ISSN 2278-3091
Volume 12, No.6, November - December 2023
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse021262023.pdf
https://doi.org/10.30534/ijatcse/2023/021262023

Network Source Identification Mechanism for IoT Devices Using Machine Learning Techniques

Isaac Terngu Adom[*1], Aamo Iorliam[1], Daniel Terkura Kumaga[1], and Samera Uga Otor[1]
[1]Department of Mathematics and Computer Science, Benue State University, Nigeria, iadom@bsum.edu.ng
[1]Department of Mathematics and Computer Science, Benue State University, Nigeria, aiorliam@bsum.edu.ng
[1]Department of Mathematics and Computer Science, Benue State University, Nigeria, danterkum16@gmail.com
[1]Department of Mathematics and Computer Science, Benue State University, Nigeria, sotor@bsum.edu.ng
[*]Corresponding author: iadom@bsum.edu.ng

Received Date : October 21, 2023 Accepted Date: November 25, 2023 Published Date: December 06, 2023

ABSTRACT

The rapid progress and evolution of the Internet of Things (IoT) have led to a significant increase in the occurrence of security gaps. Pinpointing the source of network traffic coming from IoT devices can be challenging, but doing so can reduce security risks. This study proposes a network traffic source identification mechanism that leverages machine learning (ML) techniques to accurately determine the source of network traffic. The study utilizes a diverse dataset obtained from a purpose-built IoT/IIoT testbed and employs feature extraction, model development, and evaluation techniques. By utilizing network traffic features, a range of classifiers, including LGBMClassifier (LGBM), CatBoostClassifier (CB), RandomForestClassifier (RF), ExtraTreesClassifier (ET), KneighborsClassifier (KNN), and DecisionTreeClassifier (DT), were trained and evaluated. The results demonstrate exceptional performance across the classifiers, with high accuracy, precision, recall, and F1 scores achieved in identifying the source of network traffic. Among the classifier models, LGBM achieved the best accuracy value of 0.99999857, precision value of 0.99999859, and F1 score of 0.999998803, with CB achieving the best recall of 0.999997875. Some of these results are novel, and others performed better than existing systems. The findings of this study contribute to source identification, ensure the accountability of IoT network users, and provide insights into developing better defenses against security threats in the IoT domain.

Key words: IOT, Machine Learning, Network, Security, Source Identification

1. INTRODUCTION

Numerous devices are connecting to the Internet as a result of the growing popularity of the Internet of Things (IoT). IoT devices include detectors, controllers, sensors, actuators, and other appliances that are connected to the internet. These devices, such as the Google Home Voice Controller, Amazon Dash Button, August Smart Lock, Kuri Mobile Robot, etc., are connected using wired or wireless connections and, hence, can be controlled with the aid of other computing devices like smartphones and computers (even from distant positions). Many smart devices, homes, and Personal Digital Assistants (PDAs) are ubiquitous in our society. These devices are capable of generating large volumes of data that need to be protected for confidentiality, integrity, and availability purposes. Several communication technologies and protocols are used in the context of the IoT, including Internet Protocol Version 6 (IPv6), Low Power Wireless Personal Area Networks (6LoWPAN), ZigBee, Bluetooth Low Energy (BLE), Z-Wave, and Near Field Communication (NFC) [1]. While these interconnections lead to scores of benefits, they also constitute security risks as they create a larger attack space for cybercriminals. One such security concern is the potential to spot the source of network traffic from IoT devices. For example, an attacker could use an endangered IoT device to start an assault in an attempt to damage the computer network, and it could be hard to pinpoint the device responsible for the attack without an efficacious source identification mechanism. Traditional methods of IoT device source identification, such as RFID, barcodes, and IP addresses, have been used in the past. However, with the continuous development of IoT and the increasing number of connected objects, there is a need for improved identification methods. These methods incorporate technologies like fingerprinting and ML, aiming to enhance identification [2]. This paper aims to build a source identification mechanism using machine learning techniques for IoT gadgets so as to improve the safety and accountability of IoT network users. The primary idea is to use network traffic features to train machine learning models that can correctly identify the source of network traffic from IoT devices. As a result of that, it can enhance the security of IoT networks by allowing fast and error-free identification of the source of any dubious network scheme. The proposed mechanism involves assembling network traffic data from IoT gadgets and taking out key attributes such as source and destination IP host, TCP checksum, etc. These attributes are then used to train a machine learning model to correctly identify the source of network traffic. Overall, this paper is concentrated on building an empirical solution to address a notable security challenge in the IoT realm, and it exploits machine-learning techniques to attain this goal. The rest of the paper is organized into the following sections: Section 2 presents related work to the research. Section 3 covers the methodology of our study. The implementation experiments are carried out in Section 4, with Section 5 presenting the conclusion and future research direction.

2 RELATED WORKS

Several methods have been proposed for identifying the sources of IoT devices. These methods include packet analysis, traffic correlation, and flow-based analysis. Packet analysis involves the inspection of packet headers to identify the source of the device. Traffic correlation involves the analysis of traffic patterns to identify the source of the device. The flow-based analysis involves the analysis of network flows to identify the source of the device. Non-cryptographic device identification with rogue device detection functions, in particular from the perspective of network operators and cybersecurity surveillance agents, are required to secure the IoT ecosystem in addition to conventional cryptographic mechanisms such as message authentication codes, digital signatures, challenge-response sessions, etc. [3]. [4] carried out a study to identify the vendors of IoT devices and proposed a novel and alternative method that uses widely accessible WebUI login pages with distinguishing vendor-specific characteristics as the data source and an ensemble learning model built on a combination of convolutional neural networks (CNN) and deep neural networks (DNN). The experimental results showed that the ensemble learning model can determine whether a device is from a vendor that appeared in the training dataset with 99.1% accuracy and 99.5% F1-Score, and if the answer is yes, it can identify that vendor with 98% accuracy and 98.3% F1-Score. [5] presented a new strainer feature selection technique based on NSGA-III to select effective features for IoT device identification. The technique was gauged by using an actual smart home IoT data set and three distinct ML models. A deep/dynamic flow inspection mechanism was employed to effectively take out flow-related statistical attributes based on a very detailed study. The experimental findings demonstrated the efficiency of their suggested method and the feature selection algorithm, which only requires the use of six features to achieve 99.5% accuracy over three minutes. Also, in [6], mechanisms and protocols for authenticating a device in a network by leveraging ML to classify not only if the device is IoT or not but also the type of IoT device attempting to connect to the network were implemented, with an accuracy of over 95%. [7] proposed an IoT-Portrait, a mechanism for automatically identifying IoT devices based on a transformer neural network that extracts useful information from IoT devices to accurately classify them. To address privacy concerns and optimize resource usage, the framework employs class incremental learning, which enables the integration of new devices into the network while preserving knowledge of previously used devices. [8] proposed a novel approach to source identification of IoT devices by plotting graphs of IAT values for packets and using deep learning techniques, specifically Convolutional Neural Networks (CNN), to identify the devices. The work focused on Device Fingerprinting (DFP) using Inter-Arrival Time (IAT), which is the time interval between consecutive packets. They used the Raspberry Pi as a router to capture packet information from connected Apple devices, specifically the iPad 4 and iPhone 7 Plus. They then created IAT graphs for these devices and trained a CNN model to recognize and classify the devices. The results showed an accuracy of 86.7% in device identification using the suggested method. [9] proposed AUDI, which operates autonomously after initial setup, learning without human intervention or labeled data, to identify previously unseen device types in an IoT network. AUDI is a system for quickly and effectively identifying the type of device in an IoT network by analyzing their network communications. Through systematic experiments with 33 commercial IoT devices, the authors demonstrated that AUDI is efficient (98.2% accuracy) at identifying the type of a device in any mode of operation or stage of the device's lifecycle. [10] introduced a system called System IDentifier (SysID), which uses any single packet that originated from the device to detect its kind. A genetic algorithm (GA) was used to determine relevant features in different protocol headers and then deploy various machine learning (ML) algorithms (i.e., Decision Table, J48 Decision Trees, OneR, and PART) to classify host device types by analyzing their network traffic. The researchers experimented with 23 IoT devices, and SysID identified the device type from a single packet with over 95% accuracy, allowing for fully automated classification of IoT devices using their TCP/IP packets without the need for expert input. [11] conducted a study addressing the vulnerabilities of inadequately secured Internet of Things (IoT) devices exposed by recent DDoS attacks with the goal of identifying and understanding the characteristics of IoT devices in order to gain insights into the risks associated with these attacks. To tackle this challenge, the paper proposes a novel method (IP-based) for identifying IoT devices on the Internet. The approach relies on analyzing flow-level network traffic and leveraging information from servers operated by IoT device manufacturers. The authors conducted controlled experiments using their own set of 10 IoT devices and 15 non-IoT devices behind a home router and compared the observed traffic with the device server names and IP addresses. They achieved a detection rate of more than half (6 out of 10) for inactive devices. [12] proposed a mechanism that specifies a set of discriminating features extracted from raw network traffic flows and proposes an LSTM-CNN cascade model for semantic device type identification. The method uses the rich information carried by traffic flows in IoT networks to characterize device attributes. The researchers evaluated their approach by classifying 15 IoT devices into four types with real-world collected network traffic data and achieved an accuracy of 74.8%. It focuses on automatic IoT device classification and seeks to identify new and unseen devices. [13] proposed a mechanism based on a Hierarchical Deep Neural Network (HDNN) that can classify IoT devices into their specific categories and identify new entrants with reasonable accuracy. The proposed HDNN framework distinguishes between IoT devices and non-IoT devices using a feature set specific to IoT traffic. In heterogeneous networks, the proposed HDNN model can accurately classify IoT devices into their respective categories and discriminate between IoT and non-IoT devices with an accuracy of 91.33%. Machine learning techniques have been widely used for the source identification of IoT devices. The most commonly used techniques include Logistic Regression (LR), decision trees, random forests, support vector machines (SVM), and CNN. [14] proposed a device identification method for IoT based on device profiling. The study utilized real-time data from IoT devices in a lab setting to identify the devices. The method incorporated a combination of sensor measurements, statistical feature sets, and analysis of header information for device identification. ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), were employed for classification. The results showed accuracies of 81% for RF, 86% for SVM, and 81% for LR in identifying the devices. [15] proposed a device identification method for the IoT that addresses the limitations of the passive fingerprinting approach. The existing method primarily focuses on protocol features in packet headers and overlooks the direction and length of packet sequences. In their study, the authors introduced a novel approach based on directional packet length sequences in network flows and a deep convolutional neural network. The packet length sequences capture the size and transmission direction of each packet, enabling the construction of device fingerprints. The CNN is then

employed to extract deep features from these fingerprints. The experimental results demonstrate the effectiveness of the proposed method, achieving device identification with high accuracy, recall, precision, and f1-score, all exceeding 99%. Furthermore, the approach outperforms traditional ML and feature extraction techniques, providing a more intuitive feature representation and a highly effective classification model. In this section, we present the empirical studies that were done to assess the efficacy of various source identification mechanisms for IoT devices using ML techniques.

3. METHODOLOGY

The methodology of the system encompasses the data collection process, data preprocessing, feature extraction, model development, evaluation metrics, and visualizations. Figure 1 shows the proposed methodology.

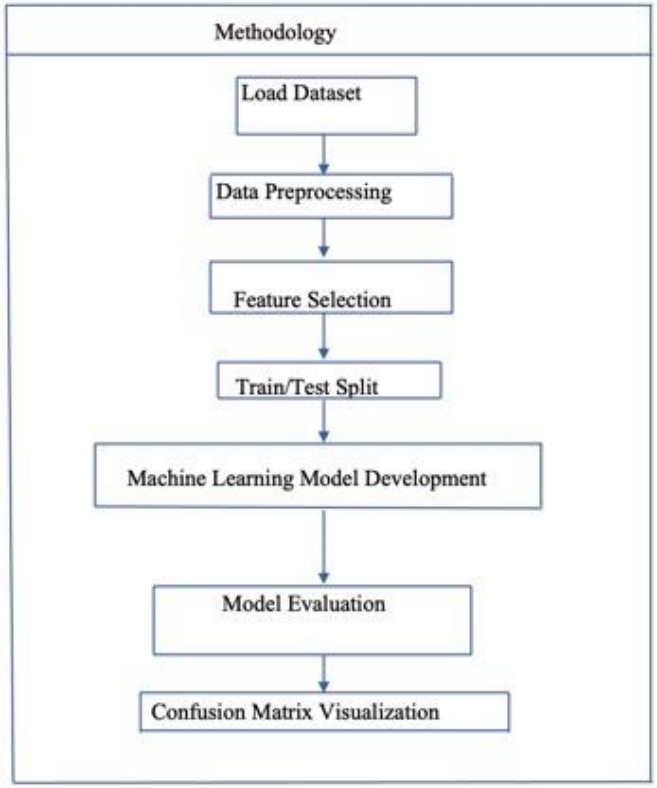

Figure 1: A general overview of the methodology

3.1 Dataset

The dataset used in the project is gotten from 'https://www.kaggle.com/. The dataset is part of the Edge-IIoTset, a comprehensive, realistic cyber security dataset for IoT and IIoT applications. It has been generated using a purpose-built IoT/IIoT testbed that incorporates a diverse set of devices, sensors, protocols, and cloud/edge configurations. The dataset contains data from various IoT devices, including more than 10 types of devices such as low-cost digital sensors for temperature and humidity, ultrasonic sensors, water level detection sensors, pH sensor meters, soil moisture sensors, heart rate sensors, flame sensors, and others. It provides a representative sample of the types of devices commonly found in IoT and IIoT environments. The dataset includes features obtained from different sources, including alerts, system resources, logs, and network traffic. In total, it comprises 1176 features. From these features, 61 new features with high correlations have been extracted for analysis and modeling purposes.

3.2 Data Preprocessing

Preprocessing steps typically involve handling missing values, data normalization, feature extraction, and any other necessary transformations to prepare the data for analysis and modeling. The dataset was shuffled to randomize the rows. Limiting the dataset to the first 100,000 rows, null values were checked, and **c**ategorical variables were converted to numerical variables using label encoding. This is achieved by applying the LabelEncoder from sci-kit-learn to encode the target variable (type) as a numeric.

3.3 Feature extraction techniques

The feature extraction model SelectKBest is used in combination with the f_classif scoring function to select the top k features from the dataset. The number of features to select is set at k = 10. This step aims to identify the most relevant features for modeling and analysis. Before applying the feature extraction model, categorical variables in the dataset are converted to numerical variables by encoding them as integer codes using the cat.codes method. The selected features are then used for training and testing the machine learning models. The feature extraction procedure used in the work is shown in figure 2.

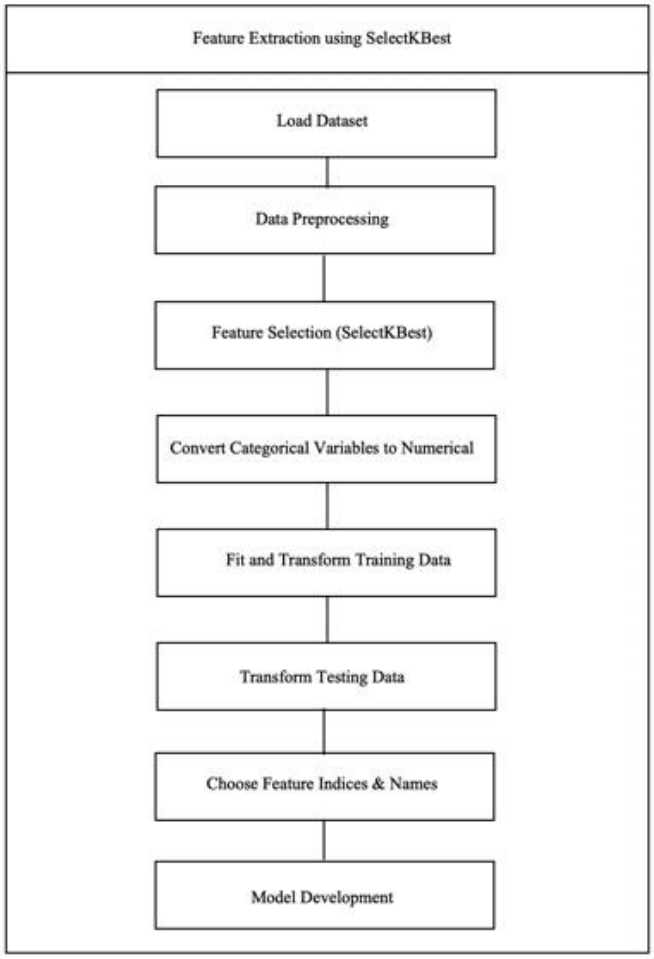

Figure 2: Description of the Feature Selection Process

4 RESULTS AND DISCUSSION

Several machine learning algorithms are implemented to build predictive models for accurate network source identification. The following classifiers are utilized as being best for this implementation: LGBM, CB, RF, ET, KNN, and DT. Each classifier is trained on the training set and evaluated on the testing set. The performance metrics, including recall, accuracy, and precision, are computed to assess the effectiveness of each classifier in accurately identifying the source of network traffic generated by IoT devices. Experiments were carried out on each of the classifiers. The performance of each classifier is evaluated based on the computed accuracy, precision, and recall scores.

4.1.1 The LGBMClassifier Model Result

The LGBMClassifier model achieved high accuracy and performance on the dataset, with an average accuracy score of 0.99999857, precision of 0.9999985949, recall of 0.999990115, and F1 score of 0.999998803. There were few insignificant misclassifications as shown in figure 3.

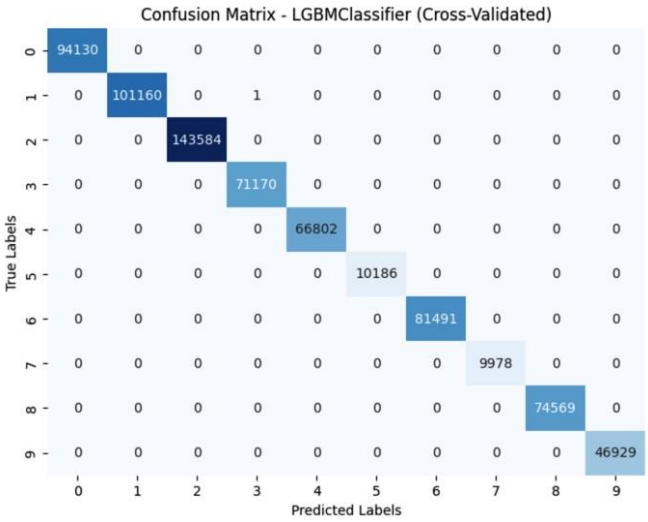

Figure 3: The confusion matrix for LGBMClassifier

4.1.2 The CatBoostClassifier Model Result

The CatBoostClassifier model also achieved perfect accuracy and performance on the dataset, with accuracy score of 0.99999714, precision of 0.99999802, recall of 0.999997875, and F1 score of 0.999997949. Figure 4 shows very few insignificant misclassifications.

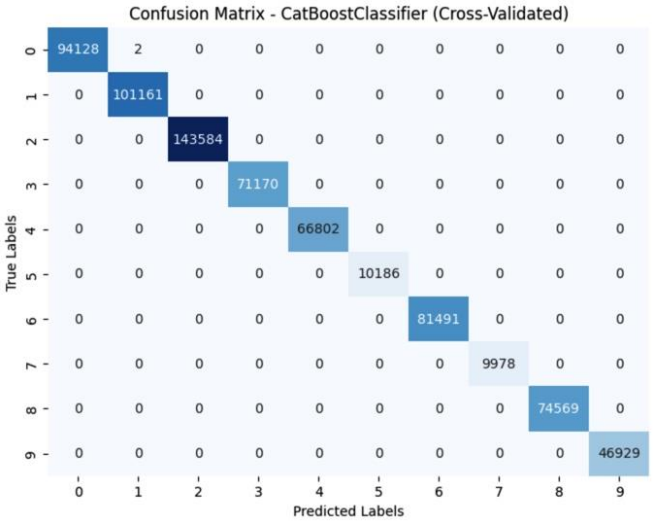

Figure 4: The confusion matrix for CatBoostClassifier

4.1.3 The RandomForestClassifier Model Result

The RandomForestClassifier model demonstrated high accuracy and robust performance. It achieved an average accuracy score of 0.999992857, precision of 0.99997705, recall of 0.9999940159, and F1 score value of 0.9999855. There were negligible misclassifications as shown in figure 5.

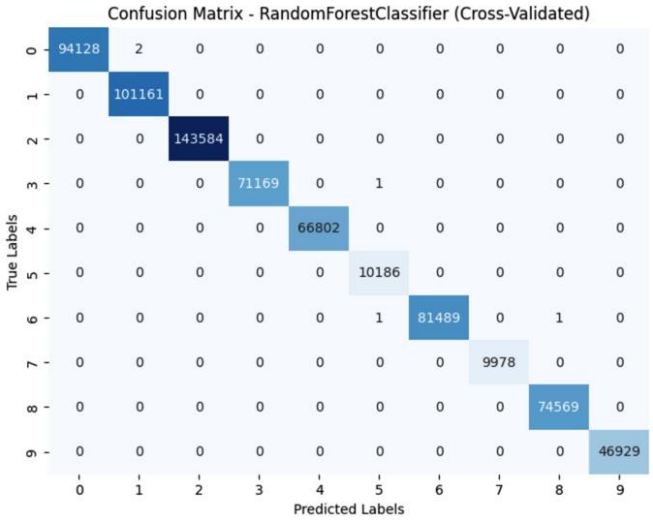

Figure 5: The confusion matrix for RandomForestClassifier

4.1.4 The ExtraTreesClassifier Model Result

The ExtraTreesClassifier model exhibited excellent accuracy and performance, with an average accuracy score of 0.9999942857, precision of 0.9999957869, recall of 0.9999157869, and F1 score value of 0.99999132596. There were negligible misclassifications as shown in figure 6.

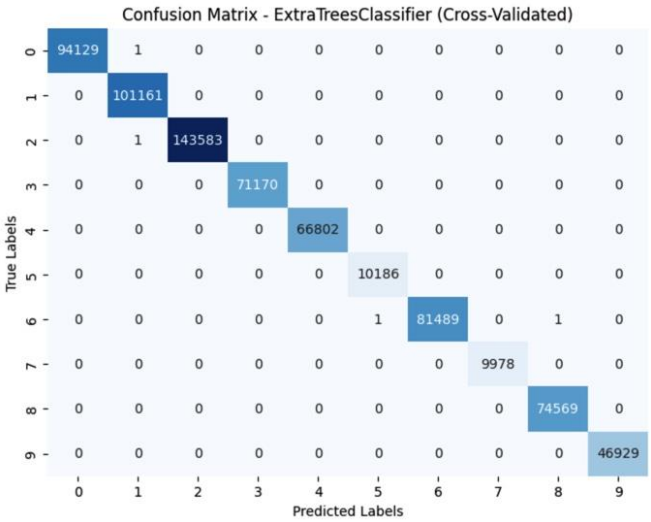

Figure 6: The confusion matrix for ExtraTreesClassifier

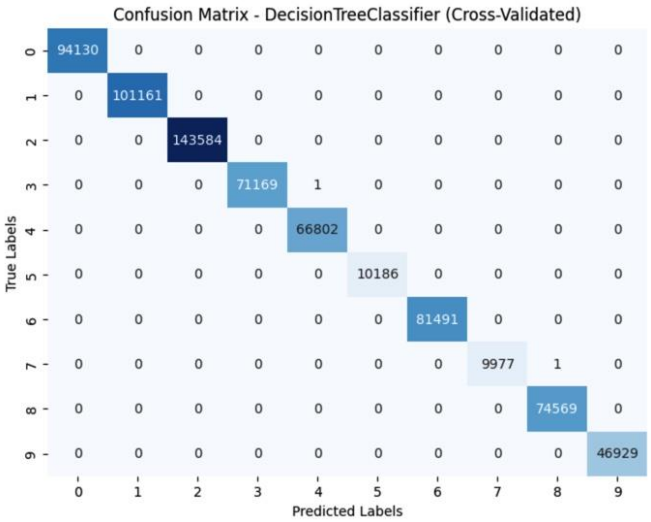

Figure 8: The confusion matrix for DecisionTreeClassifier

4.1.5 The KNeighborsClassifier Model Result

The KNeighborsClassifier model achieved high accuracy and performed well on the dataset, with an average accuracy score of 0.9998657, precision of 0.99979649, recall of 0.999801045, and F1 score value of 0.99979876. There were few misclassifications as shown in figure 7.

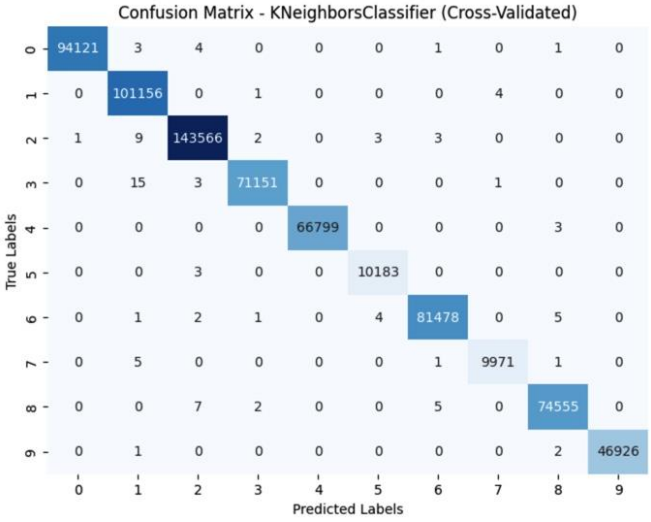

Figure 7: The confusion matrix for KNeighborsClassifier

4.1.5 The DecisionTreeClassifier Model Result

The DecisionTreeClassifier model achieved perfect accuracy and performance on the dataset, with accuracy score of 0.99999714, precision of 0.999997162, recall of 0.99998857, and F1 score value of 0.999992867. As shown in figure 8, there were negligible misclassifications.

4.2 Comparative analysis

To compare the performance of the classifiers, Table 4 presents a comparative analysis of the models' accuracy, precision, and recall scores. This provides a clear understanding of the strengths and weaknesses of each classifier in the source identification. From Table 1, LGBM, CB and DT produced the best results followed by the other classifiers.

Table 1: A Tabular Comparative Analysis of the Average Results

Classifier	Accuracy	Precision	Recall	F1 Score
LGBM	0.99999857	0.9999985949	0.999990115	0.999998803
CB	0.99999714	0.99999802	0.999997875	0.999997949
RF	0.999992857	0.99997705	0.9999940159	0.9999855
ET	0.9999942857	0.9999957869	0.9999157869	0.9999913259 6
KNN	0.9998657	0.99979649	0.999801045	0.99979876
DT	0.99999714	0.999997162	0.99998857	0.999992867

5. CONCLUSION

In this study, a source identification mechanism for IoT devices using ML techniques was proposed and evaluated. The results demonstrated the effectiveness of the mechanism in accurately identifying the source of network traffic generated by IoT devices. The LGBM, CB, RF, ET, KNN, and DT all achieved high accuracy and performance on the dataset, indicating their potential for practical implementation. The source identification mechanism contributes to the field of IoT security by providing a reliable method to detect and locate the sources of network traffic. By promptly identifying potential security threats, the mechanism enables targeted actions to mitigate risks and enhance the security of IoT networks. For future research, other feature selection techniques other than SelectKBest feature extraction should be utilized. Also,

other ML ensemble methods can be explored, and the experiment can be carried out on a large-scale IOT network.

REFERENCES

1. S. Al-Sarawi, M. Anbar, K. Alieyan and M. Alzubaidi. **Internet of Things (IoT) communication protocols: Review**, *8th International Conference on Information Technology (ICIT)*, Amman, Jordan, pp. 685-690, July 2017.
2. S. A. Bkheet, and J.I. Agbinya. **A Review of Identity Methods of Internet of Things (IOT).** *Advances in Internet of Things*, *11*(04), 153–174, October, 2021.
3. Y. Wang, J. Wang, J. Li, S. Niu, and H. Song. **Machine Learning for the Detection and Identification of Internet of Things Devices: A Survey**, *IEEE Internet of Things Journal*, *9*(1), pp. 298–320, January 2021.
4. R. Wang, H. Li, J. Jing, L. Jiang, and W. Dong. **WYSIWYG: IoT Device Identification Based on WebUI Login Pages**. *Sensors*, *22*(13), June 2022.
5. R. Du, J. Wang, and S.A. Li. **A Lightweight Flow Feature-Based IoT Device Identification Scheme.** *Security and Communication Networks*, pp. 1–10, January 2022.
6. K. Gupta. **Machine Learning-Based Device Type Classification for IoT Device Re- and Continuous Authentication.** M.Sc. Thesis, Department of Computer Science and Engineering, University of Nebraska, Lincoln, 2022.
7. J. Wang, J. Zhong, and J. Li. **IoT-Portrait: Automatically Identifying IoT Devices via Transformer with Incremental Learning**. *Future Internet*, *15*(3), March 2023,
8. S. Aneja, N. Aneja and M. S. Islam. **IoT Device Fingerprint using Deep Learning**, *IEEE International Conference on Internet of Things and Intelligence System (IOTAIS)*, Bali, Indonesia, pp. 174-179, October 2018.
9. S. Marchal, M. Miettinen, T.H. Nguyen, A. Sadeghi, and N. Asokan. **AuDI: Toward Autonomous IoT Device-Type Identification Using Periodic Communication**. *IEEE Journal on Selected Areas in Communications*, *37*(6), pp. 1402–1412, March 2019.
10. A. Aksoy and M. H. Gunes. **Automated IoT Device Identification using Network Traffic**, *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, pp. 1-7, May 2019.
11. H. Guo, and J. Heidemann. **IP-Based IoT Device Detection**, Proceedings of the 2018 Workshop on IoT Security and Privacy, pp. 36 – 42, August 2018.
12. L. Bai, L. Yao, S. S. Kanhere, X. Wang and Z. Yang. **Automatic Device Classification from Network Traffic Streams of Internet of Things,** *IEEE 43rd Conference on Local Computer Networks (LCN)*, Chicago, IL, USA, pp. 1-9, October 2018.
13. H. M. S. Zahid, Y. Saleem, F. Hayat, F.A. Khan, R. Alroobaea, F. M. Almansour, M. Ahmad, and I. Ali. **A Framework for Identification and Classification of IoT Devices for Security Analysis in Heterogeneous Network**. *Wireless Communications and Mobile Computing*, pp. 1–16. September 2022.
14. N. Yousefnezhad, A. Malhi, and K. Främling. **Automated IoT Device Identification Based on Full Packet Information Using Real-Time Network Traffic**. *Sensors*, *21*(8), April 2021.
15. Liu, Y. Han and Y. Du. I**oT Device Identification Using Directional Packet Length Sequences and 1D-CNN.** *Sensors*, 22(21), October 2022.