DETECTION OF MALICIOUS ATTACKS IN IOT NETWORK TRAFFIC USING MACHINE LEARNING TECHNIQUES

Kalaiselvi S[1], Hariharasudhan R[2], Divakar A[3], Mytheeswaran V[4], Niveditha TP[5]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

The linking of abnormal and malicious business in the network is crucial for IoT safety to monitor and restrict un-authorised IoT network traffic flows. Many machine literacy (ML) approach models have been provided over by various experimenters to aid in preventing malicious business movements in the IoT network. Due to the sloppy point selection, several ML models are vulnerable to misclassifying potentially harmful business flows. However, additional research has to be done to fully understand how to pick appropriate features for the accurate identification of malicious enterprises in IoT networks. To address the issue, a new frame model is proposed. First, we propose a new point selection measure entitled CorrAUC, and then, based on CorrAUC, we create and design a new point selection method called Corrauc. Corrauc uses a wrapper-style technique to evaluate the features directly and select the most effective features for the specified ML algorithm. In addition, we used a combined TOPSIS and bijective soft set to verify the names of features used to identify malicious enterprises in the IoT system. We evaluate the effectiveness of our suggested method using the Bot- IoT dataset and four distinct ML methods.

The examination of experimental data confirmed that our projected system is efficient and can produce > 96 consequences on average.

[1]Assistant Professor, Department of MCA, Karpagam College of Engineering, Coimbatore, India.
[2,3,4,5]Scholar, Department of MCA, Karpagam College of Engineering, Coimbatore, India.

Email: [1]kalaiselvi.s@kce.ac.in, [2]hari69357@gmail.com, [3]divakarofcl@gmail.com, [4]divyamythees@gmail.com, [5]niveditha96tp@gmail.com

DOI: 10.31838/ecb/2023.12.s3.044

1. Introduction

In order to detect accurately in an IoT environment, a brand new dataset is employed for efficient feature selection. The data set documents a wide variety of cyberattacks, including those launched via botnets, common online behaviours, and the Internet of Things. This dataset, featuring efficient information features for tracking accurate traffic, was developed using a practical test platform. Other features were retrieved in a similar fashion and added to the set of extracted features to improve the effectiveness of machine learning models and forecasting models. However, the extracted features, like assault flow, groups, and subdivisions, are labelled for better performance outcomes.

The (IoT) is a rapidly developing knowledge that links millions of devices in a matter of seconds. Using this technology streamlines and simplifies regular tasks. The (IoT) is one example of a technology that was once exclusively used in households and small businesses but is now widely adopted because of the benefits it provides in terms of efficiency and reliability. Yet, Internet of Things technology is rapidly replacing traditional methods. With the advancement of IoT technology, more than 27 million connected devices are expected by 2021. Due to their rapid development and widespread usage, (IoT) devices are increasingly the target of cyberattacks. Several studies have found that botnet attacks are more common in IoT environments than other types of attacks. Because most IoT devices lack the memory and processing capacity required for good security systems, they still contain many security problems. In addition, many rule-based monitoring systems are easily circumvented by criminals. In this study,

Using a powerful feature assortment strategy, a lightweight, highly efficient recognition system is created. A botnet is a group of bots controlled by a single individual, the botmaster, via a command and control protocol to attack a specific network. Infected computers, known as bots, are utilised to perform malicious tasks under the botmaster's virtual control without leaving any traces of infection. A botnet's size can vary widely, from a modest network of a few hundred bots to a massive one of 50,000 hosts. Hackers can spread botnet software, remain active for years, and leave no traces of their presence or activity.

Machine learning algorithms will be crucial in IoT security systems because they are effective at modelling systems that cannot be given by mathematical equations. Machine learning is a subfield of computer science that studies how machines might acquire knowledge through observation and experience. In order to detect suspicious network activity, such as an attempt at infiltration, an anomaly detection system uses machine learning to create a fresh anomaly detection model.

Machine learning algorithms (MLAs) allow computers to learn from data examples and make predictions or take actions without being explicitly programmed to do so. There are three categories of MLAs based on the type of guidance they received throughout their training. There are four different kinds of learning: unsupervised, semi-guided, reinforced, and supervised. The term "guided learning" is commonly used to describe any type of instruction that makes use of a watchful eye. Learning and prediction are a part of it, as is data that has been labelled with desired results in order to facilitate the algorithm's transition from input to output.

KNN, examples of classification methods that fall under the category of guided learning. K-Nearest Neighbor algorithms use Euclidean distance to determine how closely an object is related to K neighbours of the same class. The value of K is low, which is desirable and common. The KNN method's precision is set by the number of peers used in the selection process (the value of K). To prevent having two sets of class names tie for the same count, K is often an odd number in binary classification. Setting K to 1 will have the object simply assigned to the nearest class.

If K is too little or too large, the model may not fit the data as well, hence it's important to find the optimal value. Included in the data set are DDoS attacks, DoS attacks, OS attacks, data exfiltration attacks, key logging attacks, and additional DDoS and DoS assaults mounted on the protocol in question. The BoT-IoT dataset is constructed in an actual IoT network using a number of technologies in order to replicate numerous botnet scenarios. The BoT-IoT dataset has been analysed exhaustively and sorted according to assault types. The IoT connects devices that can talk to each other and run largely autonomously with little help from humans.

IoT areas of computing, but the reality is that it is tremendously vulnerable to a wide variety of hacking techniques because to the highly aggressive nature of the modern internet. To keep IoT networks safe, we need to create real-world countermeasures like network anomaly detection. While it's impossible to totally stop assaults, being able to respond quickly to them is crucial for keeping your data safe.

By comparing and analysing several machine learning algorithms that can be used to quickly and accurately identify attacks against IoT networks, we hope to contribute to the current body of knowledge in this area. Many methods of identification are compared using the brand-new dataset Bot-IoT. There were seven different machine learning algorithms employed during implementation, and the most of them performed well. During

deployment, unique characteristics were collected from the Bot-IoT dataset that were superior to those found in earlier studies. As people around the world become more concerned about the security and privacy of their online accounts, computer security has gone from being a nice-to-have to an absolute necessity.

Recent hacking attempts can be traced back to the widespread use of web-based apps and the popularity of cutting-edge technologies like the Internet of Things (IoT). The Internet of Things (IoT) refers to a network of computing devices that may communicate and act independently. Thanks to the Internet of Things, many devices in the healthcare, agriculture, transportation, and other sectors are now web-enabled. Appliances, lighting fixtures, bicycles, and the like are all in this category. Internet of Things apps are changing our lives by helping us save both time and money. Furthermore, there is a limitless amount of room for the dissemination of knowledge and the introduction of new ideas.

The Internet is the backbone and brain of the Internet of Things. Nodes in the IoT have restricted capabilities, low bandwidth, and no user-accessible settings, in comparison to other types of networks. Furthermore, new network-based security approaches are required due to the substantial security problems raised by the proliferation and rapid expansion of IoT devices. Existing algorithms are good at identifying certain forms of assault but not others. There is undeniable need for more sophisticated techniques to enhancing network security in light of the exponential rise of both network threats and the volume of data stored in networks. As a result, we need faster and more accurate methods of identifying threats.

Machine learning (ML) is one of the most efficient computing approaches for bringing unified intelligence to the Internet of Things (IoT). Machine learning methods are increasingly being used for critical network security functions like traffic analysis, intrusion detection, and botnet detection. Machine learning is crucial to any Internet of Things solution because it enables smart devices to learn from their experiences and make decisions about their own actions based on that knowledge. ML algorithms are used in applications like regression and classification because they can draw meaningful conclusions from data provided by humans or machines. IoT network security is another area where AI may be applied.

While machine learning has many potential uses in cyber security, there is much debate over the best way to put it to use in spotting malicious intrusions. Few studies have focused on developing efficient detection methods appropriate for IoT environments, despite the widespread use of ML algorithms by other researchers to learn how to best uncover attacks. All experiments were run in Python using the language's native machine learning tools (scikit-learn, Matplotlib, Pandas, and NumPy). To determine which machine learning algorithms were best suited to this dataset, we used the following three criteria: The algorithms were evaluated in three stages: (1) on each attack separately from the dataset; (2) on the complete dataset using a set of features that aggregated the best features for each assault; and (3) on the full dataset using the results from the individual attacks. There are many ways in which large amounts of sensitive user data might be compromised, both by insiders and by outsiders. Cyberattacks have grown increasingly sophisticated alongside the development of computing and algorithm complexity. Cybercriminals frequently target computer networks that handle confidential data, maintain important records, or provide essential services.

Detecting hostile attacks that could compromise a network requires a state-of-the-art intrusion detection system (IDS). Automatic detection and classification of attacks, policy violations, and network intrusions are all possible with the help of intrusion detection systems (IDSs).

Existing System

In order to keep tabs on and prevent undesirable business flows in the (IoT) network, it is crucial that abnormality and malicious business be identified in the IoT network. In order to prevent malicious financial transfers across the IoT network, many researchers have presented numerous machine literacy (ML) style models. Yet, because to the unfortunate point selection, a number of ML models are predisposed to misclassify significantly malicious business flows. However, additional research is needed to fully understand the critical issue of how to select efficient features for precise malicious business discovery in an IoT network. A new frame model is presented as a solution to the issue. First, we propose a new point selection metric we call CorrAUC. Then, based on CorrAUC, we build and design a new point selection technique we call Corrauc. This algorithm is based on a wrapper style to access the features directly and select effective features for the aforementioned ML algorithm. Using the Bot- IoT dataset and four distinct ML methods, we provide an estimate of the efficacy of our suggested method. Analysis of experimental data confirmed that our suggested method is efficient, with an average accuracy of > 96.

Input Data: Here, we'll use a.csv file as an input dataset, sourced from a repository of such datasets as github or the University of California, Irvine.

Output Data: Accuracy, score are some of the measures of performance that are calculated. It's not really precise.

Disadvantages

- The results is low when associated with projected
- It doesn't efficient for large capacity of data's
- Theoretical restrictions.

Proposed System

Here, malicious business insights are uncovered with the help of the Bot- Iot dataset. A point-based approach, analogous to Correlation and chi squared, is offered for making selections. While building a predictive model, the point selection helps by limiting the amount of inputs. Using the Bot- IoT dataset and four distinct ML methods, we provide an estimate of the efficacy of our suggested method. the system was built using a machine-literacy method that is comparable to SVM and Naive Bayes (NB). Using dataset, we suggest a point selection style to improve the performance of ML techniques and address the issue of efficient point selection for cyber-attacks in the IoT network industry. Based on the study of experimental findings, we know that our suggested system is efficient and can produce average outcomes for each method that are lower than 98.

Input Data: Here, we'll use a.csv file as our input dataset, sourced from a repository such as github or UCI.

Output Data: Now, we may visualize the data to determine whether or not an attack has been detected in the network. Accuracy, precision, recall, and f1-score are some of the measures of performance that will be estimated.

Advantages

1) It is efficient for large sum of datasets.
2) The experimental result is high when compared with existing system.

1. **Naïve Bayes (NB)**

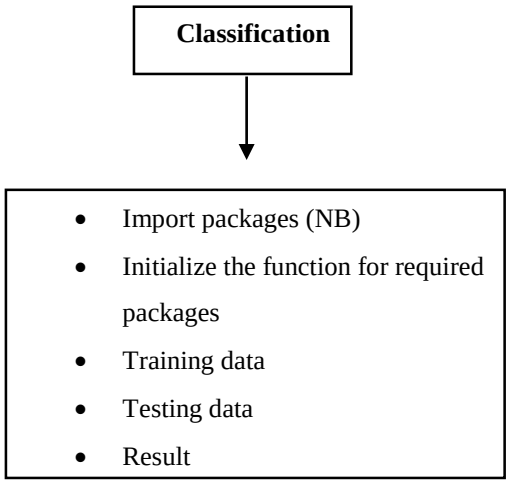


Fig 1.1

2. **Support Vector Machine (SVM)**

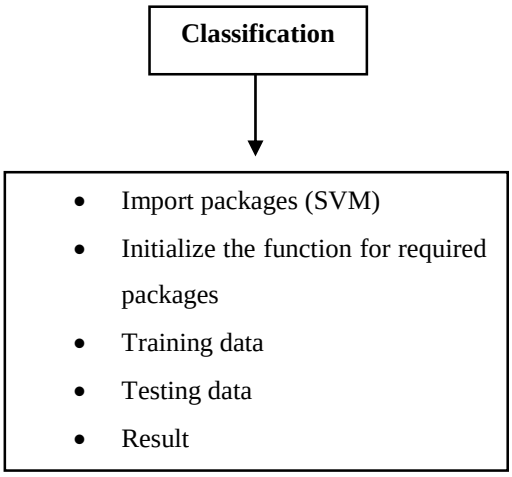


Fig 1.2

Modules

- Data selection
- Data pre-processing
- Feature selection
- Data splitting
- Classification
- Performance analysis

Data Selection

The data used as input was obtained from a data repository. We utilise the Bot- IoT dataset in our analysis. Selecting relevant data is the method used to identify dishonest enterprises. The data collection consists of several cyberattacks, business overflows, and botnet attacks, all of which have an online component. The effective information elements of this dataset were developed using a realistic test bench, allowing for reliable business tracking. More characteristics were uprooted and added with the uprooted features set to progress the performance of the machine literacy model and the efficacy of the vaticination model. nonetheless, the uprooting features are labelled like attack inflow, orders, and subclasses for superior performance outcomes.

Data Pre-Processing

Data pre-processing refers to the steps taken before a dataset is analysed. The dataset undergoes a metamorphosis during pre-processing to convert it into a format understandable by machines. At this stage, we also draw the dataset, removing any irrelevant or spoilt information that may impact the dataset's sensitivity. Lost data rubbish collection Garbling Information broken down into categories Dumping of incomplete data This procedure involves the replacement of missing or Nan values with 0. Data was cleansed of anomalies and purged of missing or duplicate values. Garbling The data is categorised into various groups. In this sense, a category is a type of data variable that can take on just a small number of possible values. The majority of algorithms used in machine learning require numerical input and affair variables.

Feature Selection

When building a predictive model, point selection is used to narrow down the inputs to a manageable set. Reducing the total number of features and instead focusing on only those features necessary to train and test the algorithms is crucial for achieving a lightweight security result suitable for IoT devices. In order to lessen the computational burden of a machine-learning algorithm, a point-selection system is being developed. Chi-square tests and correlations have been employed in this procedure.

Correlation: Because of their increased linear dependence on one another, high-correlation characteristics have about the same effect on the dependant variable.

Chi Square: The best features from a predictive model will be selected using this method.

Data Splitting

Data are required for literacy during the machine learning process. to evaluate the efficacy of the algorithm, it is necessary to collect test data in addition to the training data. Seventy percent of the Bot- IoT dataset was used for training, while the remaining thirty percent was used for testing. To split data means to divide it in half, typically for use in a cross-validator setup. A predictive model is constructed using some of the data, while the remaining data is utilised to evaluate the model's accuracy. Data mining model evaluation relies heavily on the ability to split data into separate training and testing groups. Most a subset is utilised for testing when a dataset is split into a set.

Classification

One type of supervised machine learning technique based on Bayes' is the naive bayes algorithm. The maximum likelihood of a situation occurring is used in Bayes' theorems to draw conclusions based on prior knowledge.

The SVM can be used to solve both classification and regression issues; it is a supervised machine learning technique. It performs data transformations using a method called the kernel trick and then uses those to determine an appropriate boundary between the possible outputs.

Performance Analysis

The overall categorization and prediction will be used to create the final result. Many indicators, including, are used to assess the proposed method's performance.

1. **Accuracy**

The proficiency of a classifier is referred to as its accuracy. Accurate class label predictions are made, and predictor accuracy is defined as the degree to which a particular predictor accurately predicts an attribute value for unseen data.

$$Ac= (TP+TN)/ (TP+TN+FP+FN)$$

2. **Precision**

The term "precision" refers to the ratio of correct diagnoses to total diagnoses (positive and negative).

$$Precision=TP/ (TP+FP)$$

3. Recall

The recall rate is calculated by dividing the actual number of returned results by the expected number of returned results. Recall is referred to as sensitivity in binary classification. You might think of it as the odds that your query will return a document that is useful to you.

Recall=TP/ (TP+FN)

4. F-Measure

The F measure, often known as the F1 score or the F score, is a statistical measure of how well a test predicts the true outcome.

F-measure=2TP/ (2TP+FP+FN)

I. System Architecture

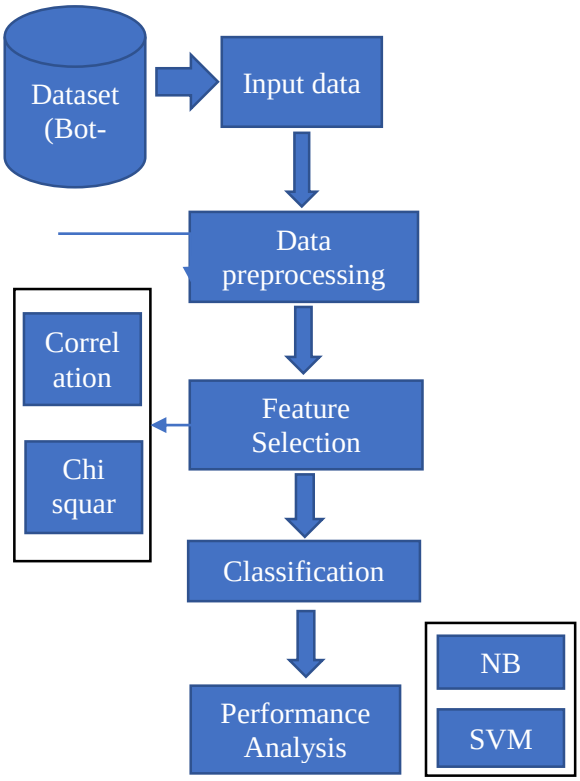


Fig 1.3

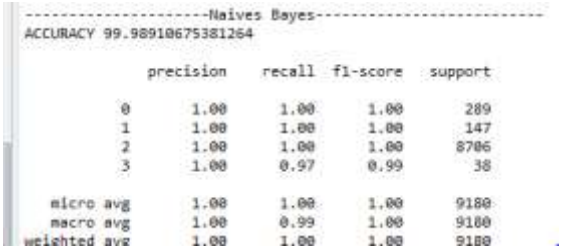
5. Conclusion

In conclusion, exploited machine learning techniques to identify malicious activity in IoT networks. The Bot IoT served as a dataset because of its consistency in updates, breadth of attack types, and variety of network protocols. Using the Bot-IoT dataset, we test our proposed method. In the end, the data was run through four prevalent machine learning algorithms with varying degrees of quality. The F-measure presentation ratios attained by these algorithms are as follows: The SVM accuracy was 95.2%, while the NB accuracy was 99.9%. Four supervised algorithms were the focus of this study. Analysis of experimental data validated our proposed strategy as effective, with an average success rate of >98%.

Future Work

It would be useful to assess the efficacy of various unsupervised algorithms in upcoming research. We also used many machine learning techniques, each in isolation. In the future, we hope to enhance the detection performance by combining many machine learning methods into a multi-layered model.

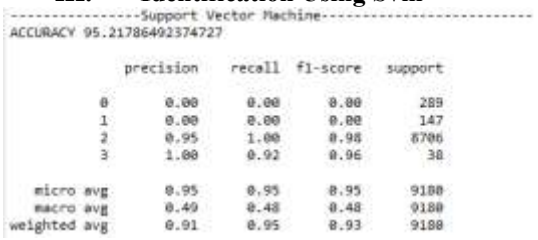
II. Identification Using Nb



```
----------------------Naives Bayes--------------------------
ACCURACY 99.98910675381264

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       289
           1       1.00      1.00      1.00       147
           2       1.00      1.00      1.00      8706
           3       1.00      0.97      0.99        38

   micro avg       1.00      1.00      1.00      9180
   macro avg       1.00      0.99      1.00      9180
weighted avg       1.00      1.00      1.00      9180
```

Fig 1.4

III. Identification Using Svm



```
---------------Support Vector Machine--------------------------
ACCURACY 95.21786492374727

              precision    recall  f1-score   support

           0       0.00      0.00      0.00       289
           1       0.00      0.00      0.00       147
           2       0.95      1.00      0.98      8706
           3       1.00      0.92      0.96        38

   micro avg       0.95      0.95      0.95      9180
   macro avg       0.49      0.48      0.48      9180
weighted avg       0.91      0.95      0.93      9180
```

Fig 1.5

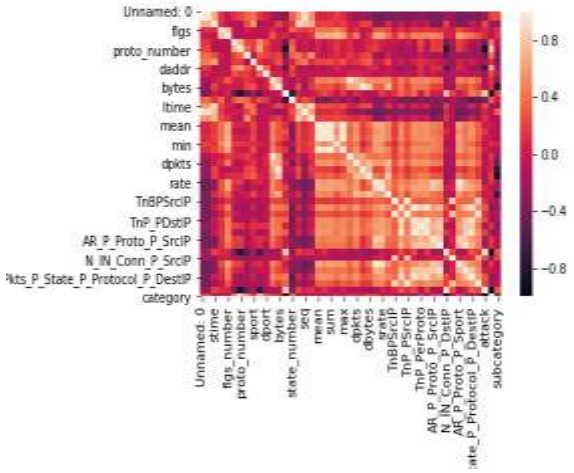


Fig 1.6

6. References

J. Qiu, Z. Tian, C. Du, Q. Zuo, S. Su, and B. Fang, "A survey on access control in the age of internet of things," IEEE Internet of Things Journal, 2020.

Y. N. Soe, Y. Feng, P. I. Santosa, R. Hartanto, and K. Sakurai, "Implementing lightweight iot-ids on raspberry pi using correlationbased feature selection and its performance evaluation," in International Conference on Advanced Information Networking and Applications. Springer, 2019, pp. 458–469.

K. Lab. (2019) Amount of malware targeting smart devices more than doubled in. [Online].

J. Qiu, L. Du, D. Zhang, S. Su, and Z. Tian, "Nei-tte: Intelligent traffic time estimation based on fine-grained time derivation of road segments for smart city," IEEE Transactions on Industrial Informatics, 2019.

J. P. Anderson, "Computer security threat monitoring and surveillance, 1980. Last accessed: November 30, 2008."

D. E. Denning, "An intrusion-detection model," IEEE Transactions on software engineering, no. 2, pp. 222–232, 1987.

L. Wu, X. Du, W. Wang, and B. Lin, "An out-of-band authentication scheme for internet of things using blockchain technology," in 2018 International Conference on Computing, Networking and Communications (ICNC). IEEE, 2018, pp. 769–773.

Z. Tian, X. Gao, S. Su, and J. Qiu, "Vcash: A novel reputation framework for identifying denial of traffic service in internet of connected vehicles," IEEE Internet of Things Journal, vol. 7, no. 5, pp. 3901– 3909, May 2020.

S. Alharbi, P. Rodriguez, R. Maharaja, P. Iyer, N. Bose, and Z. Ye, "Focus: A fog computing-

based security system for the internet of things," in 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, 2018, pp. 1–5.

Z. Tian, C. Luo, J. Qiu, X. Du, and M. Guizani, "A distributed deep learning system for web attack detection on edge devices," IEEE Transactions on Industrial Informatics, 2020. Vol 16(3): 1963-1971.

D. Ventura, D. Casado-Mansilla, J. López-de Armentia, P. Garaizar, D. López-de Ipina, and V. Catania, "Ariima: a real iot implementation of a machine-learning architecture for reducing energy consumption," in International Conference on Ubiquitous Computing and Ambient Intelligence. Springer, 2014, pp. 444–451.

M. Shafiq, X. Yu, A. K. Bashir, H. N. Chaudhry, and D. Wang, "A machine learning approach for feature selection traffic classification using security analysis," The Journal of Supercomputing, vol. 74, no. 10, pp. 4867–4892, 2018.

R. Xue, L. Wang, and J. Chen, "Using the iot to construct ubiquitous learning environment," in 2011 Second International Conference on Mechanic Automation and Control Engineering. IEEE, 2011, pp. 7878– 7880.

M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996– 2018, 2014.

M. Shafiq, X. Yu, A. A. Laghari, and D. Wang, "Effective feature selection for 5g in applications traffic classification," Mobile Information Systems, vol. 2017, 2017.