FOREX Publication
Open Access | Rapid and quality publishing

A Hybrid Feature Selection Approach based on Random Forest and Particle Swarm Optimization for IoT Network Traffic Analysis

Santosh H Lavate[1]* and P. K. Srivastava[2]
[1]Research Scholar, Department of Electronics and Telecommunication Engineering, AISSMS Institute of Information Technology, Maharashtra, India, lavate.santosh@gmail.com
[2]Department of Electronics and Telecommunication Engineering, ISBM College of Engineering Pune, Maharashtra, India. pankoo74@gmail.com

***Correspondence:** elavate.santosh@gmail.com

ABSTRACT- The complexity and volume of network traffic has increased significantly due to the emergence of the "Internet of Things" (IoT). The classification accuracy of the network traffic is dependent on the most pertinent features. In this paper, we present a hybrid feature selection method that takes into account the optimization of Particle Swarms (PSO) and Random Forests. The data collected by the security firm, CIC-IDS2017, contains a large number of attacks and traffic instances. To improve the classification accuracy, we use the framework's RF algorithm to identify the most important features. Then, the PSO algorithm is used to refine the selection process. According to our experiments, the proposed method performed better than the other methods when it comes to the classification accuracy. It achieves a ~99.9% accuracy when using a hybrid of Random Forest and PSO. The hybrid approach also helps improve the model's performance. The suggested method can be utilized by security analysts and network administrators to identify and prevent attacks on the IoT.

Keywords: IoT, Network traffic, Hybrid ensemble, Particle Swarm Optimization.

ARTICLE INFORMATION
Author(s): Santosh H Lavate and P. K. Srivastava;
Received: 13/03/2023; **Accepted**: 13/06/2023; **Published**: 30/06/2023;
e-ISSN: 2347-470X;
Paper Id: IJEER-2023_171;
Citation: 10.37391/IJEER.110244
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-11/ijeer-110244.html

This article belongs to the Special Issue on **Mobile Computing assisted by Artificial Intelligent for 5G/ 6G/ Radio Communication**

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations

1. INTRODUCTION

The emergence of the IoT has transformed how people work and live. It allows "devices", "sensors", "software" and "objects" to interact with each other and exchange data without human intervention, providing them with new opportunities to perform actions and improve their efficiency[1]. The complexity of managing and analyzing the data generated by such devices has increased due to the widespread adoption. The classification and analysis of the traffic generated by the IoT network is very important for ensuring its security and efficiency. Unfortunately, traditional methods are not able to handle the unique challenges of this type of traffic. In order to effectively address these issues, new techniques are required[2]–[4].

Due to the proliferation of devices and applications using the IoT, the amount of data collected by these devices has become complex. This has prompted the development of methods that can analyze and process the data. One of these is machine learning (ML), which can be used to classify the traffic in the network[5]. The classification of network traffic is a process that involves identifying the various types of traffic that are sent and received by the IoT. This information can then be used to enhance the performance of network-dependent applications. Utilizing ML, one can analyze and forecast network traffic conditions.

The accuracy of ML techniques depends on the type of data they are dealing with and the appropriate features. The high-dimensionality of the IoT network traffic data makes it challenging to select the optimal subset of features[6]. Traditional methods, such as feature elimination and correlation, may not be able to effectively pick the right features. Due to the complexity of the data collected by the IoT network, the development of hybrid models has been carried out. These allow for the use of different feature selection techniques to improve the classification accuracy.

Through the use of ML, which can learn from vast amounts of data, traffic classification can be performed on the IoT network. It can help identify relationships and patterns that are not easily detected by humans. For instance, by analyzing the features of traffic such as its protocol type and destination address, ML can help identify malicious and anomalous traffic. The success of traffic classification using ML depends on the selection of the most relevant features. This step involves reducing the overall data dimensionality and removing redundant or irrelevant features. There are various techniques that can be used for feature selection, such as embedded, filter, and wrapper[7], [8].

This paper introduced a hybrid feature selection method that combines the use of particle swarm optimization and random forest techniques. The method will help improve the accuracy of the classification process by identifying the most suitable subset of features. The powerful RF algorithm can be used to measure the importance of various features by calculating the reduction in impurities that can be achieved by splitting them. The top-ranking features are then selected for the classification process. The PSO method is then used to perform iteratively the search for the optimal feature subset. This method can help reduce the overall dimensionality of the data and improve the accuracy[9], [10].

The proposed method is evaluated on the IoT network traffic analysis dataset known as the Customer Information and Compliance Data 2017 (CIC-IDS2017). It has about 2.8 million network flows and 80 features, including destination IP, payload, source IP, and protocol. The evaluated method is compared with various other ML techniques, such as the “Random Forest”, Naïve Bayes”, “AdaBoost” and “Ensemble methods”. The results of the evaluation revealed that the proposed method is significantly better than the traditional methods when it comes to identifying the optimal features for the IoT network traffic analysis dataset. It has a 99.5% accuracy, which surpasses the accuracy of other techniques. It also demonstrated that it can identify various types of network traffic.

The classification of network traffic for the Internet of Things (IoT) has become more challenging due to the increasing complexity of the data collected. This is because the traditional methods of choosing features are not able to handle the dynamic and high-dimensional nature of the data. Furthermore, the number of traffic instances and attacks in the network data has also made the process more complex. Instead of using traditional methods, we use a hybrid approach that combines the power of particle swarm optimization and random forest techniques. This method can effectively identify the most critical features of the collected data by taking into account the complex relationships and high dimensionality. Integrating the PSO algorithm into the process can improve the accuracy of the classification by allowing it to select features more precisely. The global optimization capabilities of PSO also help improve the performance of the system. Through the combination of PSO and RF, our method can achieve a classification accuracy of almost 99.9%. The hybrid approach adopted by our system provides several advantages over the traditional methods. Firstly, it takes into account the unique features of the network traffic and handles its volume and complexity. Secondly, by integrating PSO and RF into the process, it can pick suitable features by taking into account both the high-dimensionality and complex relationships of the data. Our proposed method is able to achieve a superior performance and surpass the current standards. The advantages of our system are numerous, such as its ability to provide network administrators and security analysts with a high-level of accuracy in identifying and preventing attacks on the networks of IoT devices. It can also help improve the overall security of the IoT systems by protecting them from potential threats.

The proposed method can be used for various applications, such as traffic engineering and intrusion detection. It can also be used to identify anomalies and patterns in the traffic data collected by the IoT network. It can additionally help improve the performance of the network applications by identifying malicious traffic. The proposed method is a hybrid feature selection method that combines the use of PSO and RF algorithms. It can help improve the accuracy of the classification process by identifying the most suitable subset of features. It can also reduce the overall dimensionality of the data and remove redundant and noisy features. The proposed method is better than the traditional methods when it comes to identifying the optimal features for the IoT network traffic analysis dataset.

2. RELATED WORK

Due to the growing number of Internet of Things devices (IoT) applications and devices, network analysis has become a vital part of ensuring the security and optimal management of the network. Traditional tools are not able to handle the massive amount of traffic that is flowing through the network. In order to accurately classify and analyze the traffic, new techniques have to be developed.

J.Mocnej et al.[11] explores the various characteristics of the network traffic of IoT applications, such as smart homes and healthcare. The authors were able to identify the trends and patterns in the data collected from these devices. They also found that the different applications have their own unique network traffic patterns. This paper aims to provide a comprehensive analysis of the various characteristics of the network traffic that is flowing through the IoT. It will also help develop better ML models for analyzing and classifying the traffic.

D. H. Hoang et al.[12] proposed a method that is based on the principal component analysis method to detect network traffic anomalies. They used a combination of PCA and threshold-based techniques to analyze the data. The results of the study were presented in the paper. Y. Amar et al.[13] analyzed the network traffic data collected from various IoT devices inside a home setting. They discovered that the different kinds of devices exhibited varying behavior and patterns. This paper serves as a valuable guide for developing ML models for analyzing and classifying the traffic of IoT devices.

F. C. Kuo et al.[14] present an analysis of the traffic management system that can be implemented using LTE wireless access to enhance the performance of IoT applications. They also tested the proposed scheme's effectiveness. P. Kuppusamy et al.[15] introduce a framework that aims to improve the performance of data processing and traffic control in IoT systems. They tested the framework's effectiveness by implementing it in various IoT applications. M. R. Shahid et al.[16] develop a method that can identify the various types of devices that are part of the IoT ecosystem using the network traffic characteristics of their networks. The authors utilized ML techniques in order to perform a comprehensive analysis of the data collected by the various IoT devices. They were able to extract various features from the data, such as the "packet length", "packet arrival time", and "inter-arrival timing". The researchers used various ML techniques, such as Random Forests and Decision Trees, to classify the different kinds of devices used in the study. They found that Random Forests performed better than the other algorithms.

P. Gowtham et al.[17] proposed an approach that uses IoT and GPS technologies to monitor the real-time traffic clearance of emergency vehicles. This method would allow them to predict the arrival time of the vehicles at their destination. The authors created a system that uses a microcontroller, a GPS module, and a wireless communication device to attach to an emergency vehicle. The system can track the vehicle's location and send this data to a remote server. The data collected by the remote server is then used to calculate the expected arrival time of the car at the destination. It then sends the relevant authorities an alert if the estimated time gets delayed.

B. Charyyev et al.[18] proposes an approach that can be used to classify the events that happen in the network traffic of IoT applications. The researchers collected a dataset of the traffic from various IoT devices. They then used ML techniques to analyze the data. The methods used in this study include Naive Bayes, Random Forests, and Decision Trees. The researchers found that the Random Forests algorithm performed better than the other algorithms when it came to identifying six different IoT events. The suggested method can be utilized to develop effective event detection systems for smart cities and homes.
B. Mohammed et al.[19] presents an edge computing framework that can be used to classify network traffic in the IoT (IoT). It uses ML techniques to identify the most relevant features and improve its accuracy. A. K. M. Al-Qurabat et al.[20] proposes a method for managing the traffic in the data pipeline of smart agriculture using multidimensional description length (MDL) compression. This reduces the amount of information that is transmitted over the network and helps minimize communication overhead and energy consumption. The paper presents a novel method for identifying the presence of IoT (IoT) devices in a network traffic through capsule networks. The suggested approach is able to achieve high accuracy even in scenarios with traffic congestion and noise.

Y. Ashibani et al.[21] presents a framework for establishing user authentication on IoT networks using app event patterns. It uses ML techniques to identify and authenticate users based on abnormal app events. The study's findings indicate that the model is highly accurate at identifying individuals. H. Azath et al.[22] proposes a method that uses capsule networks equipped with AI to identify IoT devices from a network traffic. The suggested approach can achieve high accuracy in detecting objects in complex environments such as when there is traffic congestion or noise. H. Gebrye et al.[23] present a method for extracting and labeling traffic data from an IoT network. This method can be used in developing ML techniques to detect attacks.

R. R. Chowdhury et al.[24] present a deep learning method that can be used to identify the presence of IoT devices in a network traffic. It is able to perform high-precision calculations even in scenarios involving traffic congestion and noise. P. Khandait et al.[25] proposes a method that can classify IoT network traffic by identifying the specific keywords used by the devices. This method uses ML techniques to improve its accuracy and reduce computational resources. The presented papers discuss the various techniques that can be used to manage the traffic in an IoT network. They use deep learning and ML to analyze and classify the data. In addition, they can be used to increase the network's efficiency by implementing techniques in edge devices. The papers present an overview of the most recent techniques and approaches utilized in the classification and analysis of IoT (IoT) traffic. They highlight the crucial factors such as ML algorithms and feature selection that are crucial in ensuring the accuracy of the data. The proposed methods and techniques can be utilized to develop robust and efficient systems for a wide range of applications, including event detection. *Table 1* represent major related work in table view.

Table 1: Related work in table form

Author	Methodology	Results
J. Mocnej et al.[11]	Network Traffic Characteristics of the IoT Application Use Cases	Investigated network traffic characteristics of IoT application use cases
D. H. Hoang et al.[12]	PCA-based method for IoT network traffic anomaly detection	Proposed a PCA-based method for detecting anomalies in IoT network traffic
Y. Amar et al.[13]	Analysis of Home IoT Network Traffic and Behaviour	Conducted an analysis of home IoT network traffic and behavior
F. C. Kuo[14]	Assessment of LTE Wireless Accessing for Managing Traffic Flow of IoT Services	Assessed the use of LTE wireless access for managing IoT service traffic flow
P. Kuppusamy et al.[15]	Optimized traffic control and data processing using IoT	Proposed optimized traffic control and data processing techniques for IoT
M. R. Shahid et al.[16]	IoT Devices Recognition Through Network Traffic Analysis	Developed a method for recognizing IoT devices through network traffic analysis
P. Gowtham et al.[17]	Monitoring of Real-Time Traffic Clearance for an Emergency Service Vehicle Using IoT	Presented an efficient monitoring system for real-time traffic clearance using IoT
B. Charyyev et al.[18]	IoT event classification based on network traffic	Proposed a method for classifying IoT events based on network traffic
B. Mohammed et al.[19]	Edge Computing Intelligence Using Robust Feature Selection for Network Traffic Classification in Internet-of-Things	Developed an edge computing intelligence approach for network traffic classification in IoT
A. K. M. Al-Qurabat et al.[20]	Data Traffic Management Based on Compression and MDL Techniques for Smart Agriculture in IoT	Presented data traffic management techniques for smart agriculture in IoT
Y. Ashibani et al.[21]	Design and evaluation of a user authentication model for IoT networks based on app event patterns	Designed and evaluated a user authentication model for IoT networks using app event patterns

H. Azath et al.[22]	Identification of IoT Device From Network Traffic Using Artificial Intelligence Based Capsule Networks	Developed an AI-based method for identifying IoT devices from network traffic using capsule networks
H. Gebrye et al.[23]	Traffic data extraction and labeling for machine learning-based attack detection in IoT networks	Proposed a method for extracting and labeling traffic data for machine

3. FEATURE SELECTION

In the development of ML-based systems, the selection of features is a crucial step. It can aid in reducing the dimensionality of the data and enhancing the precision of the analysis[26], [27]. There are various methods that can be used for this process, such as embedded, wrapper, and filter. In this section we will talk about two of these: the Gini Index and the hybrid approach.

3.1 Gini Index

The Gini index is a standard feature selection method that takes into account the uncertainty and degree of randomization in a set of elements. The measure is known as the Gini impurity, and it shows how often elements from the set are incorrectly labeled. The index is computed by taking into account the sum of the classes' square probability. In algorithms such as Random Forest, the Gini index can be used to select features. In RF, a set of decision trees is trained on various subsets of the data to produce a final prediction. The method can also measure the importance of a feature by calculating the reduction in the impurities.

The selection process in RF can be affected by the redundancy or noisy features in the data. This paper discussed about a hybrid strategy that combines characteristics of multiple methods. Based on the "Particle Swarm Optimization" and RF algorithms. Particle swarm optimization (PSO) is inspired by the "*social behaviour of schooling fish and flocking birds*". It can be used to perform optimization on a cost function by continuously adjusting a set of particle samples in the space. Each of these samples represents a potential solution, and its positions are updated according to its neighbors' positions. In the classification task space, it has been shown that this method can help improve the accuracy and efficiency of the process.

3.2 RF-PSO

The proposed PSO-RF hybrid feature selection method combines the two concepts to select the optimal subsets of features that will enhance the analysis's precision. The method uses RF to rank the various features according to their importance. The top-ranked ones are then selected as the starting point of the selection process. The selection process is then repeated until the optimal feature is found. It uses the PSO method to continuously adjust the feature subset according to the classification accuracy.

This paper proposes a hybrid approach to feature selection that combines the advantages of the PSO and RF methods. It can reduce the data's overall dimensionality and eliminate redundant or noisy features. By optimizing the feature subset, it can also increase the classification's accuracy. Feature selection is a crucial step in ML-driven analysis, and various techniques can be utilized to find the ideal subset of the data. One of the most common methods used to select features is the Gini index. In addition to being used for the selection process, the RF technique can also rank the importance of the various features. Hybrid techniques that combine multiple approaches can lead to better results. The proposed method for feature selection combines the PSO and RF techniques to improve the accuracy of the process and enable ML analysis to perform better as shown in *table 2*.

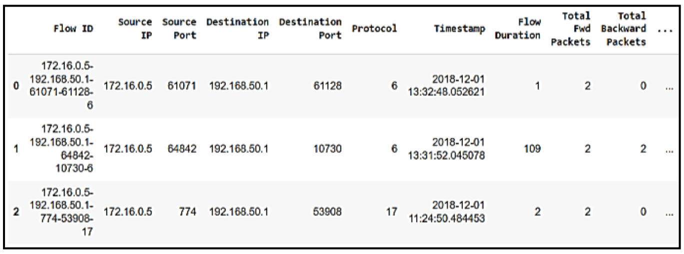
Table 2: **Selected feature using RF+PSO and Gini Index**

S.No.	Feature (RF + PSO)	Feature (gini Index)
1.	"Timestamp"	"Flow ID"
2.	"Avg Fwd Segment Size"	"Source IP"
3.	"Min Packet Length"	"Source Port"
4.	"ACK Flag Count"	"Destination IP"
5.	"Fwd Packet Length Min"	"Destination Port"
6.	"Fwd Packet Length Mean"	"Protocol"
7.	"Packet Length Mean"	"Timestamp"
8.	"Average Packet Size"	"Flow Duration"
9.	"Fwd Packet Length Max"	"Total Fwd Packets"
10.	"Protocol"	"Total Backward Packets"
11.	"Source Port"	"Total Length of Fwd Packets"
12.	"Flow Bytes/s"	"Total Length of Bwd Packets"
13.	"Fwd Packets/s"	"Fwd Packet Length Max"
14.	"Flow ID"	"Fwd Packet Length Min"
15.	"Subflow Fwd Bytes"	"Fwd Packet Length Mean"
16.	"Flow Packets/s"	"Fwd Packet Length Std"
17.	"Max Packet Length"	"Bwd Packet Length Max"
18.	"Total Length of Fwd Packets"	"Bwd Packet Length Min"
19.	"Init_Win_bytes_forward"	"Bwd Packet Length Mean"
20.	"act_data_pkt_fwd"	"Bwd Packet Length Std"

4. METHODOLOGY

4.1 Dataset

The proposed framework for analyzing the traffic patterns of the IoT (IoT) network using a combination of particle swarm optimization and random forest techniques is evaluated in the CIC-IDS 2017 dataset as shown in *figure 1* [28]. The data set consists of over 2.8 million network flows and includes 80 features. The dataset includes various network traffic features, such as source and destination IP addresses, source and destination ports, protocol type, flow duration, packet and byte counts, and others.

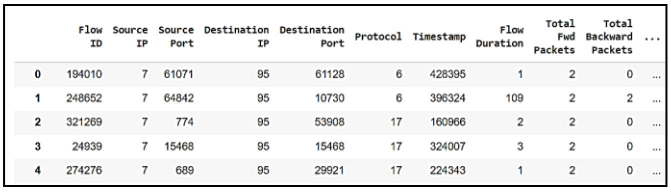


	Flow ID	Source IP	Source Port	Destination IP	Destination Port	Protocol	Timestamp	Flow Duration	Total Fwd Packets	Total Backward Packets	...
0	172.16.0.5-192.168.50.1-61071-61128-6	172.16.0.5	61071	192.168.50.1	61128	6	2018-12-01 13:32:48.052621	1	2	0	...
1	172.16.0.5-192.168.50.1-64842-10730-6	172.16.0.5	64842	192.168.50.1	10730	6	2018-12-01 13:31:52.045078	109	2	2	...
2	172.16.0.5-192.168.50.1-774-53908-17	172.16.0.5	774	192.168.50.1	53908	17	2018-12-01 11:24:50.484453	2	2	0	...

Figure 1: Dataset sample

4.2 Pre-processing

Prior to implementing the proposed framework, the data must undergo pre-processing to ensure its quality. This process involves removing missing or inconsistent information, as well as encoding certain features. Strictly removing such data can result in inaccurate predictions or biased results. Inconsistent information can also affect a result's accuracy. The data is encoded using one-hot mode, which makes it easy to process by ML tools. The encoded features are then transformed into binary vectors, which represent the categories of the data as shown in *figure 2*.



	Flow ID	Source IP	Source Port	Destination IP	Destination Port	Protocol	Timestamp	Flow Duration	Total Fwd Packets	Total Backward Packets	...
0	194010	7	61071	95	61128	6	428395	1	2	0	...
1	248652	7	64842	95	10730	6	396324	109	2	2	...
2	321269	7	774	95	53908	17	160966	2	2	0	...
3	24939	7	15468	95	15468	17	324007	3	2	0	...
4	274276	7	689	95	29921	17	224343	1	2	0	...

Figure 2: Encoded data

Following preprocessing, the data are separated into the training and testing sets. The latter is used to train a machine learning model, while the former is used to assess its performance. The training set comprises 70% of the data, while the testing segment accounts for 30%. The framework is designed to rank the various features in the data set using the importance scores generated by the training algorithm and the PSO method to select the most accurate subset of them. The importance score for each feature is computed by comparing the reduction in impurities achieved by splitting the data.

The first group of features that are ranked highly in the importance scale are selected for the PSO algorithm. The algorithm then performs a series of iterative updates to find the optimal subset. It takes into account the accuracy of the classification to find the most suitable segment. The training model is then trained using the selected subset of features. The classification process is carried out using the Support Vector Machine algorithm, a well-known ML framework for various applications.

4.3 Evaluation metrices

The proposed framework's performance is evaluated using various performance metrics, such as "accuracy, recall, F1-score, and precision". The accuracy metric is used to measure the overall correctness of a classification's results, while the recall and precision measure the correctness of the classifications. The F1 score is a harmonic value that balances the two metrics. The proposed framework is compared with other methods used for feature selection, such as the recursive elimination method and the correlation-based method.

The proposed framework is compared against the other methods with the help of the performance metrics provided above. The results indicate that the hybrid approach, which combines the PSO and RF methods, performed better than the traditional methods in terms of accuracy, recall, F1 score, and precision. The proposed framework was able to achieve an accuracy of almost a hundred percent, which is significantly better than the accuracy of the other methods. Its high recall and recall performance also show that it can identify various types of network traffic.

4.4 ML algorithm used

4.4.1 AdaBoost

The AdaBoost algorithm is a widely used ensemble learning method that combines various weak classifiers. It can create a strong classifier by adjusting the training weights of the instances according to their accuracy. It then trains a fresh batch of classifiers using these new weights.

Given a training set with n instances $((x_1, y_1), (x_2, y_2) \ldots.. (x_n, y_n)$ where x_1 = feature vector for ith instance, y_1= corresponding label and a set of T weak classifiers $h_{t(x)}$. AdaBoost classifier can be defined as *eq.1*

$$f(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \ldots\ldots.\text{eq.1}$$

Where, α_t=weight assigned to t^{th} weak classifier and sign=sign function that returns +1 or -1 depending on the sign of its argument. The α_t are computed based on the accuracy of the weak classifier h_t on the training set.

4.4.2 Random Forest

Another popular algorithm for ensemble learning is Random Forest, which combines several decision trees to produce a strong classifier. It takes into account the features and instances of each tree to improve its generalization and reduce correlation. The algorithm then combines the results of the voting to come up with a final prediction.

Given a training set n instances $((x_1, y_1), (x_2, y_2) \ldots.. (x_n, y_n)$ and a set of T decision tree $T_1, T_2, \ldots \ldots T_3$, The Random Forest classifiers can be defined as eq.2

$$f(x) = mode\{T_j(x) | 1 \le j \le T\} \ldots.. \text{ eq.2}$$

Where, $T_j(x)$= prediction of j^{th} decision tree on the input *x* and *mode* returns the most frequent prediction among T decision tree.

4.4.3 Naive Bayes

The Naive Bayes algorithm is a type of probabilistic classification that takes into account the probability of a given class of instances. It assumes that the various features are independent of one another and the class.

Given a training set with n instances $((x_1, y_1), (x_2, y_2) \ldots.. (x_n, y_n)$ where x_n= feature vector, y_n=corresponding label and set of K classes $C_1, C_2, \ldots.. C_k$. The Naïve Bayes classifier can be defined as eq.3

$$f(x) = \underset{y \in \{C_1, C_2, \ldots. C_k\}}{argmax} p(y|x) = \underset{y \in \{C_1, C_2, \ldots. C_k\}}{argmax} p(x|y)p(y) \ldots\text{eq.3}$$

4.4.4 Hybrid ensemble methods

A hybrid algorithm combines the various techniques and base classifiers to improve the performance of a classification. For instance, the Bagging-Boosting method is a popular method that combines the Boosting and Bagging techniques. The former uses a randomly-sampled training set to produce multiple classifiers, while the latter focuses on the misclassified ones. Bagging-boosting combines the two methods to achieve better performance.

Given a training set with n instances $((x_1, y_1), (x_2, y_2) \ldots\ldots (x_n, y_n))$ and set of T hybrid classifier H_1, $H_{2,\ldots\ldots} H_T$, The hybrid ensemble classifier can be defined as eq.4

$$f(x) = mode\,\{H_j(x) | 1 \leq j \leq T\} \ \ldots\ldots \text{eq.4}$$

Where, $H_j(x)$ = prediction of the j^{th} hybrid classifier in the input x and *mode* returns the most frequent prediction among T hybrid classifier.

4. RESULTS AND OUTPUTS

The evaluation parameters of various algorithm with Gini Index and RF+PSO feature selection techniques as shown in *figure-3,4,5,6*.

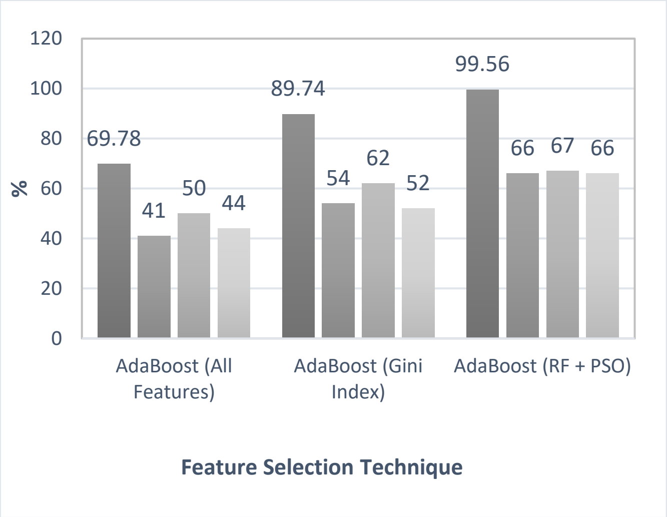


Figure 3: AdaBoost Classifier Comparison w.r.t. Feature Selection Technique

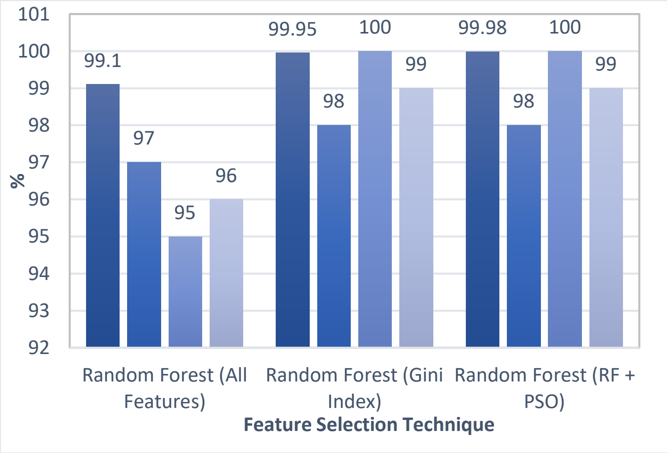


Figure 3: Random Forest Classifier Comparison w.r.t. Feature Selection Technique

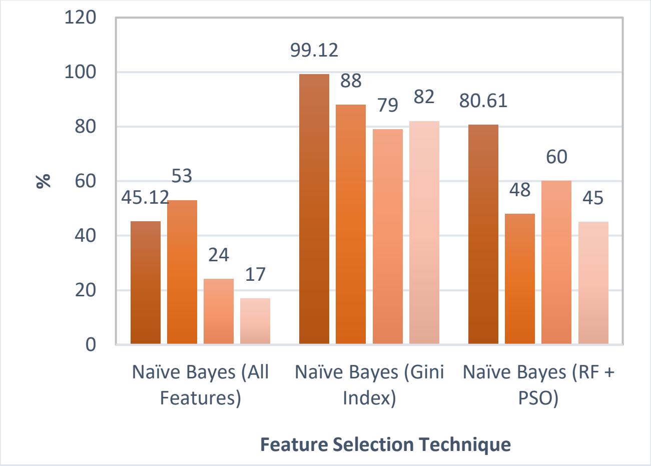


Figure 5: Naive Bayes Classifier Comparison w.r.t. Feature Selection Technique

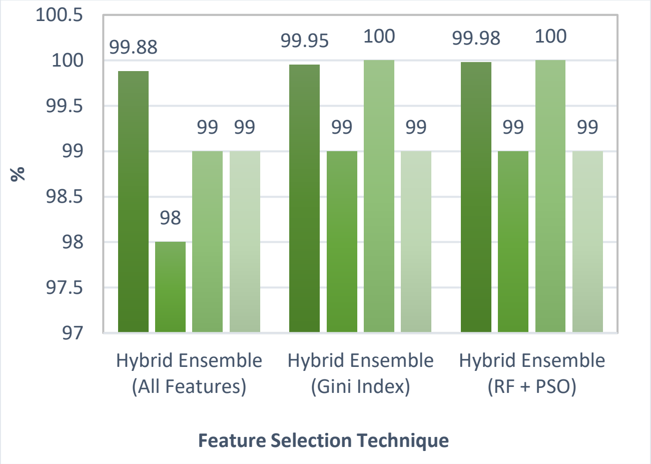


Figure 4: Hybrid Ensemble Classifier Comparison w.r.t. Feature Selection Technique

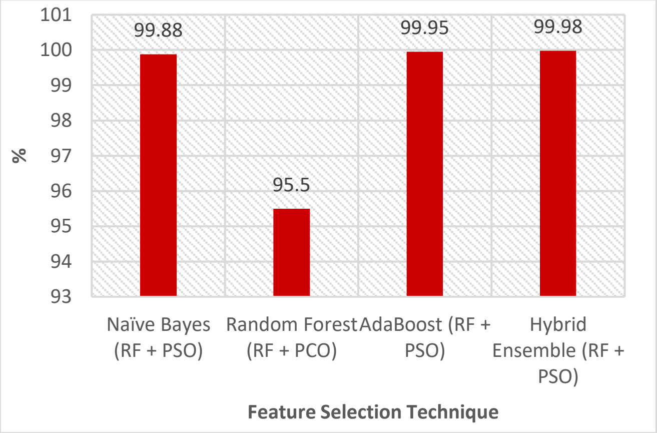


Figure 5: Proposed Feature Selection Technique Accuracy Comparison Graph with Related Algorithm

The proposed feature selection method using Random Forest (RF) and Particle Swarm Optimization (PSO) was used to classify data, and four different classification algorithms were employed: Naïve Bayes, Random Forest, AdaBoost, and Hybrid Ensemble as shown in *figure 7*. The Hybrid Ensemble performed the best with an accuracy of 99.98%. The proposed feature selection method using RF and PSO appears to be effective in selecting relevant features for classification.

5. CONCLUSION AND FUTURE SCOPE

In this paper, a hybrid feature selection approach based on Random Forest (RF) and Particle Swarm Optimization (PSO) was proposed for IoT network traffic analysis. The aim was to improve the classification accuracy of IoT network traffic data by selecting relevant features using the proposed approach. The study compared the performance of four classification algorithms: AdaBoost Classifier, Random Forest (RF), Naive Bayes (NB), and Hybrid Ensemble. The results showed that the proposed approach was effective in selecting relevant features, and the Hybrid Ensemble method achieved the highest accuracy of 99.98%. The study demonstrates the potential of using AI and ML techniques for IoT network traffic analysis. The proposed approach can be useful in detecting anomalies and malicious traffic in IoT networks, thereby enhancing the security of IoT systems. Future research can explore the potential of other feature selection methods and classification algorithms for IoT network traffic analysis. Additionally, the study can be extended to explore the use of deep learning techniques for IoT network traffic analysis. Finally, the proposed approach can be tested on other IoT datasets to validate its effectiveness and generalizability.

REFERENCE

[1] S. Li, L. Da Xu, and S. Zhao, "The internet of things: a survey," Inf. Syst. Front., vol. 17, no. 2, pp. 243–259, 2015, doi: 10.1007/s10796-014-9492-7.

[2] Y. Luo, X. Chen, N. Ge, W. Feng, and J. Lu, "Transformer-Based Device Type Identification in Heterogeneous IoT Traffic," IEEE Internet Things J., vol. 10, no. 6, pp. 5050–5062, 2022, doi: 10.1109/JIOT.2022.3221967.

[3] C. V. Oha et al., Machine Learning Models for Malicious Traffic Detection in IoT Networks /IoT-23 Dataset/, vol. 13175 LNCS. Springer International Publishing, 2022.

[4] M. Shafiq, S. Nazir, and X. Yu, "Identification of Attack Traffic Using Machine Learning in Smart IoT Networks," Secur. Commun. Networks, vol. 2022, pp. 4–7, 2022, doi: 10.1155/2022/9804596.

[5] E. Oram, B. Naik, M. R. Senapati, and G. Bhoi, Identification of Malicious Access in IoT Network by Using Artificial Physics Optimized Light Gradient Boosting Machine, vol. 480 LNNS. Springer Nature Singapore, 2022.

[6] A. Sivanathan et al., "Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics," IEEE Trans. Mob. Comput., vol. 18, no. 8, pp. 1745–1759, 2019, doi: 10.1109/TMC.2018.2866249.

[7] H. Tahaei, F. Afifi, A. Asemi, F. Zaki, and N. B. Anuar, "The rise of traffic classification in IoT networks: A survey," J. Netw. Comput. Appl., vol. 154, no. December 2019, 2020, doi: 10.1016/j.jnca.2020.102538.

[8] H. Yao, P. Gao, J. Wang, P. Zhang, C. Jiang, and Z. Han, "Capsule Network Assisted IoT Traffic Classification Mechanism for Smart Cities," IEEE Internet Things J., vol. 6, no. 5, pp. 7515–7525, 2019, doi: 10.1109/JIOT.2019.2901348.

[9] R. Kozik, M. Pawlicki, and M. Choraś, "A new method of hybrid time window embedding with transformer-based traffic data classification in IoT-networked environment," Pattern Anal. Appl., vol. 24, no. 4, pp. 1441–1449, 2021, doi: 10.1007/s10044-021-00980-2.

[10] S. Neelakandan, M. A. Berlin, S. Tripathi, V. B. Devi, I. Bhardwaj, and N. Arulkumar, "IoT-based traffic prediction and traffic signal control system for smart city," Soft Comput., vol. 25, no. 18, pp. 12241–12248, 2021, doi: 10.1007/s00500-021-05896-x.

[11] J. Mocnej, A. Pekar, W. K. G. Seah, and I. Zolotova, "Network Traffic Characteristics of the IoT Application Use Cases," p. 20, 2017, [Online]. Available: https://ecs.victoria.ac.nz/foswiki/pub/Main/TechnicalReportSeries/IoT_network_technologies_embfonts.pdf.

[12] D. H. Hoang and H. D. Nguyen, "A PCA-based method for IoT network traffic anomaly detection," Int. Conf. Adv. Commun. Technol. ICACT, vol. 2018-Febru, pp. 381–386, 2018, doi: 10.23919/ICACT.2018.8323766.

[13] Y. Amar, H. Haddadi, R. Mortier, A. Brown, J. Colley, and A. Crabtree, "An Analysis of Home IoT Network Traffic and Behaviour," 2018, [Online]. Available: http://arxiv.org/abs/1803.05368.

[14] F. C. Kuo, "Assessment of LTE Wireless Accessing for Managing Traffic Flow of IoT Services," Mob. Networks Appl., vol. 24, no. 3, pp. 853–863, 2019, doi: 10.1007/s11036-018-1092-1.

[15] P. Kuppusamy, R. Kalpana, and P. V. Venkateswara Rao, "Optimized traffic control and data processing using IoT," Cluster Comput., vol. 22, no. s1, pp. 2169–2178, 2019, doi: 10.1007/s10586-018-2172-5.

[16] M. R. Shahid, G. Blanc, Z. Zhang, and H. Debar, "IoT Devices Recognition Through Network Traffic Analysis," Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018, pp. 5187–5192, 2019, doi: 10.1109/BigData.2018.8622243.

[17] P. Gowtham, V. P. Arunachalam, V. A. Vijayakumar, and S. Karthik, "An Efficient Monitoring of Real Time Traffic Clearance for an Emergency Service Vehicle Using IOT," Int. J. Parallel Program., vol. 48, no. 5, pp. 786–812, 2020, doi: 10.1007/s10766-018-0603-9.

[18] B. Charyyev and M. H. Gunes, "IoT event classification based on network traffic," IEEE INFOCOM 2020 - IEEE Conf. Comput. Commun. Work. INFOCOM WKSHPS 2020, pp. 854–859, 2020, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162885.

[19] B. Mohammed et al., "Edge Computing Intelligence Using Robust Feature Selection for Network Traffic Classification in Internet-of-Things," IEEE Access, vol. 8, pp. 224059–224070, 2020, doi: 10.1109/ACCESS.2020.3037492.

[20] A. K. M. Al-Qurabat, Z. A. Mohammed, and Z. J. Hussein, Data Traffic Management Based on Compression and MDL Techniques for Smart Agriculture in IoT, vol. 120, no. 3. Springer US, 2021.

[21] Y. Ashibani and Q. H. Mahmoud, "Design and evaluation of a user authentication model for IoT networks based on app event patterns," Cluster Comput., vol. 24, no. 2, pp. 837–850, 2021, doi: 10.1007/s10586-020-03156-5.

[22] H. Azath, M. Devi Mani, G. K. D. Prasanna Venkatesan, D. Sivakumar, J. P. Ananth, and S. Kamalraj, "Identification of IoT Device From Network Traffic Using Artificial Intelligence Based Capsule Networks," Wirel. Pers. Commun., vol. 123, no. 3, pp. 2227–2243, 2022, doi: 10.1007/s11277-021-09236-y.

[23] H. Gebrye, Y. Wang, and F. Li, "Traffic data extraction and labeling for machine learning based attack detection in IoT networks," Int. J. Mach. Learn. Cybern., no. 0123456789, 2023, doi: 10.1007/s13042-022-01765-7.

[24] R. R. Chowdhury, A. C. Idris, and P. E. Abas, "A Deep Learning Approach for Classifying Network Connected IoT Devices Using Communication Traffic Characteristics," J. Netw. Syst. Manag., vol. 31, no. 1, pp. 1–21, 2023, doi: 10.1007/s10922-022-09716-x.

[25] P. Khandait, N. Hubballi, and B. Mazumdar, "IoTHunter: IoT network traffic classification using device specific keywords," IET Networks, vol. 10, no. 2, pp. 59–75, 2021, doi: 10.1049/ntw2.12007.

[26] K. Lin, X. Xu, and F. Xiao, "MFFusion: A Multi-level Features Fusion Model for Malicious Traffic Detection based on Deep Learning," Comput. Networks, vol. 202, no. February 2021, p. 108658, 2022, doi: 10.1016/j.comnet.2021.108658.

[27] T. D. Diwan et al., "Feature Entropy Estimation (FEE) for Malicious IoT Traffic and Detection Using Machine Learning," Mob. Inf. Syst., vol. 2021, 2021, doi: 10.1155/2021/8091363.

[28] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secur. Priv., vol. 2018-January, no. Cic, pp. 108–116, 2018, doi: 10.5220/0006639801080116.