]. Due to the large amount of data generated by IoT devices, these characteristics of DNN have made it the most preferable approach to be used for an IDS created for an IoT network [9]. In this paper, we concentrate on the potential use of DNN to suggest a productive NIDS solution in the context of IoT

The process of feature selection is carried out due to two primary reasons, Firstly IoT devices generate large amounts of data which increases computation complexity. Secondly, an overbalanced of features in the dataset results in increased dimensionality which affects the performance of the intrusion detection system. This paper proposes a novel two-phase feature selection method called XGBoost with Recursive Feature Elimination (XGBoost-RFE), which makes use of Information Gain (IG) and XGBoost. This method aims to solve the problems that feature selection in IoT-based NIDS presents. In phase one IG is used in the initial stage to find the most relevant features for intrusion detection. In the second phase of feature selection XGBoost-RFE is used to remove most redundant or less significant features, thus further refining the feature set for better detection accuracy. This paper aims to improve the feature selection process in NIDS for IoT networks, which will result in increased detection accuracy and reduced FAR. The following are the major contributions made by this work.

- A literature on feature selection methods in the context of IoT-based NIDS.
- The introduction of a two-phase feature selection technique that combines IG and XGBoost-RFE.
- Experimental results and comparative analysis to evaluate the proposed technique's performance.

The paper is organized as follows: Section 2 gives a brief literature review related to NIDS, feature selection, IG, and XGBoost. the proposed methodology is given in section 3 which includes data collection, data preprocessing, and the two-phase feature selection algorithm. Section 4 describes the experimental setup. Section 5 presents the results and discusses their implication followed by a conclusion.

2. Related Work

This section provides a review of current studies in the fields of feature selection, deep learning, and machine learning that relate to network intrusion detection systems (NIDS) in the Internet of Things (IoT). Many studies have been carried out on the topic of NIDS which uses machine learning. IoT networks have been protected from anomaly detection using traditional machine-learning techniques. Garcia-Teodoro et al. looked into many ML techniques for NIDS, including Naive Bayes, Random Forests (RF), Decision Trees (DT), and Support Vector Machines (SVM) [10]. NIDS has improved significantly as a result of deep learning's growing popularity. Both long short-term memory (LSTM) networks and recurrent neural networks (RNNs) can identify sequential relationships in network data, making them viable options for NIDS. IoT network abnormalities have been identified using RNN-based techniques [11]. By automatically extracting important and most relevant features from the dataset. Different DL approaches for intrusion detection have been explored for their performance. To detect network anomalies, Kim et al. [12] suggested using Deep Neural Networks (DNNs). They demonstrated in their study how well DL models can extract complex patterns from network traffic data. The significance of feature selection in improving NIDS performance was presented in their study [13].

One of the important elements of NIDS is the process of feature selection, especially in the high-dimensional dataset received from IoT networks. Different feature selection techniques, such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), Mutual Information (MI), Chi-squared, and Information Gain (IG), have been investigated by the author in this study [14]. A detailed analysis of feature selection techniques for NIDS and their effect on accuracy was carried out by Lin et al. [15]. The significance of choosing pertinent traits to improve NIDS performance was highlighted by their findings. Attention has been drawn to hybrid methods that combine ML/DL with feature selection methodologies. A Hybrid Intrusion Detection System (HIDS) that combines feature selection and ML techniques was presented by Khan. [16]. By using

ML for classification and choosing the most useful characteristics, our hybrid technique demonstrated increased NIDS accuracy.

NIDS IoT networks are changing rapidly, with an increase in focus on a combination of feature selection with DL. The feature selection process solves the problem of a high-dimensional dataset [17][18]. After a thorough analysis of previous works, a framework to solve the issue of a high-dimensional dataset is proposed in this study. We proposed a two-step feature selection method for successfully detecting intrusions in IoT networks by reducing the computational complexity.

3. Proposed Methodology

We proposed a phase feature selection technique that focuses on challenges caused by high dimensional datasets received from various IoT sensors, processing this dataset requires more computational complexity. In our proposed features selection, we combine the two techniques first is information gain (IG) for raking the feature in a dataset with Recursive Feature Elimination (XGBoost-RFE) and XGBoost to find the most appropriate features from the dataset. Our proposed model as shown in Figure 2 is designed to remove redundant data from the dataset and extract the more informative features in the dataset.

3.1. Data preprocessing

To ensure the reliability and accuracy of the data used for training and testing in NIDS for IoT networks, the dataset preprocessing step is important. In this paper, we ensure consistency in feature scaling, in the first stage the standardization of data is done using Z-score normalization and handling missing data with advanced stopping techniques. We use the Synthetic Minority Over-sampling Technique (SMOTE) to address the common problem of class imbalance in NIDS datasets [19]. Moreover, in this proposed approach we added controlled variability to the dataset through data augmentation. Our suggested two-phase feature selection methodology, which uses XGBoost with Recursive Feature Elimination (XGBoost-RFE) and Information Gain (IG) to prepare the data for deep neural network (DNN) implementation, would not be possible without these carefully designed preprocessing steps.

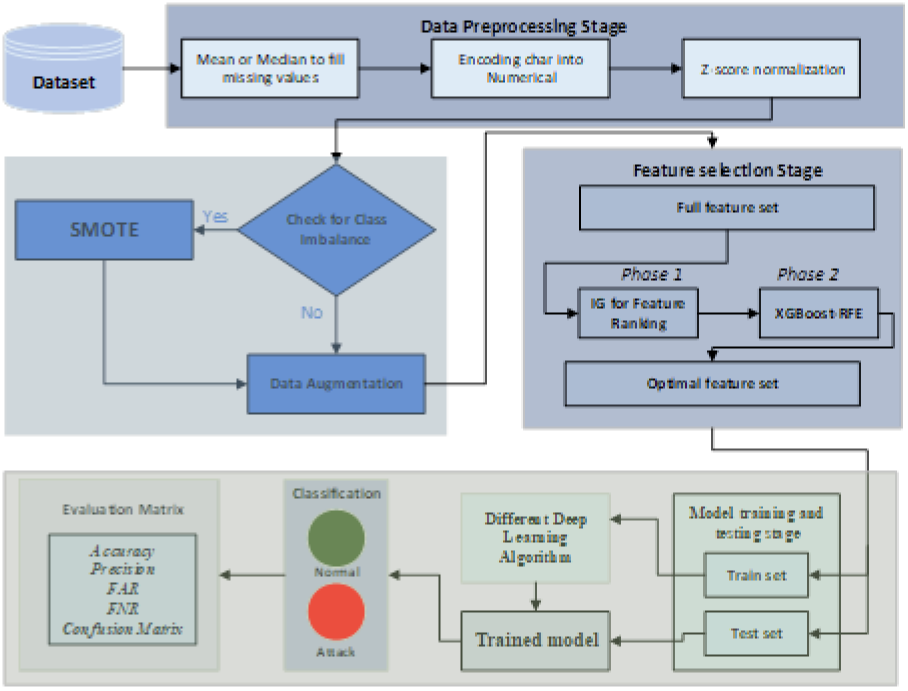

Fig 2 Proposed Methodology Framework

3.2. Phase 1: Information Gain (IG) for Feature Ranking

We use Information Gain (IG) as a feature selection method in the first stage. Each feature's information content is quantified by IG, allowing us to rank the features according to their significance. For NIDS, features with higher IG scores are considered to be more informative, whereas features with lower values are seen to be less important. Information Gain (IG) evaluates the decrease in entropy or uncertainty when a feature is used to divide a dataset. For feature selection, it is often used in feature ranking. Entropy serves as a framework for the IG, entropy on the dataset measures the data instability or impurities. The impurity of the entire dataset is measured by entropy(D).

$$Entropy(D) = -\sum_{i=1}^{c} p(i \mid D) * log_2(p(i \mid D)) \qquad (1)$$

Where D is the entire dataset with all feature sets, and what is the number of classes in this case we have 2 classes 1 as anomaly and 0 as normal in the dataset. $p(i \mid D)$ is the probability of an instance

belonging to a class i in the dataset D. Simplified equation is given as

$$Entropy(D) = -(p(0 \mid D) * log_2(p(0 \mid D)) + p(1 \mid D) * log_2\,(p(1 \mid D))) \tag{2}$$

For each feature, $\boldsymbol{f}$, we calculate the information gain $IG(D, \boldsymbol{f})$ as follows

$$K = IG(D, \boldsymbol{f}) = Entropy(D) - \sum v \in Values(\boldsymbol{f}) |D||D_v| \cdot Entropy(Dv) \tag{3}$$

Where value $(\boldsymbol{f})$ represents the possible values of feature $\boldsymbol{f}$. $|D|$ is total number of instances in the dataset. $|D_v|$ is the number of instances D that have value v for feature $\boldsymbol{f}$. we calculate $IG(D, \boldsymbol{f})$ for each of the 79 features in the dataset to rank them based on their Information Gain. This ranking will help to select the most relevant features in Phase 1 of our methodology. Finally, **K** is the top feature selected in the first phase.

By applying IG in the initial phase, we narrow down the feature space, improving computational efficiency and preparing the dataset for further refinement.

3.3. Phase 2: XGBoost with Recursive Feature Elimination (XGBoost-RFE)

The reaming subset of features goes on to the second phase, where we add XGBoost with Recursive Feature Elimination (XGBoost-RFE), after the IG-based feature ranking. XGBoost is a potent gradient-boosting algorithm that is well-known for how well it assesses feature importance. In the setting of NIDS, XGBoost-RFE trains the model by iteratively removing the least significant features while focus keeping the most useful features. The process keeps on until the best possible feature set for NIDS is found.

K is the set of top selected k-features from phase 1. To obtain the importance scores for these features, we use XGBoost. The importance scores *(I)* for each feature in K can be calculated using XGBoost as follows:

$$I(\boldsymbol{K}) = \sum_{i=1}^{N} \frac{Gain_i}{Gain_{Total}} * \;\; I_i(\boldsymbol{k}) \tag{4}$$

Where N is the number of trees in the XGBoost ensemble, $Gain_i$ is the improvement in accuracy find by feature $\boldsymbol{k}$ in tree number i. $Gain_{Total}$ is the total improvement in accuracy provided by all features in all trees and $I_i(\boldsymbol{k})$ is the gain of feature $\boldsymbol{k}$ in tree number i.

After evaluating the importance scores for the top-k selected features, we identify the least important feature based on these scores, from the list of selected features, the least significant features are eliminated until we reach the optimal features for NIDS. The final selected features set is given by $I(\boldsymbol{k}) = \{k_1, k_2 \ldots..k_n\}$**,** this selected features now subjected to DNN for testing and training process.

3.4. Deep Nurel Network implementation

The selected features of the model we propose are tested, trained, and validated using deep learning in the last stage. We used Deep Neural Networks (DNN) to evaluate the performance of our proposed model and compared it with some of the standard DL models like Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). The DNN model falls under the umbrella of supervised learning and is trained over several layers. Based on the idea of an FFAN our proposed model consists of several hidden layers to enhance the obtained features to a higher capacity, the DNN employed in this work.

As seen in Figure 3, this proposed model consists of three dense layers: an input layer, a hidden layer that contains the ReLU as the activation function, and an output layer that uses the sigmoid function. $K = \{k_1, k_2 \ldots..k_n\}$**,** t are the selected n number features from our proposed feature selection approach from a total 79 input features from our dataset. while the output from the output layer which $Y = \{y_1, y_2\}$, output vector binary classification 1 represents anomaly and 0 attack. Mathematically, each hidden layer H_i output computation is represented as follows

$$H_i(K) \;=\; \sigma_i(W_i * K \;+\; B_i) \tag{5}$$

The weight metrics $W_i.B_i$, which is added to the weighted total, indicates the bias vector. The weighted sum of the inputs from the layer before to it F is represented as $W_i \; * K$. σ_i is the activation function that is not linear. The output of each hidden layer in a deep learning model is mostly determined by the weights, biases, and activation functions; the mathematical representation of this process is shown in the following equation. ReLU is a frequently used activation function. By producing the input if it is positive and 0 otherwise, it produces non-linearity. It is shown in the equation below as

$$f(K) \;= max(0, K) \tag{6}$$

The output of each neuron is given by $f(k)$ and s is the weighted sum of the input metrics. Because of its better probability prediction, sigmoid is used in the output layer for binary classification, restricting the output inside the interval [0, 1]. Mathematically, it is given as

$$f(k) \;=\; \frac{1}{1 + exp\,(-k)} \tag{7}$$

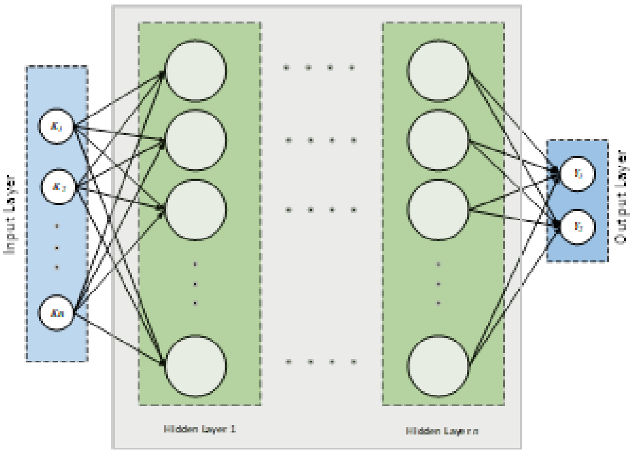

Fig 3 Deep learning model

3.5. Methodology

This study consists of three stages. The first stage includes the data preprocessing, the second stage is the feature selection stage, and finally implementation in DL for binary classification. The different steps involved in implementing and evaluating the proposed model include,

Step 1 Preprocess the data, which includes handling missing values using techniques like imputation (e.g., filling missing values with the mean or median). Normalizing the features using methods like Z-score normalization. Encoding categorical variables if necessary and also converting categorical data into numerical format.

Step 2 Check for Class Imbalance by analyzing the distribution of class labels to determine if there's class imbalance. If the class imbalance is present, implement the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes. This step helps in creating a more balanced dataset.

Step 3 Implement data augmentation to increase the diversity and size of our dataset. This can include techniques like Adding noise to data points, randomly modifying or perturbing features, and creating augmented data samples.

Step 4 Feature Selection in Phase 1 using Information Gain (IG) for Feature Ranking:

Calculate Information Gain scores for each feature. Information Gain measures the reduction in entropy provided by a feature, indicating its importance.

Select the top-k features with the highest Information Gain scores.

Step 5 The second Phase of feature selection includes XGBoost with Recursive Feature Elimination (XGBoost-RFE):

Apply XGBoost on the remaining features. Rank features based on their importance scores provided by XGBoost.

Eliminate the least important features in each iteration until you obtain the optimal feature subset.

Step 6 Split dataset into training and testing sets. Common splits are 80% training and 20% testing on our dataset.

Step 7 After feature selection, we proceeded with implementing a Deep Neural Network (DNN) model for NIDS and compared it with other DL methods

4. Implemenation

We use the BOT-IoT 2020 which is publicly available dataset in our proposed approach to evaluate the performance of deep learning model [20]. This dataset consists of 79 features and 1 is labeled to detect anomaly or normal [21]. The dataset we are using is in CSV format and it has a greater number of entries, making it perfect for testing our proposed method which is combination of information Gain (IG) and XGBoost with Recursive Feature Elimination (XGBoost-RFE), carefully select the best features for better results.

Preprocessing the dataset to improve data quality, fill in missing values, and balance feature scales is the first step in our proposed approach. We then employ the two-phase feature selection strategy to reduce dimensionality and improve computing performance by methodically identifying and keeping the most important features. Our deep neural network (DNN) model is optimized to identify network intrusions in the Internet of Things (IoT) context using these specific attributes. We do experiments by adjusting the hyperparameters of the DNN and assessing its accuracy, precision, recall, and F1 score. Table 1 shows the confusion metrics. We also do a comparison analysis to demonstrate the superiority of our method for IoT network security. We show through these studies how well our suggested technique works to handle security issues in IoT contexts.

Table 1: Confusion metrics

		Prediction	
		Attack	Normal
Attack	Attack	TP	FN
	Normal	FP	TN

Figure

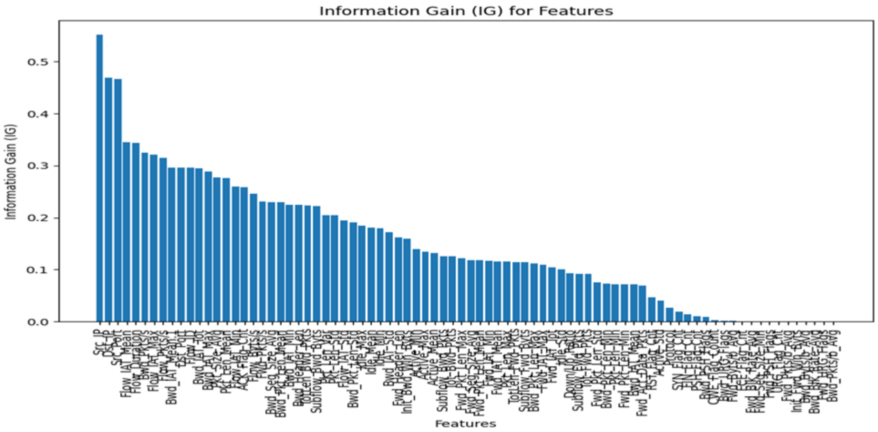

Fig 4 Features arranged in descending order based on IG

The best numerical features are arranged in descending order as shown in figure 4 selected according to the IG value. To determine the ideal collection of relevant and significant features that may be utilized to train the DL model with a low loss in detection accuracy, we took the top 76, 32, 16, and 8 features based on the greatest IG scores as shown in Figure 4. The selected top features in phase 1 are now further fine-tuned by XGBoost-RFE for further removal of redundant and correlated features from the previous phase

Table 2 Specification for the experiment

Parameters	**Specification**
Programming language	Python
Processor Model	13th Gen Intel(R)
CPU	Core (TM) i7-1360P 2.20 GHz
RAM	16 gigabytes
Temporary memory (Cache)size	56320KB
No. of cores in CPU	1
Windows	11

Table 3 Performance evaluation metrics

Model	***Performance evaluation metrics***			
	Accuracy	**Precision**	**Recall**	**F1 Score**
LSTM	0.5854	0.531	0.668	0.592
GRU	0.942	0.931	0.955	0.943
RNN	0.96525	0.967	0.9633	0.965
CNN	0.9712	0.969	0.991	0.98
DNN	0.998	0.998	0.999	0.989

5. Results and Discussion

We use the BOT-IOT 2020 dataset to evaluate our proposed methodology, which proposes a two-phase feature selection method that is followed by the deep neural network (DNN). The proposed method shows good results when compared to other feature selection methods which results in improved security in IoT networks. Our proposed research's experimental setup is built using Google Colab, a cloud-hosted Jupyter Notebook [22]. Table 2 specifies the parameters of the experimental setup.

The selected features from features selection are now subjected to different DL-based IDS techniques that were carefully used based on specific parameter selections in this study. The Adam optimizer was used, the learning rate was set at 0.01, and the batch size was fixed at 2^7. Binary cross-entropy was the loss function used, and the DL model activation functions were Sigmoid and Rectified Linear Unit (ReLU). Table 3 presents the performance evaluation metrics and provides an in-depth summary of the findings. Four different supervised DL algorithms were compared to the suggested DNN-based IDS methodology: the 1-dimensional CNN (CNN-1D), Recurrent Neural Network (RNN), and its variants, Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM). The proposed DNN-based IDS performs exceptionally well for an IoT network, compared to other DL methods in the evaluation. The CNN-1D model is the second-best. To improve, we adjusted model settings through hyperparameter tuning, refining them for better performance.

Table 4 FAR and FNR for different DL techniques

DL Algorithms	**LSTM**	**GRU**	**RNN**	**CNN**	**DNN**
FAR	2.1	1.9	0.45	0.1	0.0001
FNR	1.8	1.4	0.23	0.01	0.001

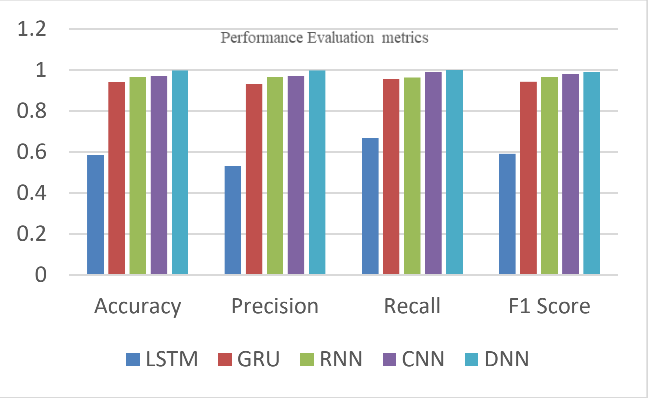

Fig 5 Performance of different DL techniques

A moderate performance was shown by the LSTM model, with an accuracy of 0.5854. It performed worse than other models, although it showed that it could capture certain time correlations. This suggests that managing complex patterns in the IoT network traffic data might be difficult. Accuracy of 0.942 indicates increased performance of the GRU model. However advanced models performed better than they could. The standard RNN exhibited its ability to capture sequential dependencies in the network traffic data by obtaining an acceptable accuracy of 0.96525. Its general performance was impacted when compared to more advanced models, though, as it shared LSTM's difficulties with long-term memory retention. The CNN model performed well, exhibiting an accuracy of 0.9712 as given in Table 3

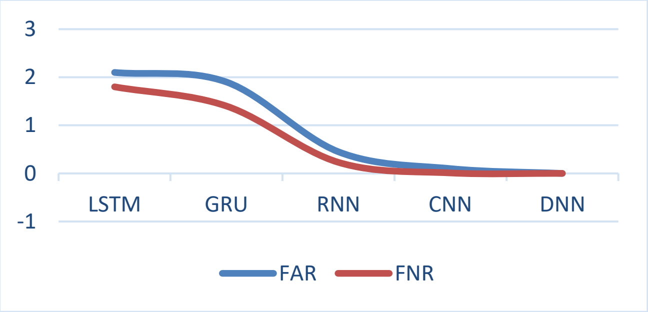

Fig 6 FAR and FNR curve

With an outstanding accuracy of 0.998, the DNN proved to be the most effective. Its deep design made it possible to learn complicated models efficiently, which improved the ability to recognize complex correlations in the data. In particular, it exhibits outstanding efficiency while utilizing features chosen by REF-XGBoost, demonstrating its capacity to utilize optimized feature sets for improved intrusion detection in Internet of Things networks. Even after tuning the paraments, LSTM performance was worse compared to other DL techniques as shown in Figure 5.

The False Acceptance Rate (FAR) and False Negative Rate (FNR) measures show different performance levels for different Deep Learning (DL) algorithms, such as LSTM, GRU, RNN, CNN, and DNN. DNN is the model with the lowest FAR of all of them, at 0.0001, indicating that it is better at minimizing the acceptance of false positives. In the same way, DNN's remarkably low FNR of 0.001 shows how well it reduces false negatives as shown in Figure 6. These findings show the DNN model's durability in finding an advantageous equilibrium between recall and accuracy. The percentage improvement of the DNN-based NIDS over other DL-based systems is shown in Figure 7.

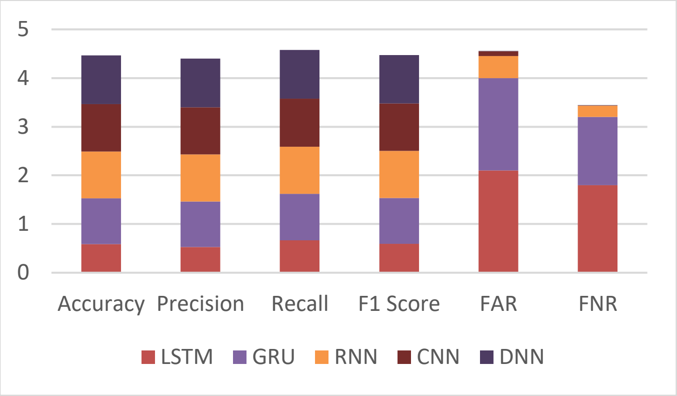

Fig 7 Performance improvement of DNN in comparison to previous DL-based systems.

We noticed a minimal decline in accuracy for the DNN model across feature sets of 76, 32, and 16. The model showed a slight drop of in detection accuracy when utilizing an 8-feature set. When we select 8 features accuracy should be compromised, so we make a tradeoff between computational complexity and accuracy by selected 16 features for training. Figure 8 shows the confusion matrix of proposed DNN model for 76,32,16 and 8 features set. From this confusion matrix it can be observed that anomalies are correctly detected for 76, 32 and 16 feature set, where as there is more miss classification for 8 features.

Figure 9 shows the confusion metrics of the different DL-based NIDS for IoT. It was found that over time, all of the DL approaches detection performance was enhanced by the all-type features like Floating or integer. The DNN and CNN-1D exhibited nearly perfect accuracy in detecting both anomaly and benign samples. Additionally, the GRU, RNN, and LSTM demonstrated enhanced detection performance when compared to the results obtained from experiments using all feature set. Among the considered DL-based methodologies, we noted that LSTM showed a higher number of incorrect predictions compared to others. Furthermore, it was observed that the DNN achieved similar results with just two hidden layers, effectively reducing the overall model complexity.

The present investigation provides a comprehensive comparative examination of different Deep Learning (DL)-based Network Intrusion Detection Systems (NIDS) specifically designed for Internet of Things (IoT) networks. In these contexts, the Deep Neural Network (DNN) outperformed alternative DL methodologies, attaining the highest detection accuracy at 99.8%.

Our study is different from and more focused than the study carried out by Ullah et. al. [21], our study firstly focuses on the DL approach as they only favor ML approaches. Secondly, we achieve the same accuracy with 16 features only, which results to reduce in computational complexity and system requirements

Our study differs from others in several aspects. Firstly, they favored a Machine Learning (ML) approach, while our research is distinctly focused on employing the Deep Learning (DL) paradigm. Additionally, their achievement of 100% accuracy was contingent upon utilizing 20 features, whereas we achieved accurate predictions with just 16 features [23]. This shows the value of the DL technique for IoT networks and emphasizes how well it can analyze massive amounts of data, leading to precise and useful predictions.

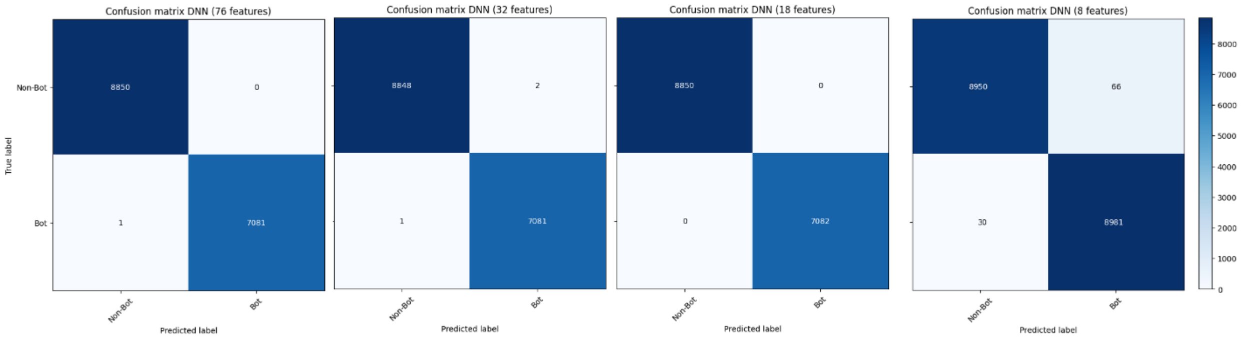

Fig 8 Confusion matrix for DNN with different features

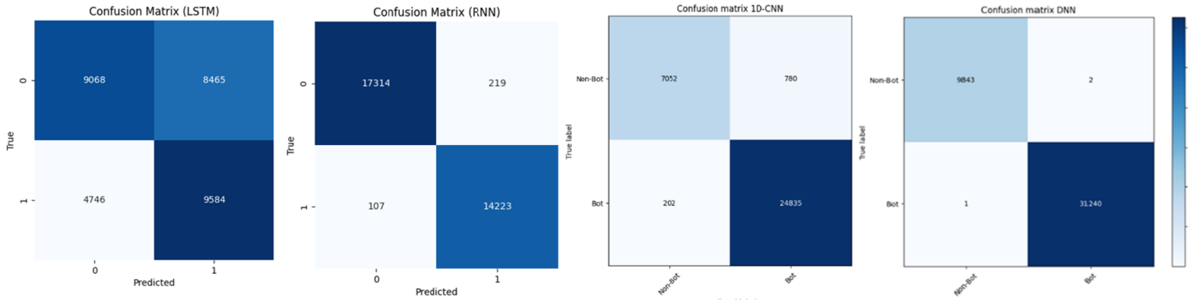

Fig 9 Confusion matrix for different DL approach

6. Conclusion

This paper proposed an effective novel two-phase feature selection method combining Information Gain and XGBoost-RFE for anomaly detection in based upon DL to secure the IoT. This feature selection approach successfully solves the problem which is raised by high dimensional IoT dataset IoT, this approach outperforms other feature selection techniques. The experiment was conducted using the BoTIoT dataset, which contains 79 features. These 79 features were narrowed down to 16 features by the proposed feature selection technique. The deep neural network (DNN) model was trained on the selected features; we achieved an outstanding detection accuracy of 99.8% superior compared to other DL approaches whereas LSTM archives the least accuracy. The proposed DNN achieves improvement in accuracy up to 0.56–2.7%, also we achieve 1% of FAR which shows our proposed method is more effective compared to the present study. It also observed model accuracy to detect anomaly traffic is increased compared to the existing method. The results show the importance that feature selection for maximizing NIDS performance and lowering computing complexity. In general, this study offers important insights into IoT security and offers an appropriate approach for improving NIDS performance against new types attacks.

Author contributions

Mohammed Sayeeduddin Habeeb received a master's degree in communication engineering from Osmania University, Hyderabad, India in 2010. He completed his master's thesis and internship at DOFI, RCI, DRDO laboratory (Ministry of Defense, India) and is currently pursuing a Ph.D. degree in Communication Engineering from the College of Engineering and Technology, Acharya Nagarjuna University, Andhra

Pradesh, India. His current research interests focus on

Internet-of-Things and Intrusion detection in networks and Device to device communication and trust management issues.

Tummala Ranga Babu obtained his Ph.D. in Electronics and Communication Engineering from JNTUH, Hyderabad, M.Tech in Electronics & Communication Engineering (Digital Electronics & Communication Systems) from JNTU College of Engineering (Autonomous), Anantapur, M.S. (Electronics & Control Engineering) from BITS, Pilani and B.E. (Electronics and Communication Engineering) from AMA College of Engineering (Affiliated to University of Madras). Served at different positions at different colleges. He is currently working as Professor and head of the Department of Electronics & Communication Engineering, acting as Chairman, Board of studies for ECE board for RVR & JC College of Engineering (Autonomous) and member of the Executive Council of RVR & JC College of Engineering (Autonomous). He is a member in various professional bodies like IEEE, IETE, ISTE, CSI, and IACSIT. His research interests include Image Processing, Embedded Systems, Pattern Recognition, and Digital Communication.

Conflicts of interest

The authors declare no conflicts of interest

References

[1] K. Albulayhi, Q. A. Al-Haija, S. A. Alsuhibany, A. A. Jillepalli, M. Ashrafuzzaman, and F. T. Sheldon, “IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method,” Applied Sciences 2022, Vol. 12, Page 5015, vol. 12, no. 10, p. 5015, May 2022, doi: 10.3390/APP12105015.

[2] B. Xu, L. Sun, X. Mao, R. Ding, and C. Liu, “IoT Intrusion Detection System Based on Machine Learning,” Electronics 2023, Vol. 12, Page 4289, vol. 12, no. 20, p. 4289, Oct. 2023, doi: 10.3390/ELECTRONICS12204289.

[3] N. V. Sharma and N. S. Yadav, “An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers,” Microprocess Microsyst, vol. 85, p. 104293, Sep. 2021, doi: 10.1016/J.MICPRO.2021.104293.

[4] M. Ahmed, A. Naser Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” Journal of Network and Computer Applications, vol. 60, pp. 19–31, Jan. 2016, doi: 10.1016/J.JNCA.2015.11.016.

[5] M. Ahmed, A. N. Mahmood, and M. R. Islam, “A survey of anomaly detection techniques in financial domain,” Future Generation Computer Systems, vol. 55, pp. 278–288, Feb. 2016, doi: 10.1016/J.FUTURE.2015.01.001.

[6] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, “Variable selection using random forests,” Pattern Recognit Lett, vol. 31, no. 14, pp. 2225–2236, Oct. 2010, doi: 10.1016/J.PATREC.2010.03.014.

[7] G. P. Dubey and D. R. K. Bhujade, “Optimal feature selection for machine learning based intrusion detection system by exploiting attribute dependence,” Mater Today Proc, vol. 47, pp. 6325–6331, Jan. 2021, doi: 10.1016/J.MATPR.2021.04.643.

A. Rashid, M. J. Siddique, and S. M. Ahmed, “Machine and Deep Learning Based Comparative Analysis Using Hybrid Approaches for Intrusion Detection System,” 3rd International Conference on Advancements in Computational Sciences, ICACS 2020, Feb. 2020, doi: 10.1109/ICACS47775.2020.9055946.

[8] M. A. Khan, “HCRNNIDS: Hybrid Convolutional Recurrent Neural Network-Based Network Intrusion Detection System,” Processes 2021, Vol. 9, Page 834, vol. 9, no. 5, p. 834, May 2021, doi: 10.3390/PR9050834.

[9] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, “Anomaly-based network intrusion detection: Techniques, systems and challenges,” Comput Secur, vol. 28, no. 1–2, pp. 18–28, Feb. 2009, doi: 10.1016/J.COSE.2008.08.003.

[10] Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” 2018, doi: 10.5220/0006639801080116.

[11] Fotiadou, T. H. Velivassaki, A. Voulkidis, D. Skias, S. Tsekeridou, and T. Zahariadis, “Network Traffic Anomaly Detection via Deep Learning,” Information 2021, Vol. 12, Page 215, vol. 12, no. 5, p. 215, May 2021, doi: 10.3390/INFO12050215.

[12] R. Vijayanand and D. Devaraj, “A Novel Feature Selection Method Using Whale Optimization Algorithm and Genetic Operators for Intrusion Detection System in Wireless Mesh Network,” IEEE Access, vol. 8, pp. 56847–56854, 2020, doi: 10.1109/ACCESS.2020.2978035.

[13] G. P. Dubey and D. R. K. Bhujade, “Optimal feature selection for machine learning based intrusion detection system by exploiting attribute dependence,” Mater Today Proc, vol. 47, pp. 6325–6331, Jan. 2021, doi: 10.1016/J.MATPR.2021.04.643.

[14] M. Gheisari, G. Wang, and M. Z. A. Bhuiyan, “A Survey on Deep Learning in Big Data,” Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017, vol. 2, pp. 173–180, Aug. 2017, doi: 10.1109/CSE-EUC.2017.215.

[15] M. A. Khan, "HCRNNIDS: Hybrid Convolutional Recurrent Neural Network-Based Network Intrusion Detection System," Processes 2021, Vol. 9, Page 834, vol. 9, no. 5, p. 834, May 2021, doi: 10.3390/PR9050834.

[16] M. S. Habeeb and T. R. Babu, "Network intrusion detection system: A survey on artificial intelligence-based techniques," Expert Syst, vol. 39, no. 9, p. e13066, Nov. 2022, doi: 10.1111/EXSY.13066.

[17] Z. Ahmad et al., "S-ADS: Spectrogram Image-based Anomaly Detection System for IoT networks," Proceedings - AiIC 2022: 2022 Applied Informatics International Conference: Digital Innovation in Applied Informatics during the Pandemic, pp. 105–110, 2022, doi: 10.1109/AIIC54368.2022.9914599.

[18] T. Wu, H. Fan, H. Zhu, C. You, H. Zhou, and X. Huang, "Intrusion detection system combined enhanced random forest with SMOTE algorithm," EURASIP J Adv Signal Process, vol. 2022, no. 1, pp. 1–20, Dec. 2022, doi: 10.1186/S13634-022-00871-6/TABLES/6.

[19] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," Future Generation Computer Systems, vol. 100, pp. 779–796, Nov. 2019, doi: 10.1016/J.FUTURE.2019.05.041.

[20] Ullah and Q. H. Mahmoud, "A Technique for Generating a Botnet Dataset for Anomalous Activity Detection in IoT Networks," Conf Proc IEEE Int Conf Syst Man Cybern, vol. 2020-October, pp. 134–140, Oct. 2020, doi: 10.1109/SMC42975.2020.9283220.

[21] "Welcome To Colaboratory - Colaboratory." Accessed: Dec. 25, 2023. [Online]. Available: https://colab.research.google.com/

[22] Ullah and Q. H. Mahmoud, "A Two-Level Flow-Based Anomalous Activity Detection System for IoT Networks," Electronics 2020, Vol. 9, Page 530, vol. 9, no. 3, p. 530, Mar. 2020, doi: 10.3390/ELECTRONICS9030530.