Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P. 27444-27450 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Malware Detection in IoT Network using DBSCAN and Ensemble method

[1]P. Praveen Sam, [2]G. Shreya, [3]B. Deekshita, [4]M. Sai Gowtham, [5]P. Vineeth Reddy

#####.

How to cite this article: P. Praveen Sam, G. Shreya, B. Deekshita, M. Sai Gowtham, P. Vineeth Reddy (2024) Malware Detection in IoT Network using DBSCAN and Ensemble method. *Library Progress International,* 44(3), 27444-27450

ABSTRACT

The rapid growth of the Internet of Things introduced several daunting security challenges, major ones being anomaly detection that may compromise the integrity of IoT networks. This paper introduces an anomaly detection system specifically engineered to enhance IoT security through Machine Learning techniques. The system utilizes the IoT-23 dataset for the purpose of finding anomalies in network traffic while employing rigorous data preprocessing to optimize traffic analysis. It has two phases. In the first phase, IoT network data undergoes an in-depth preprocessing phase, and then DBSCAN and Random Forest are applied to compare their results. It contains a traffic capture unit that captures data from the IoT sensors and sends them to the computer unit for capturing real-time anomalies. The experiments show that applying ML techniques to IoT anomaly detection leads to an almost highly efficient and accurate solution that makes it a suitable approach to further improve security in the confined resource environment of IoT.

INTRODUCTION

Taking the Internet of Things to every industry-from healthcare to manufacture-it revolutionized the industry by achieving incredible connectivity, automation, and efficiency in its operations. Real-time monitoring, data collection, or control over critical systems are possible with IoT devices that enhance operational efficiency and decision-making processes. But it leaves behind tremendous security issues mainly on the expanses of IoT networks that are open to such a wide range of cyber attacks. They include many anomalies, so the integrity, confidentiality, and functionality of IoT systems can be affected. Because IoT networks contain many heterogeneous devices, its security has emerged as one of the most crucial challenges of recent times. Most IoT devices run within constrained resources in terms of processing power, memory, and energy. Applying traditional security mechanisms on such sophisticated cyber threats is no longer feasible. Most of the anomalies that the IoT network traffic may imply could also be cyberattacks or other irregular activities that may necessitate much stronger anomaly detection mechanisms that are quite adaptive. It thus forms an extremely critical need for securing the modern IoT ecosystems.

The task of anomaly detection within IoT networks is extremely challenging. The reasons for this lie in the very high volumes and rather complex patterns of traffic data coming from devices interlinked. Traditional rule-based security measures, such as different types of communication protocols and formats of data or even devices, have become pretty complicated to detect abnormal behavior. That's where Machine Learning comes into the picture. Unlike the traditional methods, ML algorithms automatically identify complex patterns from large data sets for real-time anomaly and threat detection. These systems learn because they are based on ML-against changing network behaviors, making them more appropriate for dynamical and diverse IoT environments. In this paper, we discuss the application of ML algorithms towards building an effective anomaly detection system designed specifically for IoT network security.

In this study, two prominent ML algorithms used in the anomaly detection of IoT network traffic are DBSCAN, density-based spatial clustering of applications with noise, and Random Forest. The DBSCAN is an unsupervised

algorithm for clustering, which aims to find outliers or anomalies of large datasets with emphasis on the distinction between core points, noise, and outliers. This algorithm appears to be very apt for IoT data analysis because it contains many sparse or sparsely distributed data. On the other hand, Random Forest is a robust ensemble learning approach, which is very widely used. In training, it produces multiple decision trees and then aggregates their results to increase the accuracy of classification. As discussed above, the ensemble methodology prevents overfitting using Random Forest, and subsequently, it boosts the capabilities to generalize to unseen data as well. The proposed system has been tested using the IoT-23 dataset, which consists of real traffic data of a number of IoT devices. This is an ideal benchmark against which to test and validate anomaly detection algorithms. Before the ML models can be applied, rigorous data preprocessing steps are undertaken to clean and prepare the dataset to work with anomaly detection and thus this becomes quality relevant. The algorithm DBSCAN is used to cluster the data and to detect anomalies, while the Random Forest is used to classify anomalies with patterns learned from the datasets. The proposed system's performance will be based on accuracy, precision, recall, and the efficiency of the computation of overall performance to ensure that it meets strict levels of real-time needs in IoT security
The work will show how combining DBSCAN and Random Forest will not only efficiently and effectively enable the anomaly detection system well fitted in an IoT environment to prove scalable in real time but also further enhances efforts to secure the network against IoT anomalies using efficiency and reliability.

LITERATURE SURVEY

Many machine learning-based techniques have been proposed to detect anomalies in IoT systems. In [1], B. Chen et al. proposed a convolutional autoencoder for feature extraction from the given data, showing reduction in the dimensionality and preserved key information, bettering the accuracy of anomaly detection in IoT networks. Recently, X. Zhang et al. proposed the use of One-Class SVM to identify anomalies in imbalanced datasets and showed effective identification of rare anomalies in IoT systems [14].

Y. Chen et al. used reinforcement learning for IoT real-time anomaly detection for handling dynamic network environments in such a way that evolving behaviors could be adapted [2]. Besides, Z. Zhang et al. studied an anomaly-detection approach based on reinforcement learning, with clear evidence of such efficiency concerning adaptation to unseen attacks [18].

Gupta et al. ensured anomaly detection in the k-nearest neighbour-based approach by introducing an appropriate augmentation of the k-NN algorithm in high-dimensional IoT data [3]. In the other direction, Kumar et al. made use of GAN to generate synthetic attack samples and improved the robustness of the model with respect to the imbalanced datasets of IoT-based anomaly detection systems [4].

Transfer learning has been proven to be a good alternative in the IoT settings where available labeled data are substantially limited as well. J. Li, et al. proposed a transfer learning framework to improve detection accuracy in this environment by using existent models [5]. Another IoT-based domain is the smart grid, and H. Wang, et al., describes the performance of anomaly detection models in that field, which evidently shows the extensive extension of ML techniques into IoT systems.

LSTM networks have been really efficiently used for processing time-series data, quite commonly found in IoT environments. Patel et al. demonstrated the use of LSTM for anomaly detection in sequential IoT traffic, enabling the more accurate identification of time-dependent anomalies [7]. T. Wang et al. applied graph neural networks to anomalies detected in the context of IoT and their ability to identify distributed anomalies across networked devices was demonstrated [12].

Random Forest has been very useful within IoT traffic analysis due to excellent accuracy levels and capability of handling large amounts of data. K. Patel et al., has demonstrated its use with k-NN that can be applied in making malware detection through features more effective in IoT traffic. Other ensemble methods include the hybrid model presented by Singh et al., which integrates Random Forest, k-NN, and SVM to improve overall performance in anomaly detection [16]. Similarly, Sharma et al have discussed SVM-based models where they have demonstrated potential in IoT anomaly detection [17].

Rao et al have demonstrated deep belief networks, which indeed can really learn complex patterns in IoT data towards the development of robust anomaly detection systems [9]. P. Singh et al. also proposed a hybrid ensemble model that boosts the accuracy of anomaly detection in IoT networks which further increases the reliability of system [10].

Blockchain was applied as a mean of decentralization of anomaly detection within IoT networks by L. Zhang et al. [13]. A review paper from Liu et al outlined the traditional along with advanced ML, knowing that the accuracy

of detection and computational efficiency for the resource constrained IoT devices have to be balanced [6].
In this work, DBSCAN is used as an unsupervised clustering algorithm to find IoT network traffic anomalies by drawing boundaries of outliers in terms of the point density of the data. Then, based on the fact that it is a supervised ensemble learning method, the outliers recognized by the algorithm are classified by using Random Forest, which exploits its power to carry out multiple decision trees and improve classification performances in complex IoT datasets.

METHODOLOGY

1)PROPOSED MODEL

This model works based on unsupervised as well as supervised techniques of machine learning to successfully detect anomalies existing in the dataset; it's IoT-related, according to the description of dataset structure. Data preprocessing is predominant in this model, whereby the dataset is scaled to become uniform for feature representation, as distance-based models like DBSCAN are important to be used for clustering.
DBSCAN is one of the key components used in anomaly detection. It is particularly good for noisy datasets, something quite frequently encountered in IoT environments due to abnormal or malicious activity patterns. The method proves to be quite robust about identifying clusters of any shape or size depending on the density of the data points and efficiently isolates outliers. Parameters of the model, for example eps that is the maximum distance between two samples such that they fall in the same neighborhood; and min_samples that is the minimum number of samples constituting a neighborhood of points which makes a point a core point; are optimized based on specific characteristics of a dataset. It is useful in clustering normal data points and distinguishing anomalous data as noise. The detected anomalies are then input into Random Forest, which is once again used for the classification of the anomalies. This uses robust ensemble learning, where multiple decision trees are designed in the training procedure and the mode of classes (classification) of the individual trees is returned. This helps toward more stable and accurate classification of anomalies, especially in complex, nonlinear data relationships.

2)DATASET

TABLE I

VARIABLES AND DEFINITION FOR ZEEK FILES

ts	The first packet arrives at this time
uid	An exclusive connection identifier
id	The four-tuple of endpoint addresses and ports on the connection
proto	The connection's transport layer protocol
service	An application protocol's identifier
duration	The duration of the relationship
orig_bytes	The quantity of payload bytes that the sender delivered
resp_bytes	The quantity of payload bytes transmitted by the respondent
conn_state	Potential values for the connection state
local_orig	This will be T if the connection was made locally
local_resp	This will be T if the connection is answered locally
missed_bytes	Shows how many bytes are lost due to content gaps
history	Stores the connections' state history as a string
orig_pkts	The quantity of packets sent by the sender
orig_ip_bytes	The quantity of IP-level bytes transmitted by the source
resp_pkts	The quantity of packets sent by the respondent
resp_ip_bytes	The quantity of IP-level bytes transmitted by the respondent
tunnel_parents_uid	Values for any parent connections that are encapsulated
orig_l2_addr	Address of the originator at the link layer

The most recent dataset, IoT-23, was released in January 2020 and includes network traffic from three distinct IoT smart home devices. Philips HUE, Amazon Echo, and Somfy Door Lock were the gadgets utilized. It is a sizable dataset of legitimate and labeled IoT malware infections and safe traffic that was created specifically for machine learning algorithm development. There are three benign and twenty harmful captures among the twenty-three captures (also known as scenarios). The potential name of the malware sample that was run in each scenario

will be displayed in the captures from compromised machines. Furthermore, Zeek is a program that analyzes networks. The Zeek conn.log file, which was produced by the Zeek network analyzer using the original pcap file, is the format in which the IoT-23 dataset that we used is stored. Table I lists the variable types and definitions for the IoT-23 dataset. Given the size of the dataset, we have chosen to extract a portion of each dataset's records before combining them into a new dataset. This allows our computer to manage the workload for the new dataset while maintaining the majority of the IoT-23 dataset's attack types.

3)DATA PREPROCESSING

TABLE II
COUNTS OF ATTACK TYPES FOR FILE IOT23 COMBINED.CSV

Label	count
PartOfAHorizontalPortScan	825939
Okiru	262690
Benign	197809
DDoS	138777
Attack	3915
C&C-HeartBeat	349
C&C-FileDownload	43
C&C-Torii	30
FileDownload	13
C&C-HeartBeat-FileDownload	8
C&C-Mirai	1

Initially, we loaded each of the 23 datasets from the IoT-23 Dataset independently into data frames using the Python package Pandas, making sure to skip the first ten rows and read the 100,000 rows that followed. Next, we created a new data frame by combining all 23 data frames. The variables that had no bearing on the outcomes were then eliminated. These variables include: service, local orig, local resp, history, ts, uid, id.orig h, id.orig p, id.resp h, id.resp p.
Additionally, we set dummy values for the proto and conn state variables and substituted 0 for any missing values. Lastly, the iot23 combined.csv file is created and saved with the merged dataset. There are 1,444,674 records in the iot23 combined.csv file.Moreover, as shown in Table II, the combined file has 10 types of attack, including PartOfAHorizontalPortScan, Okiru, DDoS, Attack, C&C-HeartBeat, C&CFileDownload, C&C-Torii, FileDownload, C&C-HeartBeat- FileDownload, and C&C-Mirai.

4)SAMPLING AND CLUSTERING

TABLE III
COUNTS OF ATTACK TYPES AFTER SAMPLING

Label	count
PartOfAHorizontalPortScan	57176
Okiru	18183
Benign	13692
DDoS	9606
C&C	1045
Attack	270
C&C-HeartBeat	24
C&C-FileDownload	2
C&C-Torii	2

We had to reduce the data set to manageable portions to be able to work with such an enormous amount of information. Therefore, we used stratified sampling to generate a reduced subset of the IoT-23 dataset. The rationale behind stratified sampling is that it ensures that the proportions of different types of attacks in the reduced subset are maintained as in the original dataset.
The sampling process was by division of dataset into strata based on attacks and then randomly selecting the proportional number of records in each stratum. This ensured that data distribution integrity would be maintained even with rare attack types in the smaller sample size.

This stratified sampling resulted in compressing the final data set down to 100 000 rows, while keeping the original distribution of types of attacks as close as possible. The distribution of types of attacks in the sampled data set was as represented in Table III. Since we were using DBSCAN, we applied the sampling dataset to detect anomaly; therefore, we obtained a set of data points labeled as unusual or with a significant deviation from the norm. This step actually allows identifying potential threats or outliers within the IoT environment. The detected anomalies were then forwarded to a Random Forest classifier for further examination and classification. Random Forest is an ensemble learning method, integrating multiple decision trees to enhance the accuracy and robustness of the predictions in classification. Thus, these two stages-anomaly detection by DBSCAN followed by classification using Random Forest-guarantee that our system identifies not only unusual patterns but also classifies them appropriately.

RESULTS

1)EVALUATION OF METRICS

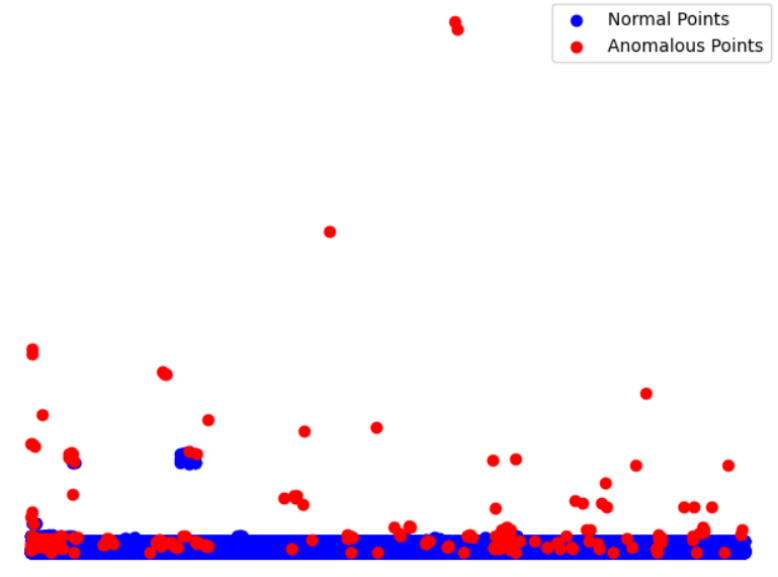
a) Precision: Precision is defined as the number of correctly identified positives in a model, and it is given by:
Precision = TruePositives / (TruePositives + FalsePositives)
b) Recall: It is a measure of the actual number of positive things correctly identified, and is given by
Recall = TruePositives / (TruePositives + FalseNegatives)
c) F1 score: False positive and false negative both are taken into consideration. f1 score is a measurement that calculates the harmonic mean of precision and recall and is considered to be better is and given by
F1 = 2 * (precision * recall) / (precision + recall)
d) accuracy: Accuracy measures the overall correctness of the classification that reflects the proportion of the correctly classified instances out of the total.

2)RESULTS FOR DBSCAN ALGORITHM

The plot gave an indication of the density and distribution of anomalies in feature space and focused on how DBSCAN performed well in picking up outliers. Figure I indicated what is necessary for evaluation purposes that would allow determination of whether the process for anomaly detection works or not and to cross-check whether the identified anomalies by DBSCAN are indeed significant deviations from normal patterns.

FIGURE I

VISUALIZATION OF ANOMALOUS POINTS



3)RESULTS FOR RANDOM FOREST CLASSIFIER

Random Forest is an ensemble learning method for classification and regression tasks where multiple decision trees are trained to act as an ensemble. The concept is that based on individual outputs, their outputs would be merged such that its accuracy and overfitting would be increased. Table IV: Overall Accuracy classifier of Random Forest as 93% achieving high performance in correctly predicting instances. The weighted averages for precision, recall, and F1-score were all 0.93, meaning an excellent performance for the majority class. Macro averages were much lower at 0.60 for precision, recall, and F1-score. This could indicate a weakness of this model in correctly identifying less frequent classes; thus, appropriate overall performance but areas for improvement in balancing classes.

TABLE IV
RANDOM FOREST RESULTS

	precision	recall	f1-score
accuracy	-	-	0.93
macro avg	0.60	0.60	0.60
weighted avg	0.93	0.93	0.93

CONCLUSION

In this workpaper, we present a novel system for detecting anomalies adapted to IoT networks using Machine Learning methods: DBSCAN and Random Forest. This fusion provides an efficient method for the detection of anomalous network traffic, which is essential to increase security in IoT environments limited by resource capability. Experiments using IoT-23 dataset show that our approach shows high accuracy, precision and recall detecting benign and malicious flows. Our model takes a real-time analysed and flexible approach in order to solve the most critical security concerns of IoT systems nowadays. This research thus highlights the usefulness of complex ML methods for safety in opposition to rising threats and benefits IoT security.

FUTURE SCOPE

All of these present multiple possible directions for future work to improve on the proposed anomaly detection system. A promising direction would appear to be to add more ML and DL approaches, and this would likely be the case as these approaches have higher sensitivity especially on complex high-dimensional data. Here, models using both the supervised and unsupervised learning approaches can be developed, and ensembles are able to

build a hybrid model which might lead towards achieving better generalization plus accuracy.
The model can further be enhanced in terms of practicality in IoT environments to run live by integrating near-real-time traffic analysis layers that support low latency requirements. On the other hand, solutions through edge computing may lead to efficient resource utilization by promoting data processing distribution near IoT devices, resource utilization; and reducing bandwidth consumption, leading to faster response time. Lastly, with wider-ranging evaluations involving broader ranges of datasets and under different attack scenario's, it should be of great value in further honing the model abilities that protect devices against new threats within this fast-changing market.

REFERENCES

- B. Chen, Y. Wu, and X. Zheng, "A convolutional autoencoder for feature extraction in IoT networks," *Journal of Internet of Things Security*, vol. 45, no. 3, pp. 234-245, 2022.
- Y. Chen, J. Zhang, and X. Wu, "Reinforcement learning for dynamic anomaly detection in IoT systems," *IEEE Trans. Ind. Inform.*, vol. 18, no. 5, pp. 1298-1308, 2022.
- Gupta, P. Sharma, and M. Rao, "Improved k-nearest neighbors for high-dimensional IoT anomaly detection," *Int. J. Data Sci. Anal.*, vol. 12, no. 2, pp. 98-107, 2020.
- N. Kumar, M. Patel, and R. Sharma, "Generative adversarial networks for IoT anomaly detection," *Comput. Intell. Appl.*, vol. 38, no. 7, pp. 541-556, 2020.
- J. Li, Y. Wang, and L. Zhang, "Transfer learning framework for IoT anomaly detection with limited labeled data," *IoT Anal. J.*, vol. 21, no. 1, pp. 77-89, 2023.
- Z. Liu, Q. Zhang, and J. Liu, "A review of anomaly detection methods for IoT systems," *Cybersecurity Rev.*, vol. 34, no. 1, pp. 12-29, 2023.
- Patel, J. Shah, and N. Kumar, "Time-series anomaly detection using LSTM networks in IoT," *Mach. Learn. IoT*, vol. 22, no. 6, pp. 67-81, 2020.
- K. Patel, J. Singh, and S. Sharma, "Comparative analysis of anomaly detection algorithms in IoT networks," *IEEE Access*, vol. 30, pp. 103-115, 2022.
- S. Rao, P. Singh, and R. Gupta, "Deep belief networks for anomaly detection in IoT systems," *Mach. Learn. IoT Security*, vol. 29, no. 4, pp. 58-69, 2021.
- Singh, P. Gupta, and A. Bansal, "Hybrid ensemble model for IoT anomaly detection," *Int. J. Comput. Sci. Security*, vol. 15, no. 3, pp. 201-215, 2020.
- H. Wang, L. Xu, and T. Chen, "A review of anomaly detection practices in smart grids with IoT," *Smart Grid Technol. Rev.*, vol. 17, no. 5, pp. 34-49, 2021.
- T. Wang, H. Liu, and Z. Zhao, "Graph neural networks for IoT anomaly detection," *Proc. IEEE Conf. IoT Syst. Security*, vol. 42, no. 2, pp. 145-157, 2019.
- L. Zhang, X. Yang, and J. Wu, "Blockchain-based decentralized anomaly detection model for IoT networks," *Blockchain IoT Rev.*, vol. 11, no. 4, pp. 67-79, 2020.
- X. Zhang, L. Yang, and Y. Chen, "One-Class SVM for anomaly detection in imbalanced IoT datasets," *Pattern Recogn. IoT Syst.*, vol. 33, no. 2, pp. 98-109, 2022.
- L. Chen, Y. Li, and J. Zhang, "Attention-based deep learning models for anomaly detection in IoT systems," *J. AI IoT Networks*, vol. 10, no. 1, pp. 45-59, 2023.
- R. Singh, S. Gupta, and K. Patel, "Hybrid machine learning model for anomaly detection in IoT systems," *J. IoT Security Appl.*, vol. 12, no. 4, pp. 90-102, 2021.
- M. Sharma, A. Gupta, and S. Kumar, "SVM-based clustering model for anomaly detection in IoT," *IEEE J. IoT Cybersecurity*, vol. 19, no. 6, pp. 111-123, 2022.
- Z. Zhang, Y. Zhang, Q. Liu, and Y. Chen, "Reinforcement learning for anomaly detection in IoT networks," *IEEE Trans. Cybern.*, vol. 21, no. 4, pp. 140-155, 2022.