Intrusion Detection System in IoT Network using ML

Soumya Bajpai and Kapil Sharma
Department of Computer Science and Engineering
Amity University Madhya Pradesh, Maharajpura Dang, Gwalior, MP, India
soumyabajpai2013@gmail.com and ksharma@gwa.amity.edu

Brijesh Kumar Chaurasia
Department of Computer Science and Engineering
Pranveer Singh Institute of Technology
Kanpur, UP, India
brijesh.chaurasia@psit.ac.in

Abstract - **Due to the Internet of Things (IoT) device's exponential growth and the resulting expansion of their attack surfaces, hackers can launch increasingly damaging cyber-attacks. The intrusion attempted to drain the target IoT network's resources through malicious activities. For IoT networks, we are utilizing machine learning methods with a carefully categorized dataset to detect intrusion. Firstly, a variety of intrusion detection datasets are presented. Second, we are utilizing the IoTID20 dataset and describing classification features. In the third step, we take some significant characteristics from the dataset. Finally, we apply machine learning methods, such as the linear algorithm, random forest, gradient boost algorithm, and many more, over the dataset to identify anomalies with high accuracy.**

Index Terms - Flow-based intrusion detection, Internet of Things, Intrusion detection, IDS Dataset, Anomaly detection system.

I. Introduction

Computer systems have permeated every aspect of our daily lives, and the IoT has recently attracted a lot of attention in the IT industry due to its many benefits [1]. Actual objects from many industries are connected to the Internet through the IoT. IoT has grown significantly in importance for creating smart infrastructure because of its analytics and interconnection of machinery and personal electronics [2]. Smart infrastructure would have a profound impact on how we manage vital services and run businesses. Due to its exponential growth, the IoTs will turn into a smart object that hackers may exploit to carry out malicious actions, expanding the attack surface of IoT networks. The deployment of IoT devices enables smart infrastructure to operate more reliably and effectively [3]. While substantially cutting operational costs. Anomaly detection in massive data sets is becoming a challenge in network security [4]. Data mining and machine learning techniques have played a significant part in the creation and development of anomaly-based intrusion detection systems. Data mining techniques [5] include regression, classification, clustering, the Learning Association rule, and visualization. The most popular method used in supervised data mining is classification. To properly classify entities, classification aims to build a model using objects that have been effectively categorized. Based on the fusion of smart objects and the Internet, the IoT provides a wide range of applications. The physical components of the infrastructure are combined with smart objects to provide better resource consumption and management. The Internet of Things will be beneficial to many organizations' offerings. The main goal of the IoT is to provide a digital response to social expectations in daily interactions [6]. The main problems in IoT networks include data management, security, cost, scalability, efficiency, application development, and availability. One of these concerns that are thought to pose the biggest challenges to the development of smart infrastructure is security. The most popular datasets for intrusion detection are the DARPA 98/99 [7], [8] datasets, which were produced at MIT Lincoln Lab utilizing an emulated network environment. The network traffic in the DARPA98 dataset spans seven days, while the network traffic in the DARPA99 dataset spans five weeks. The DARPA98/99 datasets, on the other hand, are frequently employed for network intrusion detection, however, they are frequently criticized for having several duplicate items.

The rigorous process used to manufacture both legitimate and malicious network traffic was used to construct the ISCX dataset [9]**.** The dataset contains statistics from seven days of SMTP, FTP, IMAP, POP3, HTTP, and SSH network activity. The dataset's harmful element was created using a multistage attack process. A labeled dataset called ISCX contains both actual network traffic and several intrusion scenarios. The dataset may be downloaded for free from the internet. The NSL-KDD [10] dataset was produced using the KDD99 dataset. Duplicate entries were removed from the KDD99 dataset by the NSL-KDD. In the training data, KDD99 has 78 percent

redundant instances, while in the testing data, it has 75 percent redundant instances.

CICIDS2017 dataset included both legitimate and harmful network traffic from today. The dataset shows both legitimate and malicious network traffic in a trustworthy manner and comprises 80 network characteristics. Five days were spent gathering the data. The dataset's CSV files were created from the Pcap files using the ISCXFlow meter [11] .

The IoT botnet dataset was discussed in[12]using real and fake IoT networks. IoT services including weather station, smart fridg, smart thermostat, motion-activated lighting, and remotely operated garage door were created using the Node-Red tool. Five IoT devices are often used in a smart home setup. These devices are controlled locally and are linked to the cloud infrastructure using the Node-Red tool to provide consistent traffic. So, to detect malicious behavior, we use machine learning algorithms on the IoTID20 dataset is needful.

The organization of this work is as follows. The machine learning approaches to detect intrusion and datasets is evaluated in section II. Proposed methodology is presented in section III. Results are evaluated in section IV and the work was concluded with a discussion of expectations for the future in section V.

II. Related Work

In 2008, a flow-based intrusion detection dataset for high-speed networks is introduced in [13]. The first publicly accessible label dataset for flow-based intrusion detection was created with this dataset. The CICDDOS2019 dataset was created by [9] and contains up-to-date normal and malicious DDOS network traffic. To produce realistic regular background network traffic, they employed a B-Profile system. The collection contains 12 DDoS assaults that are harmful. The CICDDOS2019 dataset may be found on the website [14] . This dataset contains detailed information on IP addresses and harmful network traffic. An IoT botnet dataset was created by [15]. Nine commercial IoT devices and two IoT-based botnets, BASHLITE and Mirai, were used to create the dataset. A total of 115 network features are included in the dataset, which gives a reliable normal and malicious network flow. To assure normal network behavior for the training dataset, the dataset contains independent benign network traffic for each commercial device. The author[16]organized a battle to create a tagged dataset for intrusion detection. The collection covers four days of warfare competition network traffic. They went through the benefits and drawbacks of employing wartime competition to generate a current labeled dataset for intrusion detection in great detail. The collection includes both typical user behaviors and numerous forms of assaults. The data is open to the public. In this work, we have considered a dataset as an IoTID20 dataset [1], which is a combination of IoT devices and interconnecting structures. A typical smart home environment was implemented, which consists of an EZVIZ Wi-Fi camera and smart home device SKTNGU. These two Internet-of-Things devices were linked to a smart home Wi-Fi router. Laptops, tablets, and smartphones are among the other devices linked to the smart home router.

TABLE I
CORRELATED FEATURES OF BINARY LABEL IN IOTID20 DATASET

S. No.	Attributes			
	Binary	Binary Level Distribution	Extract Features	Features Name
1	Normal	40073	12	Active_Max, Idle Max Bwd_IAT_Max, Subflow Bckwd Byts, Bckwd Seg Size Avg, Fwd_IAT_Max, Pkt Size Avg, Subflow Bwd Pkts, Subflow Fwd Byts, Fwd Seg Size Avg, Subflow Fwd Pkts, PSH Flag Cnt
2	Anomaly	585710		

All the other devices in the testbed are attacking devices, whereas the SKT NGU and EZVIZ Wi-Fi cameras are IoT victim devices [17]. There are 80 network characteristics and three label features in the IoTID20 dataset. Binary labels of the IoTID20 dataset are shown in Table I.

The IoTID20 dataset's main advantages are that it mimics a current trend in IoT network traffic and that it is one of the few freely accessible IoT intrusion detection datasets. The IoTID20 dataset has 83 network features and three label features in its final form. The IoTID20 dataset was evaluated by features correlation and various machine learning Classification algorithms.

Due to the exponential growth of IoT devices in our daily lives and in every field. IoT devices are using heterogeneous communication channels and open shared communication mediums, so security has become a prime concern. To deploy security mechanisms, the most important phase is identification of anomalies and intrusion. IoT devices have heterogeneous capacities, low latency, and enormous volumes, so intrusion detection in IoT is necessary.

III. Proposed Methodology

The work flow of the proposed methodology of IDS for IoT home databases using ML is illustrated by Fig. 1. Initially, Data pre-processing is done using one-hot encoding once data is acquired from the IoTID20 dataset. The tagged data is usually not in a machine-readable format, thus it must be transformed into a numerical format. As a result, the unstructured data will be converted into structured data [18].

eISSN 1303-5150

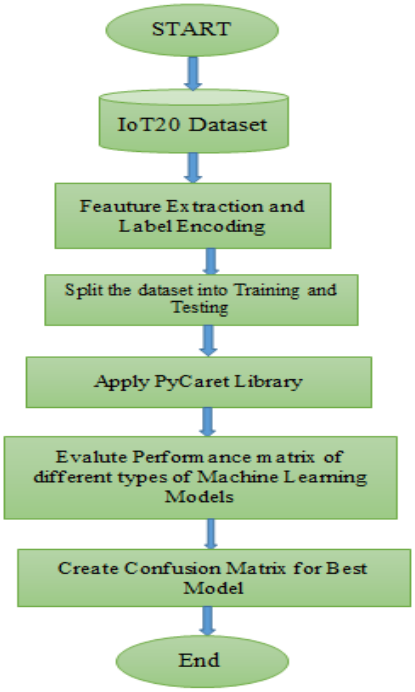

Fig. 1 Flow Chart of Proposed Methodology

As a result, a One-Hot Encoding technique is used to overcome the ordering concerns. The category value that is changed to a column with a value of 0 or 1 is addressed by the One hot encoding. The row values that show true and value as '1' will be in the first column value. The other columns' values indicate false, which is represented by the number '0'. If the rows and column values match, the whole value is assigned a value between 0 and 1. As a result, pre-processing one hot encoding approach is used to convert labels into numeric values. Afterwards the PyCaret library is a low-code machine learning framework that may be used for regression and prediction. It's an open-source machine learning library that was released to the public in 2020. The 12 ML methods were chosen automatically after an exhaustive search of ML algorithms using the IoTID20 dataset. The study has several limitations size of data, models were trained using 10-fold cross-validation, which used all of the samples as both training and testing. This method analyses five classification accuracy, Area under the ROC Curve (AUC), Recall, Precision, and F1 Score to assess the performance of the classifier. Classification accuracy is the likelihood that data will be accurately categorized, whereas sensitivity is the likelihood that attacking data will be correctly labeled as negative. Specificity is the likelihood that normal data will be correctly classified as positive. The formula below can be used to describe these indicators:

The number of true positives over the sum of the true positives and the false positives is known as precision (PRE).

$$\mathcal{PRE} = \frac{\mathcal{TP}}{\mathcal{TP} + \mathcal{FP}} \tag{1}$$

The number of true positives over the number of true positives and addition of the number of false negatives is how recall is calculated.

$$\mathcal{REC}\ \frac{\mathcal{TP}}{\mathcal{TP} + \mathcal{FN}} \tag{2}$$

A measured value's accuracy is how closely it resembles a reference or predetermined value. The accuracy is denoted as $\mathcal{ACC}$.

$$\mathcal{ACC} = \frac{\mathcal{TP} + \mathcal{TN}}{(\mathcal{TP} + \mathcal{TN} + \mathcal{FN} + \mathcal{FP})} \tag{3}$$

The "Cohen's kappa is a measurement of the level of agreement between two raters who decide which category a limited number of subjects fall within. There are no levels of disagreement between the two raters *(i.e.*, no weightings); their ratings are either in agreement (*i.e.*, the category to which a topic is assigned) or they are in disagreement."

$$Pr(\mathcal{ACC}) = \frac{\mathcal{TP} + \mathcal{TN}}{(\mathcal{TP} + \mathcal{TN} + \mathcal{FN} + \mathcal{FP})} \tag{4}$$

$$Pr(e) = \frac{(\mathcal{FP} + \mathcal{TN}) * (\mathcal{TP} + \mathcal{TN}}{(\mathcal{FP} + \mathcal{FN}) * (\mathcal{TP} + \mathcal{FN})} \tag{5}$$

$$K = \frac{Pr(\mathcal{ACC}) - Pr(e) + \mathcal{TN}}{1 - Pr(e)} \tag{6}$$

The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers. Suppose that classifier A has a higher recall, and classifier B has higher precision."

$$F1 - Score = 2 * ((precision * recall)/(precision + recall)). \tag{7}$$

Confusion matrices are used to display crucial predictive parameters including recall, F1 Score, AUC, accuracy, and precision. Confusion matrices are useful because they offer

eISSN 1303-5150

direct comparisons of variables like True Positives, False Positives, True Negatives, and False Negatives. So here we figure out the confusion matrix and the ROC curve for the best model *i.e.* XGBoost Model.

PyCaret [19] is extremely simple to install and just takes a few minutes in a virtual environment by using the command 'pip install PyCaret'. The 'setup' is the first and most important step in using the PyCaret package. This step takes care of all of the data preparation that is required before creating models. It saves all of the information types of information highlights after presenting the setup. To proceed, simply hit 'Enter' if the information types highlighted are correct. PyCaret uses a 70:30 split ratio for the train-test split, which may be changed by supplying a boundary 'train size' when configuring the models. It may be fascinating to see the model show a different split proportion. Because the highlights of the dataset are scaled differently, there is a clear need to standardize the dataset to provide a better result while creating and testing the model. Allow us to set 'normalize=True' for PyCaret to perform the normalization. A shape is a tool for interpreting the results of machine learning models.

IV. Result and Discussion

In this section, we have analysed the intrusion detection approaches with different models in terms of accuracy, etc. After training the dataset, performance is measured for binary-level distribution. The simulation ran on the HP Notebook with a 1.80GHz CPU with 4-cores and 8GB of RAM and PyCaret package [10]. The results are compared and analysed in terms of performance metrics. The accuracy of SVM is lower than other models, and the highest accuracy is provided by the XGBoost model in Fig.2. Similarly, it is also observed that the performance of the SVM model in terms of recall, precision, and F1score is also comparatively less than that of other models. Fig. 2 shows the optimal confusion matrix of XGBoost for each chunk of IDS in home IoT.

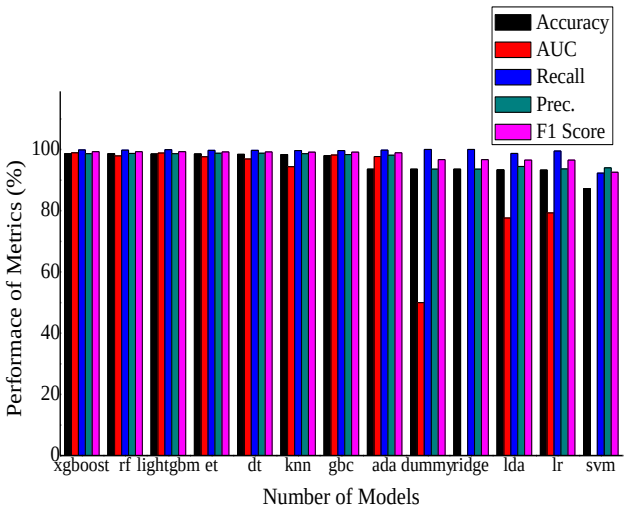

Fig. 2 Comparative analysis of Binary Label Distribution for IDS in IoT

Fig. 3 shows the overall confusion matrix after averaging the results of all chunks of of IDS in home IoT dataset.

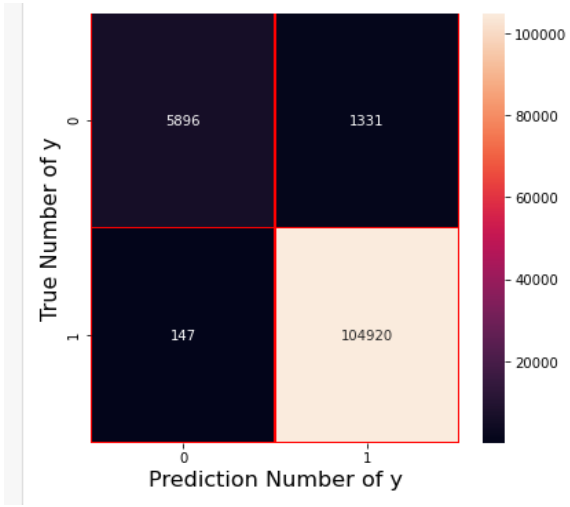

Fig. 3 The optimal confusion matrix of XGBoost.

In Fig. 4, receiver operating characteristic curve (ROC) of XGBoost is presented. The orange dotted line shows that true positive rate is increasing exponentially, however at 0.2 it is constant upto 1.0 false positive rate.

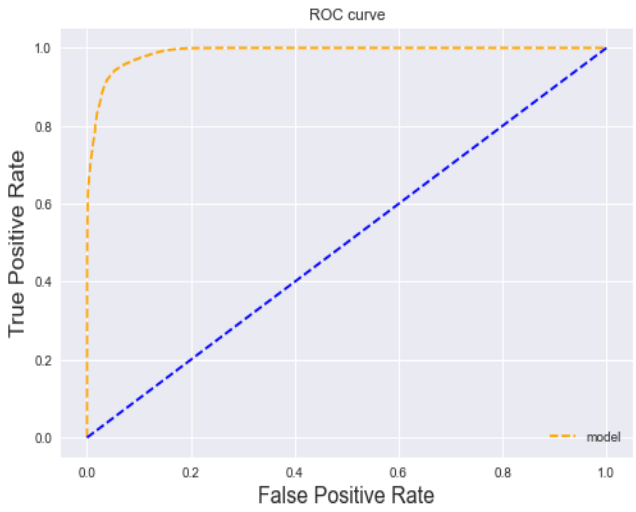

Fig. 4 ROC curve of XGBoost

V. CONCLUSION

This paper addressed the feasibility of using machine learning approaches to detect intrusions in the IoT home dataset

(IoTID20). The IoTID20 collection includes a wide variety of IoT threat kinds and families. We also analyzed the shortcomings of a few additional intrusion detection datasets. This dataset is used to evaluate intrusion detection strategies in IoT networks. We also provided a list of strongly related and relevant features for feature importance and ranking. The 12 most well-liked machine learning algorithms were employed to analyse the data. The outcome of the comparative analysis showed that the XGBoost ML approach provided better accuracy compared to other ML approaches.

REFERENCES

[1] I. Ullah and Q. H. Mahmoud, "A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks," in Advances in Artificial Intelligence, pp. 508–520, 2020.
[2] S. Li, L. Da Xu, and S. Zhao, "The internet of things: a survey," Information Systems Frontiers, vol. 17, no. 2, pp. 243–259, 2015.
[3] U. Pandey, V. K. Srivastav, B. K. Chaurasia and Neelu, "IoT conceptual Model and Application," Internet of Things Frameworks for Enabling and Emerging Technologies, Ch 1, pp. 1-22, CRC Press, Taylor & Francis Group, USA, April 29, 2022.
[4] M. B. Mohamad Noor and W. H. Hassan, "Current research on Internet of Things (IoT) security: A survey," Computer Networks, vol. 148, pp. 283–294, 2019.
[5] A. Chaturvedi, D. N. Goswami, R. Soni and B. K. Chaurasia, "Secure Multi-party Communication in Data-mining Applications," International Journal of Database Theory and Application, vol. 8, no. no.4, pp. 299–306, 2015.
[6] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The cluster between internet of things and social networks: Review and research challenges," IEEE Internet of Things Journal, vol. 1, no. 3, pp. 206–215, 2014.
[7] R.P. Lippmann, D.J. Fried, I. Graf, J.W. Haines, K.R. Kendall, D. McClung, D. Weber, S.E. Webster, D. Wyschogrod, R.K. Cunningham, and M.A. Zissman, "Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation," in Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00, vol. 2, pp. 12–26, 2000.
[8] J. W. Haines, L. M. Rossey, R. P. Lippmann, and R. K. Cunningham, "Extending the DARPA off-line intrusion detection evaluations," In Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01, vol. 1, pp. 35–45, 2001.
[9] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," Computers & Security, vol. 31, no. 3, pp. 357–374, 2012.
[10] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009, pp. 1–6, doi: 10.1109/CISDA.2009.5356528.
[11] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection," IEEE Access, vol. 8, pp. 132911–132921, 2020.
[12] M. Almshari, A. O. Khadidos, G. Tsaramirsis, F. Q. Khan, A. Khadidos, I. K. Giannopoulos, A. K. Leros, D. Piromalis, and M. Bukya, "Setting up Local Private Smart Grids with Data Collection Sensors for Scientific Experiments using MQTT and Node-Red," in Intelligent Computing Techniques for Smart Energy Systems, vol. 862, pp. 743–755, 2022.
[13] A. Sperotto, R. Sadre, F. van Vliet, and A. Pras, "A Labeled Data Set for Flow-Based Intrusion Detection," In IP Operations and Management, pp. 39–50, 2009.
[14] M. Shurman, R. Khrais, and A. Yateem, "DoS and DDoS attack detection using deep learning and IDS," International Arab Journal of Information Technology, vol. 17, no. 4A Special Issue, pp. 655–661, 2020.
[15] Y Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," IEEE Pervasive Computing, vol. 17, no. 3, pp. 12–22, 2018.
[16] B. Sangster, T. J. O'Connor, T. Cook, R. Fanelli,E. Dean, W. J. Adams, C. Morrell, and G. Conti, "Toward instrumenting network warfare competitions to generate labeled datasets," 2nd Workshop on Cyber Security Experimentation and Test, CSET 2009, 2009.
[17] M. G. Karthik and M. B. M. Krishnan, "Detecting Internet of Things Attacks Using Post Pruning Decision Tree-Synthetic Minority Over Sampling Technique," International Journal of Intelligent Engineering and Systems, vol. 14, no. 4, pp. 105–114, 2021.
[18] M. Adhikari, A. Hazra, V. G Menon, B. K. Chaurasia, and S. Mumtaz, "A Roadmap of Next-Generation Wireless Technology for 6G-enabled Vehicular Networks," IEEE Internet of Things Magazine, Vol. 4, No. 4, pp. 79-85, 2021.
[19] F. Bin Iqbal, S. Biswas, and R. Urba, "Performance Analysis of Intrusion Detection Systems Using the PyCaret Machine Learning Library on the UNSW-NB15 Dataset," In journal of Big Data, Article no. 105, 2020.