ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 3(02), February 2024, pp. 16-25

Article Title: **Classification of IoT Network Traffic using Random Forest Classifier**

Classification of IoT Network Traffic using Random Forest Classifier

S. Ragul[1,*], S. Tamilselvi[2]

[1]Assistant Professor, Department of Electrical and Electronics Engineering, Chettinad College of Engineering and Technology, Karur, ragulsambath22@gmail.com

[2]Associate Professor, Department of Electrical and Electronics Engineering, SSN College of Engineering, Kalavakkam, tamilselvis@ssn.edu.in

ABSTRACT

The swift advancement of the Internet of Things (IoT) has ushered in a wealth of benefits, allowing countless interconnected devices to interact and exchange data effortlessly. Previously, network traffic including unusual patterns, was mainly produced by established, secure endpoints with strong security features, like smartphones. With the advent of the Internet of Things (IoT) no matter how small or intricate device, now has the capability to produce unusual levels of network activity. One of the biggest challenges facing the IoT industry is network traffic, which can have a negative impact on the overall performance of IoT devices and systems. To address this issue, a random forest classifier has been developed specifically for classifying IoT data. Extra Trees offer a significant benefit by minimizing bias. This is achieved by randomly sampling from the entire dataset when building the trees. Random Forest is a widely recognized machine learning technique which favours accuracy, reliability, flexibility and scalability. The process of data preprocessing involves transforming unrefined data into a refined dataset. Chi-square based feature extraction is utilized to extract relevant information and this technique enhances classification by selecting the most important features from the extraction regions. In the end, the chosen characteristics are inputted into both an extra tree and random forest classifier to ensure precise categorization and the implementation of this endeavor is carried out utilizing Python programming.

keywords: Chi-square, Extraction, data pre-processing, random forest classifier, Extra trees, Categorization.

1 Introduction

IoT also known as the Internet of Things, pertains to the interconnected system of devices and the technology enabling communication between devices and the cloud, as well as between the devices themselves. The IoT utilizes sensors and communication technologies to gather and transmit real-time data, facilitating quick data processing and informed decision-making. Additionally, IoT has the potential to revolutionize the energy industry by transitioning it from a centralized to a decentralized, intelligent, and interconnected energy network. Traditional networks have several limitations when it comes to routing, flexibility, manageability and extensibility and these shortcomings will pose a significant challenge for IoT traffic [1]. The

modern Internet of Things (IoT) gadgets gather data from the physical world and offer insights into the surrounding environment. As a result, IoT traffic is expected to make up a significant portion of overall Internet activity. Despite this, the exact effects of this increased traffic on networks remain largely uncertain. Due to limited resources and potential misconfigurations, IoT devices are vulnerable to cyberattacks [2-3]. An intrusion detection system (IDS) can offer additional security to an IoT network by detecting and preventing security attacks. However, a common issue with existing IDSs is the rise in false alarms when identifying zero-day anomalies [4]. Due to this reason, random forest algorithm is taken into account and it is a versatile tool that creates a multitude of decision tree models on different subsets of the data, ultimately enhancing predictive accuracy and reducing the risk of overfitting [5]. The issue of overfitting in a single decision tree is effectively addressed by implementing the Random Forest algorithm. Studies have shown that the Random Forest algorithm outperforms the Extra Trees algorithm in terms of prediction accuracy [6]. IoTTGen (Internet of Things Traffic Generator) is a tool that is used to simulate diverse IoT environments by modeling the behavior of sensor devices [7]. Enhancing IoT Traffic Classification in Smart Environments through Network and Statistical Features deals with an advanced machine learning method with a dual-stage process is employed, utilizing two distinct sets of features. The first set consists of statistical features like packet inter-arrival time, burstiness rate and traffic flow rate. The second set encompasses more detailed network features such as IP addresses, IP protocol type, port numbers, MAC addresses, Time-to-Live (TTL) and packet size. Moving forward, enhancements involves incorporating additional machine learning strategies like Kmeans clustering in conjunction with classification algorithms [8]. The utilization of SDN in FiWi-IoT smart environment network traffic classification with supervised machine learning models offers numerous advantages. By leveraging the overarching perspective provided by SDN and addressing the requirements of traffic flows, a refined system has been developed to enhance the multipath transmission of IoT applications. SDN-Enabled FiWi-IoT Smart Environment Network Traffic Classification Using Supervised ML Models gives a merits of Using the global view of SDN and the need for traffic flows, an optimized scheme is built for the multipath transmission of IoT applications, Detecting anomalies plays a crucial role in uncovering valuable business insights and ensuring the smooth operation of essential functions. For future, the development of anomaly detection serves as a vital mechanism for identifying fraudulent activities, network disruptions and other noteworthy irregularities [9].

Detecting Malicious Bot-Generated Traffic in IoT Networks through Machine Learning Approaches is proposed and this technique is highly effective and consistently delivers over 96% success rates on average. The C4.5 decision tree and Random Forest machine learning algorithms show great potential when applied to the Bot-IoT dataset. While SVM and Naive Bayes algorithms also perform well, they fall slightly short in comparison to the robust performance of the C4.5 decision tree and Random Forest algorithms [10]. A novel method for

classifying malware using vision-based techniques is proposed to overcome the limitations of existing malware detection systems. This approach leverages the advantages of deep transfer learning techniques, along with fine-tuning and ensemble strategies, to enhance detection and classification accuracy without the need to build models from the ground up. In the future, consider implementing federated learning, a method that allows for the deployment of machine learning and deep learning models on IoT sensors. This approach is beneficial due to its simplicity and distributed nature, as it avoids the need to process IoT data on a centralized server [11-12]. Analyzing the length of data packets to classify flows in IoT using ensemble learning techniques that evaluate the runtime efficiency of the classifier and demonstrate that a C implementation, capable of multi-threading and generated directly from the trees, that effectively manage high-speed networks. For future, Investigate the correlation in timing between various flows associated with a single device where a challenge lies in identifying patterns within the opening and closing of flows by the device [13-14]. The system assesses the security implications of Distributed Denial of Service (DDoS) attacks by leveraging Machine Learning algorithms on network traffic data. The previous research highlights the effectiveness of using the Random Forest Regressor (RFR) feature selection technique in enhancing the accuracy of Machine Learning models for detecting malicious attack traffic. However, it delves into the boundaries and prospective advancements for systems that detect and address network anomalies in the Internet of Things (IoT) [15]. This paper proposes a new approach for Categorizing Internet of Things (IoT) network data with Random Forest Algorithm. Building upon existing research, this paper introduces a novel contribution:

- Random Forest favours exceptional accuracy, reliability, flexibility and scalability which widely recognized machine learning technique.
- Extra Trees offer a significant benefit by minimizing bias. This is achieved by randomly sampling from the entire dataset when building the trees.

2 Proposed System Description

As cyberattacks become more advanced and elusive on IoT networks, it is crucial to enhance intrusion detection methods to effectively combat evolving threats. The Internet of Things (IoT) demands heightened focus due to its pivotal role in the ongoing digital transformation, extending beyond services into physical product industries.

Data preprocessing is the essential step of converting real-world data into a digital format that easily analyzed, displayed and stored in a computer system. This transformation is crucial for effective processing in machine learning algorithms. To achieve accurate classification results, the extra tree and random forest classifiers are utilized. These classifiers have shown to produce reliable predictions. However, the process requires extensive training time, especially when dealing with large and unbalanced datasets at the packet level.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 3(02), February 2024, pp. 16-25

Article Title: Classification of IoT Network Traffic using Random Forest Classifier

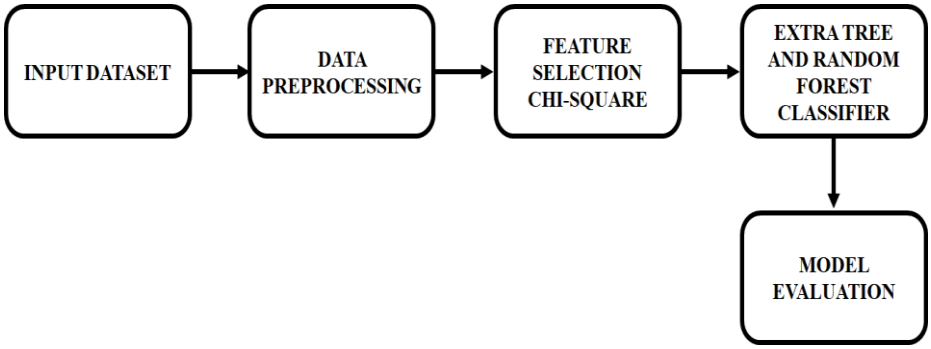

Figure 1: *Proposed Block Diagram*

In order to extract relevant features from the data, the Chi-square Method is employed. This method helps in identifying the most significant attributes that contribute to the classification process.

3 System Modeling

3.1 Data preprocessing

The initial phase of Machine Learning involves a crucial process known as Data Pre-processing.

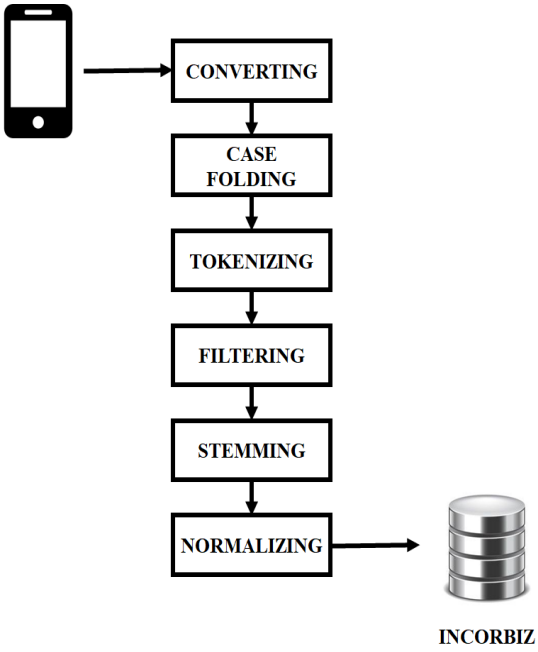

Figure 2: *Steps in data preprocessing*

This step involves transforming or encoding the data to make it suitable for efficient analysis by the machine. Essentially, this step ensures that the model algorithm effectively interpret and

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 3(02), February 2024, pp. 16-25

Article Title: Classification of IoT Network Traffic using Random Forest Classifier

analyze the data features. Data pre-processing plays a crucial role in enhancing the overall performance of a Supervised Machine Learning Algorithm. The quality of training data has a significant impact on the algorithm's ability to generalize effectively, especially as the input space dimension increases exponentially.

Data processing is carried out for the following reasons:

- Addressing data-related issues result in discrepancies and hinder the ability to conduct thorough data analysis.
- Deriving essential and valuable insights from the provided dataset.

3.2 Chi-square test for Feature Selection

The chi-square ($\chi 2$) statistic is a tool used to evaluate the goodness of fit of a model against real-world data. It is important that the data utilized for calculating the chi-square statistic is both random and raw, as well as being mutually exclusive and drawn from independent variables. Additionally, the sample size must be large enough to ensure the accuracy of the results. When considering the possibility that there are no variations between the groups within the population, the calculated test statistic from the data will adhere to a chi-square distribution. The formula for chi-square is:

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

c = degrees of freedom
O = observed values
E = Expected values

3.3 Random Forest Classifier

The Random Forest algorithm is widely used in machine learning and falls under supervised learning. This versatile algorithm is suitable for addressing both Classification and Regression tasks within the field of ML. Its foundation lies in ensemble learning, a method that involves blending multiple classifiers to tackle intricate problems and enhance model performance. A Random Forest is a classification model that consists of multiple decision trees built on different subsets of the dataset. By averaging the predictions of these trees, the model enhances the accuracy of its predictions.

Below are some of the points for which random forest classifier:

- It requires a shorter training period compared to alternative algorithms.
- It accurately forecasts results with precision, even when handling extensive datasets, demonstrating efficient performance.
- It is capable of preserving precision even in the absence of a significant amount of data.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 3(02), February 2024, pp. 16-25

Article Title: Classification of IoT Network Traffic using Random Forest Classifier

4 Result and discussion

The proposed paper focuses on categorizing IoT network traffic through the utilization of a random forest classifier. Python, a dynamically typed language renowned for its implementation approach, is a versatile high-level programming language with multiple applications.

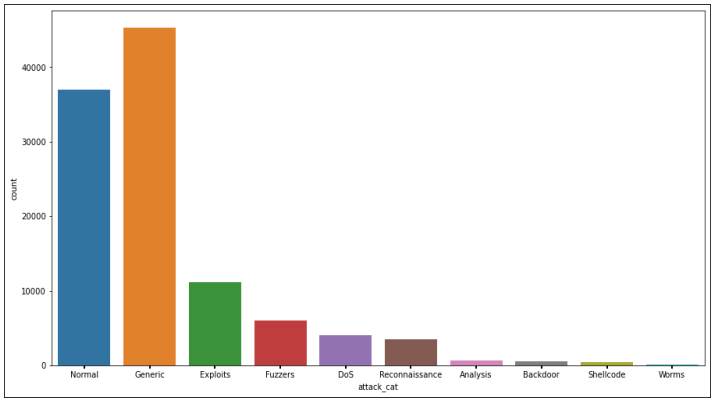

Figure 3: *Attack Cat*

In the illustration provided in Figure 3, different levels of attacks on IoT networks are depicted. This visual representation takes into account both the frequency of attacks and the categories of attacks.

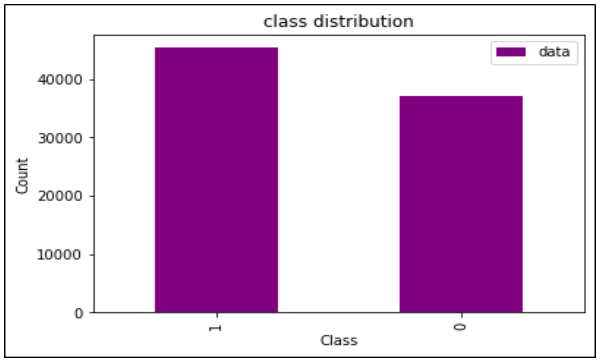

Figure 4: *Class Distribution*

Figure 4 illustrates the graph of class distribution. A class distribution is essentially a mapping of class values (0 or 1) to the corresponding number of randomly generated examples that should be included in the dataset.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 3(02), February 2024, pp. 16-25

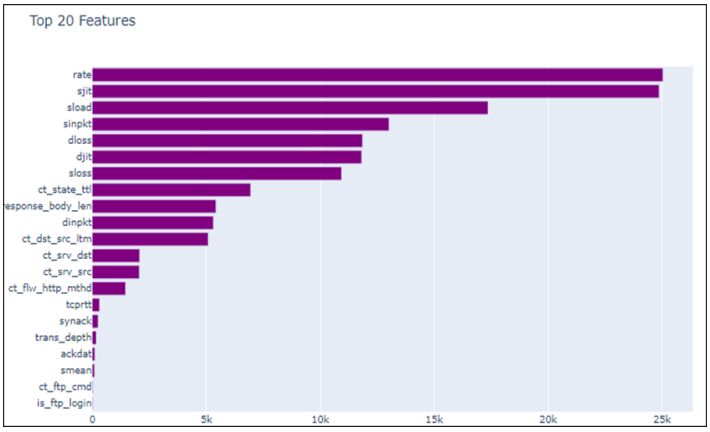

Figure 5: *Top Features*

In the illustration provided in Figure 5, a clear visual representation is chosen using the CHI-square method. The CHI-square formula is a statistical tool used to compare multiple data sets.

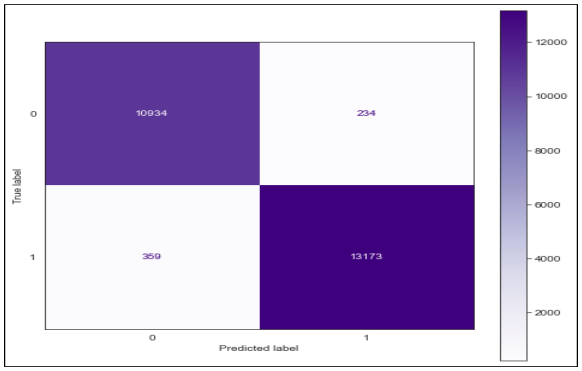

Figure 6: *Performance matrix for extra tree*

Figure 6 depicts the concept of Extra Trees, which is a technique in supervised machine learning that leverages decision trees. The diagram showcases the comparison between the true label and the predicted label generated by the Extra Trees algorithm.

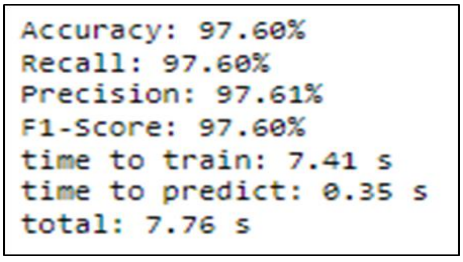

Figure 7: *Confusion matrix for extra tree*

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 3(02), February 2024, pp. 16-25

Article Title: Classification of IoT Network Traffic using Random Forest Classifier

In the illustration presented as Figure 7, the confusion matrix for the extra tree model is displayed. Various performance metrics such as accuracy, recall, precision and F1-score observed in this visual representation.

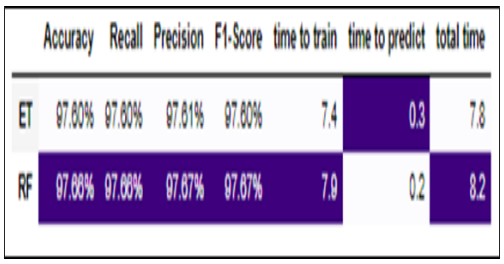

Figure 8: *Model Performance*

Figure 8 depicts the model performance of the confusion matrix for IoT network by using the random forest classifier.

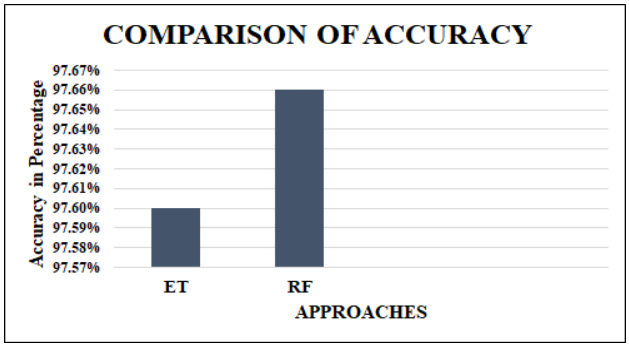

Figure 9: *Comparison of accuracy*

Figure 9, a comparison is shown between the accuracy of Extra trees Classifier (ET) and Random Forest Classifier (RF). The accuracy of ET classifier ranges of 97.6% and with the implementation of the proposed system, the accuracy of RF classifier improves in the range of 97.66%.

5 Conclusion

The proposed approach for this paper involves utilizing a random forest classifier to categorize IoT network traffic. During a span of 26 weeks, the equiption of a smart environment with 28 distinct IoT devices and gathered traffic data non-stop. In situations where devices experience limited bandwidth, the importance of their data flows within the dataset is consequently decreased. Data preprocessing involves transforming data into a format that is more easily and efficiently processed by machine learning algorithms. The use of additional tree and random forest classifiers has been shown to be effective in achieving accurate classification results. The findings presented above vividly demonstrate the difference between the confusion and

performance matrices, with an impressive increase in accuracy of approximately 97.66% achieved when utilizing the random forest classifier. This successful demonstration showcases the effectiveness of the classifier in enhancing accuracy.

References

1. Naser Hossein Motlagh; Mahsa Mohammadrezaei; Julian Hunt; Behnam Zakeri, Year: 2020, "Internet of Things (IoT) and the Energy Sector", Energies, Vol: 13, pp. 494.

2. Hung Nguyen-An; Thomas Silverston; Taku Yamazaki; Takumi Miyoshi, Year: 2021, "IoT Traffic: Modeling and Measurement Experiments", IoT, Vol: 2, pp. 140 – 162.

3. Qasem Abu Al-Haija; Ahmad Al-Badawi, Year: 2022, "Attack-Aware IoT Network Traffic Routing Leveraging Ensemble Learning", Sensors, Vol: 22, pp. 241.

4. Zeeshan Ahmad; Adnan Shahid Khan; Kashif Nisar; Iram Haider; Rosilah Hassan; Muhammad Reazul Haque; Seleviawati Tarmizi; J. J. P. C. Rodrigues, Year: 2021, "Anomaly Detection Using Deep Neural Network for IoT Architecture", Applied Science, Vol: 11, pp. 7050I.

5. Ibbad Hafeez; Markku Antikainen; Aaron Yi Ding; Sasu Tarkoma, Year: 2020, "IoT-KEEPER: Detecting Malicious IoT Network Activity Using Online Traffic Analysis at the Edge", in IEEE Transactions on Network and Service Management, Vol: 17, no: 1, pp. 45 – 59.

6. S. M. Abu Adnan Abir; Adnan Anwar; Jinho Choi; A. S. M. Kayes, Year: 2021, "IoT-Enabled Smart Energy Grid: Applications and Challenges", in IEEE Access, Vol: 9, pp. 50961 – 50981.

7. Elaiyasuriyan Ganesan; I-Shyan Hwang; Andrew Tanny Liem; Mohammad Syuhaimi Ab-Rahman, Year: 2021, "SDN-Enabled FiWi-IoT Smart Environment Network Traffic Classification Using Supervised ML Models", Photonics, Vol: 8, pp. 201.

8. Muhammad Shafiq; Zhihong Tian; Ali Kashif Bashir; Xiaojiang Du; Mohsen Guizani, "CorrAUC: A Malicious Bot-IoT Traffic Detection Method in IoT Network Using Machine-Learning Techniques", in IEEE Internet of Things Journal, Vol: 8, no: 5.

9. Safa Ben Atitallah; Maha Driss; Iman Almomani, Year: 2022, "A Novel Detection and Multi-Classification Approach for IoT-Malware Using Random Forest Voting of Fine-Tuning Convolutional Neural Networks", Sensors, Vol: 22, pp. 4302.

10. M. Ganesh Karthik; M. B. Mukesh Krishnan, Year: 2021, "Hybrid random forest and synthetic minority over sampling technique for detecting internet of things attacks", Journal of Ambient Intelligence and Humanized Computing.

11. Gennaro Cirillo; Roberto Passerone, Year: 2020, "Packet Length Spectral Analysis for IoT Flow Classification Using Ensemble Learning", in IEEE Access, Vol: 8, pp. 138616 –138641.

12. Hayelom Gebrye; Yong Wang; Fagen Li, Year: 2023, "Traffic data extraction and labeling for machine learning based attack detection in IoT networks", International Journal of Machine Learning and Cybernetics, Vol: 14, pp. 2317 – 2332.

13. Rami J. Alzahrani; Ahmed Alzahrani, Year: 2021, "Security Analysis of DDoS Attacks Using Machine Learning Algorithms in Networks Traffic", Electronics, Vol: 10, pp. 2919.

14. Thanh Noi Phan; Verena Kuch; Lukas W. Lehnert; Year: 2020, "Land Cover Classification using Google Earth Engine and Random Forest Classifier—The Role of Image Composition", Remote Sensor, Vol: 12, pp. 2411.

15. V. Jackins; S. Vimal; Madasamy Kaliappan; Mi Young Lee, Year: 2021, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes", The Journal of Supercomputing, Vol: 77, pp. 5198 – 5219.