# IDS Model for Detecting IoT Network Attacks

Meenakshi Maharana[1] Ashalata Panigrahi[2]
[1]M.Tech Scholar, NIST University, Berhampur, India
[2]Associate Professor, NIST University, Berhampur, India

## Abstract

The Internet of Things (IoT) has become a critical technology in modern industries, driving innovations in automation, healthcare, smart cities, and more. As IoT networks expand, ensuring the security and reliability of these systems becomes increasingly vital. Intrusion Detection Systems (IDS) play a pivotal role in identifying and mitigating security threats within IoT environments. Due to dynamic nature and volume of data in IoT networks, the use of advanced classification techniques is essential for effective anomaly detection and threat analysis. The RT-IoT Dataset 2022 provides a valuable resource for developing and testing intrusion detection models in real-time IoT applications. In this study four machine learning techniques namely, Random Forest, Extra Tree, Logistic Regression and Support vector Machine have been employed to built an anomaly-based network intrusion detection model. Further, in order to remove irrelevant features from the RT-IoT 2022 dataset two feature selection techniques namely, information gain and gain ratio have been applied. The efficiency of various combinations of four classifiers and two feature selection methods was analyze based on five evaluation matrix such as accuracy, precision, recall, f1-score and false alarm rate. Experimental results showed that Information Gain with Random Forest classifier is the best model with Accuracy 0.9913, Precision 0.9913, Recall 0.9913 and F1 Score 0.9913. Gain Ratio with Random Forest classifier is the best model with Accuracy 0.9915, Precision 0.9915, Recall 0.9915 and F1 Score 0.9915.Both Information Gain and Gain Ratio with Extra Tree give lowest False Alarm Rate of 0.0052 and 0.0057 respectively.

**Keywords:** Random forest, intrusion detection system, information gain, F1-score, extra tree.

## 1. Introduction

With the vast growth of the Internet of Things (IoT) applications such as smart homes, health care, manufacturing, transportation, agriculture the number of intelligent devices connected to the IoT is increasing exponentially. These devices, which range from household appliances like smart refrigerators to industrial sensors in manufacturing plants and they are capable of collecting, processing, storing , communicating data with each other over the internet and enabling intelligent decision making. The rapid proliferation of IoT devices brings significant benefits, such as automation, enhanced efficiency and productivity, and data-driven insights. However, it also introduces several security challenges due to the large scale, diversity of devices, and often limited built-in security features of many IoT devices. With billions of IoT devices expected to be connected globally in the coming years, the security of IoT networks is a growing concern. Threats such as unauthorized access, data breaches, and denial-of-service (DoS) attacks are becoming more prevalent. As IoT devices are often deployed in critical infrastructure, their compromise could lead to significant consequences. Ensuring the security and integrity of IoT networks is, therefore, a fundamental requirement for their continued growth and acceptance. An Intrusion Detection System (IDS) is a critical security tool used to monitor network or system activities for signs of unauthorized access or malicious behavior. An IDS specifically focuses on identifying suspicious patterns and alerting the security team so they can take appropriate action to mitigate or investigate the threat. IDS plays a vital role in cyber security by continuously monitoring the network and system activities, identifying potential threats, and providing alerts for further analysis. Intrusion detection can be broadly classified as signature based and anomaly based [1]. Signature-based

IDS detects attacks by matching traffic or system activity with predefined signatures or known patterns of malicious behavior. But, anomaly based IDS has the ability to detect new and unknown types of intruders.

The objective of the study is to develop an efficient Intrusion Detection System(IDS) using different classification techniques for IoT applications that can detect harmful attacks with low false positive rate and high accuracy.

The rest of the article is structured as follows: Section 2 presents previously published related work on intrusion detection, Section 3 describes proposed model, Section 4 presents the classification techniques, Section 5 introduces the experimental design and setup, finally conclusion and future work are presented in Section 6.

## 2. Related Work

Elzaghmouri, Bassam Mohammad, et al.[2] presents an innovative hybrid deep learning architecture that excels at detecting IoT threats in real-world settings. The proposed model combines Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU), and Attention mechanisms into a cohesive framework. This integrated structure aims to enhance the detection and classification of complex cyber threats while accommodating the operational constraints of diverse IoT systems. The model evaluated using the RT-IoT2022 dataset. The model surpassed traditional machine learning algorithms and the state-of-the-art, achieving over 99.6% precision. In [3] the authors have proposed IDS based on a deep learning model called Pearson-Correlation Coefficient - Convolutional Neural Networks (PCC-CNN) to detect network anomalies. The model applied for both binary and multi-class classification for detection of various types of attacks. The model is evaluated on three publicly available datasets: NSL-KDD, CICIDS-2017, and IOTID20. For performance evaluation five machine learning algorithms namely, SVM, Logistic regression, Linear Discriminant Analysis, KNN, and CART are applied on the datasets. PCC-CNN model performed better as compared with the state-of-the-art ML approaches. In [4] authors proposed anomaly based IDS model using different ML methods namely Random Forest (RF), Adaptive Boosting (AB), Logistic Regression (LR), and Neural Network (NN). For experiment they used Canadian Institute for Cybersecurity (CIC) dataset. In the dataset 33 types of attacks are present and are divided into seven categories. The result shows that RF performs 99.55% accuracy. In [5] the authors proposed anomaly based IDS to detect cyber anomalies within IoT system. To detect anomalies they have applied ML algorithms Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (DT), Logistic Regression (LR), and K-Nearest Neighbors (k- NN). Results demonstrate that ANN performs better than other models.

Adel Abusitta [6] presents a deep learning-powered anomaly recognition for IoT. The proposed model is designed based on a denoising autoencoder. Also, the denoising autoencoder allows the system to obtain features. Finally, experiments were conducted using the DS2OS traffic dataset.

## 3. Proposed Model

This work focus on anomaly based Network IDS. This study consists of 2 steps.

**Step 1:**

For selection of important features two feature selection methods namely Information Gain and Gain Ratio are applied on the dataset before the training process.

**Step 2:**

The classification stage evaluates the performance of the four machine learning techniques namely Random Forest, Extra Tree, Logistic Regression and Support Vector Machine on the selected subset of features. The proposed model was evaluated using Splitting Criteria(80% training and 20% testing). Confusion Matrix was used to compare the performance of different metrics namely accuracy, precision, recall, f1-score and false alarm rate.

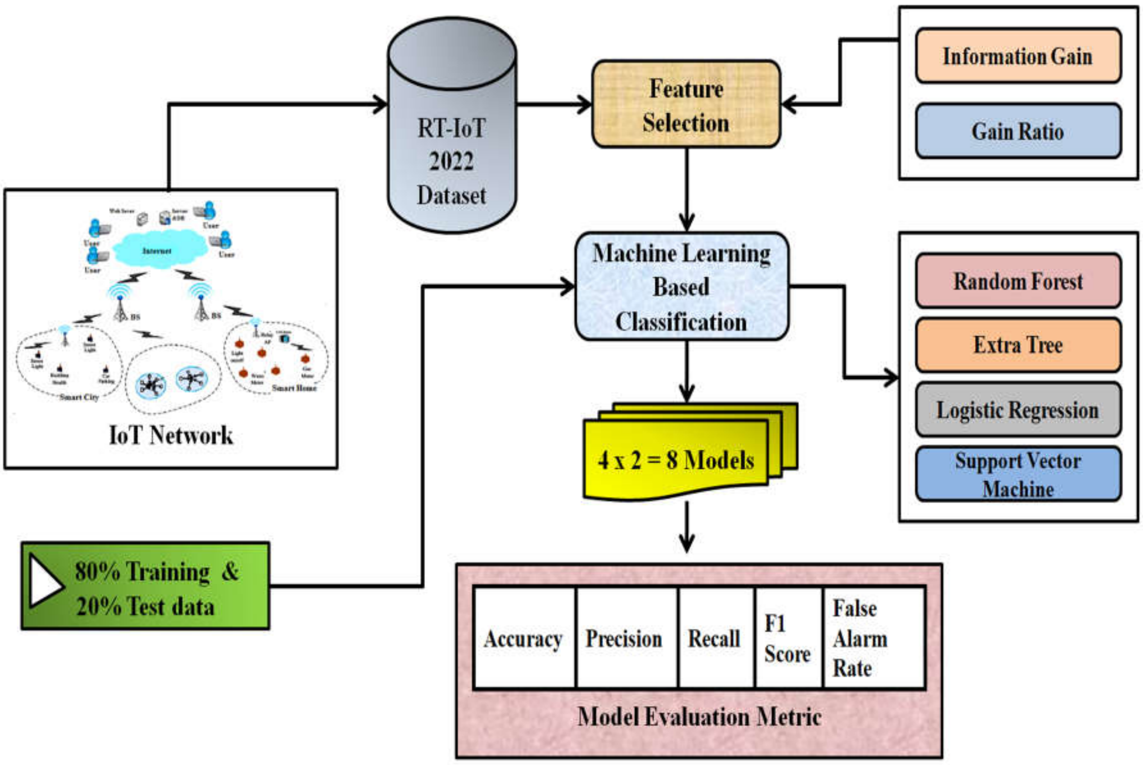


**Fig. 1** IDS Model for the IoT Network Attacks

# 4. Classification Techniques

## 4.1 Random Forest

Random forest technique randomly select a subset of the features from the dataset and final result obtained by training the model using multiple decision trees as base learners. The algorithm uses the concept of Gini index to determine which features are most important for classification [ 7]

## 4.2 Extra Tree

Extra tree is a tree-based ensemble method for supervised classification and regression problems [8]. It builds multiple decision trees from the training data. For each split in a tree a random subset of features are used without searching for optimal thresholds. A random value within the feature's range is chosen as the split point. The strength of the randomization can be tuned to problem specifics by the appropriate choice of a parameter.

## 4.3 Logistic Regression

Logistic regression used for binary classification problems (when target is categorical). logistic regression uses the sigmoid function also called the logistic function to map predicted values to a probability between 0 and 1 [9].

## 4.4 Support Vector Machine

Support vector machine (SVM) is a supervised machine learning technique based on statistical learning theory [10]. The idea of SVM is to find a hyperplane in the *n*th number of features that distinctly classifies the data points. Many possible hyperplanes are available, but only one hyperplane is chosen to classify two classes of the data points. The orientation and the position of the hyperplane is influenced by some data points that are closer to the hyperplane. These data points are called the support vectors of the hyperplane. SVM is a binary classifier but it can handle problems with many classes and also numerical prediction tasks .

# 5. Experimental Setup

## 5.1 Dataset Description

The RT-IoT2022 dataset has been used for experimentation [11]. The dataset comprises 123,117 instances and 83 features. It includes both normal and adversarial network behaviors. The dataset contains both numerical and categorical features. Nine different attack types are present in the dataset namely, DOS_SYN_Hping, ARP_poisoning, NMAP_UDP_SCAN, MAP_XMAS_TREE_SCAN, NMAP_OS_DETECTION, NMAP_TCP_scan, DDOS_Slowloris, Metasploit_Brute_Force_SSH, NMAP_FIN_SCAN. Dataset description depicted in Table 1.

Table 1 RT-IoT 2022 Dataset ( Total Records)

| RT-IoT Dataset(2022) | |
|---|---|
| Category | No. of Records |
| Normal | 12507 |
| DoS_SYN_Hping | 94659 |
| ARP_Poisioning | 7750 |
| NMAP_UDP_SCAN | 2590 |
| NMAP_XMAS_TREE_SCAN | 2010 |
| NMAP_OS_Detection | 2000 |
| NMAP_TCP_SCAN | 1002 |
| DDoS_Slowloris | 534 |
| Metasploit_Brute_Force_SSH | 37 |
| NMAP_FIN_SCAN | 28 |

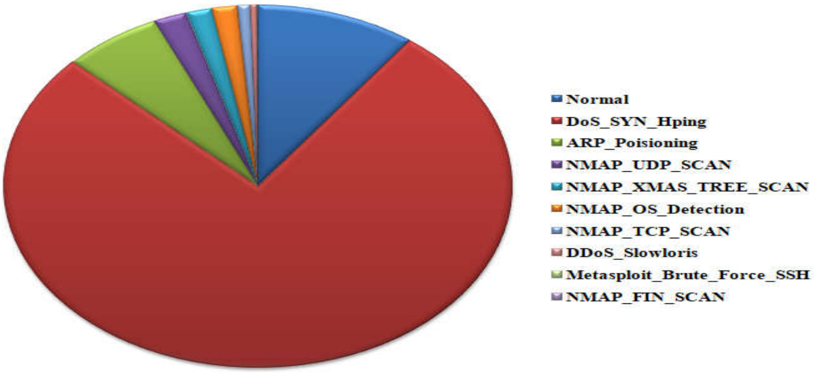


Fig. 2 RT-IoT Dataset Representation

In this work randomly selected 23000 records and all the 83 features from the original dataset for experimental study. Out of 23000 records 12507 are normal and 10493 are attack.

Table 2. RT-IoT 2022 Dataset Distribution

| Category | Percentage(%) of Records |
|---|---|
| Normal(12507) | 54.37% |
| Attack(10493) | 45.62% |

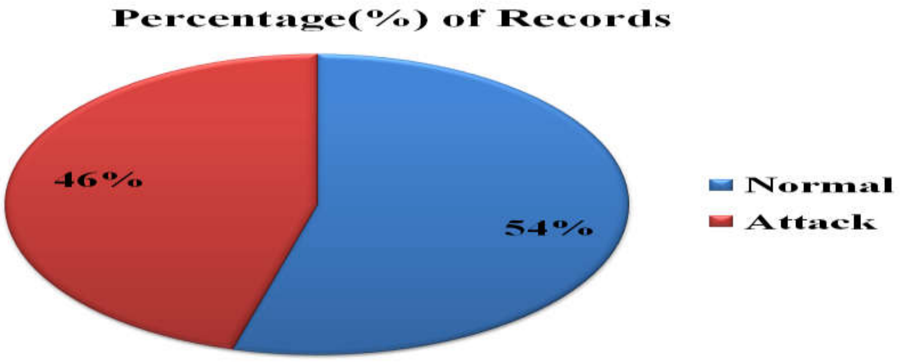


Fig. 3 RT-IoT Dataset Distribution Representation of 23000 records

### 5.2 Feature Engineering

The data set contains 83 features but all the features are not important for model building and also increases computational time. So, it is important to select most important features for classification. In this study two effective feature selection methods namely, information gain and gain ratio have been applied on the dataset for selection of important features.

### 5.3 Evaluation Matrices

A confusion matrix [12] is a performance measurement tool for classification problems, used to assess how well a machine learning model is performing. It is a table that describes the performance of a classification model by comparing the predicted class labels with the actual class labels from a test set. The four components of matrix are True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN). Based on these values the following performance measurements can be made:

Accuracy: Measures the overall correctness of the model.

Accuracy = ( TP + TN ) / ( TP + TN + FP + FN )

Precision (or Positive Predictive Value): Measures how many of the predicted positive cases are actually positive.

Precision = ( TP ) / ( TP + FP )

Recall (or Sensitivity, True Positive Rate): Measures how many of the actual positive cases were correctly identified by the model.

Recall = ( TP ) / ( TP + FN )

F1-Score: The harmonic mean of precision and recall, providing a balance between them.

F1-Score = 2 * ( Precision * Recall ) / ( Precision + Recall )

False Alarm Rate: Rate of incorrect positive alerts among all actual negatives.

False alarm rate = ( FP ) / ( FP + TN)

## 6. Results and Discussion

In this study performance of machine learning-based classifiers are evaluated using confusion matrix. The metrics, namely, accuracy, precision, recall, f1-score, and false alarm rate of four machine learning-based classifiers are compared with two types of feature selection methods, namely, information gain and gain ratio that are reported in Table 3. In the table, values in boldface represent the highest value as compared to the rest.

Table 3 provides the values of evaluation metrics namely accuracy, precision, f1-score, and false alarm rate for different models. The model Information Gain feature selection with Random Forest achieves highest accuracy of 0.9913, precision 0.9913, recall 0.9913 and f1-score. Information Gain with Support Vector Machine model reports lowest accuracy of 0.5785, precision 0.7466, recall 0.5785 and f1-score 0.4581. Gain Ratio with Random Forest model reports highest accuracy of 0.9915, precision 0.9915, recall 0.9915 and f1-score 0.9915. Gain Ratio with Support Vector Machine model reports lowest accuracy of 0.5607, precision 0.7405, recall 0.5607 and f1-score 0.4208. In both Information Gain with Extra Tree and Gain Ratio with Extra Tree Model gives the lowest false alarm rate of 0.0052 and 0.0057 respectively. Figure 4 and 5 provides a comparative analysis of the performance of accuracy and recall of four classifiers respectively. Figure 6 and 7 represents confusion matrix of Information Gain With Random Forest Model and Information Gain With Extra Tree Model respectively.

**Table 3.** Accuracy, Precision, f1-score, and False Alarm Rate for different Machine Learning-based classifiers with both Information Gain and Gain Ratio feature selection (The values in boldface represent the highest value as compared to other values)

| Feature Selection Method | Classification Technique | Evaluation Metric | | | | |
|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1-Score | FAR |
| Information Gain | **Random Forest** | **0.9913** | **0.9913** | **0.9913** | **0.9913** | 0.0057 |
| | Extra Tree | 0.9904 | 0.9905 | 0.9904 | 0.9904 | **0.0052** |
| | Logistic Regression | 0.8615 | 0.8673 | 0.8615 | 0.8599 | 0. 2293 |
| | Support Vector Machine | 0.5785 | 0.7466 | 0.5785 | 0.4581 | 0.9214 |
| Gain Ratio | **Random Forest** | **0.9915** | **0.9915** | **0.9915** | **0.9915** | 0.0062 |
| | Extra Tree | 0.9909 | 0.9909 | 0.9909 | 0.9909 | **0.0057** |
| | Logistic Regression | 0.7567 | 0.7912 | 0.7567 | 0.7435 | 0.4695 |
| | Support Vector Machine | 0.5607 | 0.7405 | 0.5607 | 0.4208 | 0.9619 |

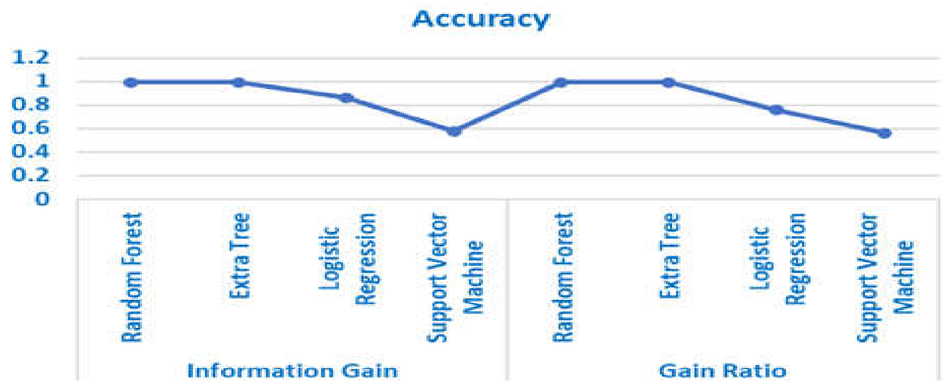


Fig. 4 Analysis of Accuracy

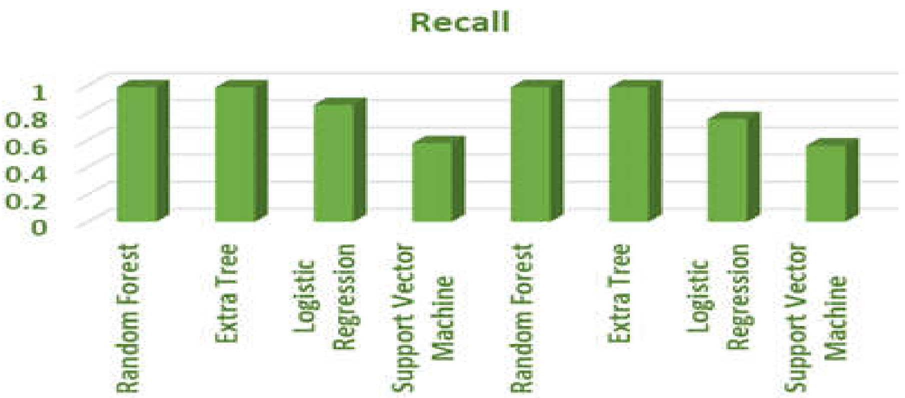
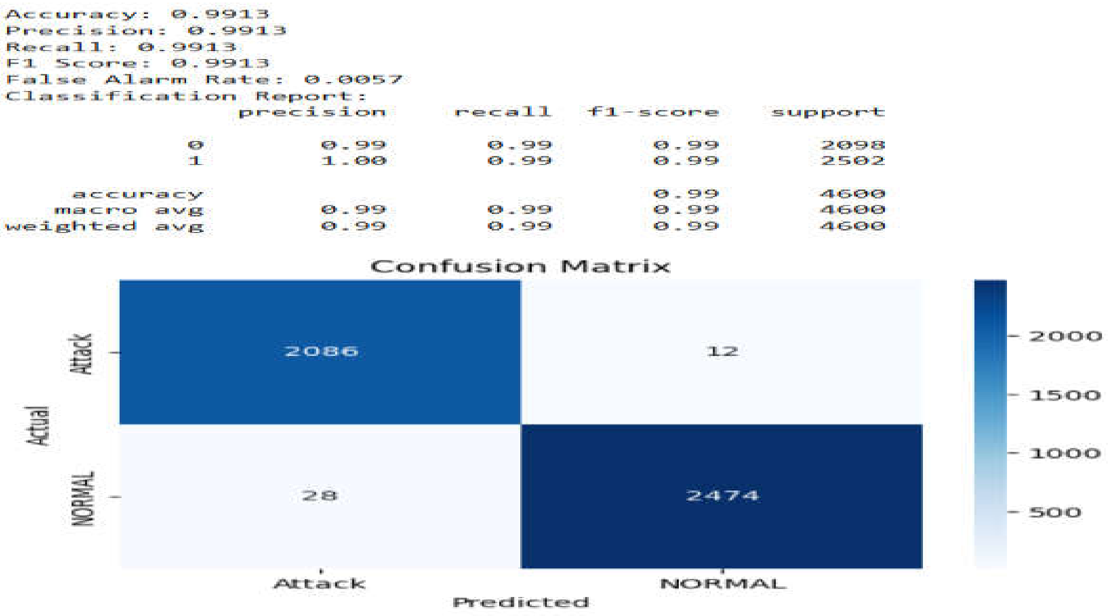


Fig. 5 Analysis of Recall

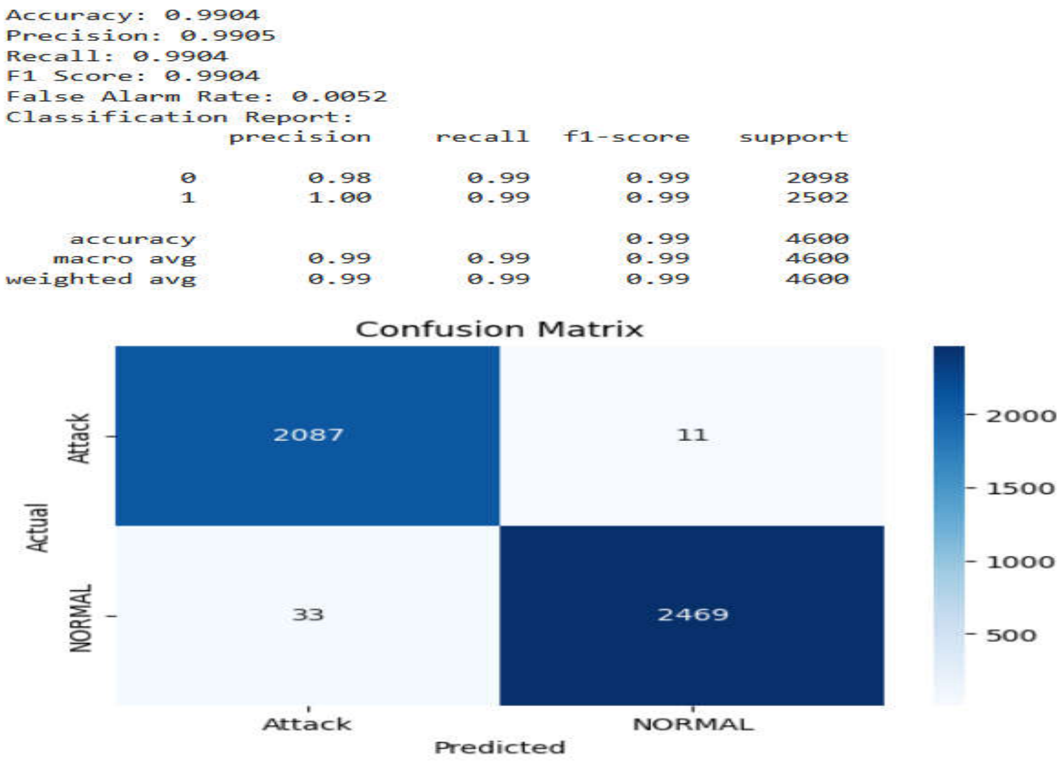


```
Accuracy: 0.9913
Precision: 0.9913
Recall: 0.9913
F1 Score: 0.9913
False Alarm Rate: 0.0057
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      2098
           1       1.00      0.99      0.99      2502

    accuracy                           0.99      4600
   macro avg       0.99      0.99      0.99      4600
weighted avg       0.99      0.99      0.99      4600
```

Fig. 6 Confusion Matrix and Result of Information Gain With Random Forest Model



```
Accuracy: 0.9904
Precision: 0.9905
Recall: 0.9904
F1 Score: 0.9904
False Alarm Rate: 0.0052
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      2098
           1       1.00      0.99      0.99      2502

    accuracy                           0.99      4600
   macro avg       0.99      0.99      0.99      4600
weighted avg       0.99      0.99      0.99      4600
```

Fig. 7 Confusion Matrix and Result of Information Gain With Extra Tree Model

## 6.Conclusion and Future Work

In this work both Information Gain and Gain Ratio feature selection method employed to select the best features from the dataset. Four Machine Learning based techniques namely Random Forest, Extra Tree, Logistic Regression and Support vector Machine(SVM) were applied to build the classification model. The proposed model was evaluated using Splitting Criteria(80% training and 20% testing).Confusion Matrix was used to compare the performance of different four machine learning techniques. Experimental result shows that Random Forest classifier is the best model with Accuracy 0.9913, Precision 0.9913, Recall 0.9913 and F1 Score 0.9913.Both Information Gain and Gain Ratio with Extra Tree give lowest False Alarm Rate of 0.0052 and 0.0057. In future studies the aim is to investigate the performance of the proposed model on individual classes contain within the dataset. Also planned to experiment with different hybrid models in order to enhance the accuracy of intrusion detection.

## References

[1]. W. Alhakami, A. ALharbi, S. Bourouis, R. Alroobaea, and N. Bouguila, "Network Anomaly Intrusion Detection Using a Nonparametric Bayesian Approach and Feature Selection," IEEE Access, vol. 7, pp. 52181-52190, 2019.

[2] Elzaghmouri, Bassam Mohammad, et al. "A Novel Hybrid Architecture for Superior IoT Threat Detection through Real IoT Environments." *Computers, Materials & Continua* 81.2, 2024.

[3] M. Bhavsar, K. Roy, J. Kelly, and O. Olusola, ‘‘Anomaly-based intrusion detection system for IoT application,’’ Discover Internet Things, vol. 3, no. 1, 2023.

[4] M. M. Khan and M. Alkhathami, “Anomaly detection in IoT-based healthcare: machine learning for enhanced security,” *Scientific reports* , vol. 14, no. 1, p. 5872, 2024.

[5] M. M. Inuwa and R. Das, “A comparative analysis of various machine learning methods for anomaly detection in cyber attacks on IoT networks,” Internet Things, vol. 26, p. 101162, 2024.

[6] A. Abusitta, G. H. de Carvalho, O. Abdel Wahab, T. Halabi, B. C. M. Fung, and S. Al Mamoori, “Deep learning-enabled anomaly detection for IoT systems,” *Internet of Things* , vol.21, 2023.

[7] L. Breiman Random Forest, Machine Learning, 45, p.5–32, 2001

[8] O. Sagi, L. Rokach, “Explainable decision forest: transforming a decision forest into an interpretable tree”, *Information Fusion,* 61 , p.124-138, 2020.

[9] D. Caigny, K. Arno, W. Koen, and D. Bock, “A new hybridclassifcation algorithm for customer churn prediction basedon logistic regression and decision trees,” European Journal ofOperational Research, vol. 269, no. 2, pp. 760–772, 2018.

[10] Vapnik, V. (1996). The nature of statistical learning theory. Springer, New York

[11FF] https://archive.ics.uci.edu/dataset/942/rt-iot2022

[12] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, San Francisco, 2006.