A ROBUST DETECTION OF CYBER INCIDENTS UTILIZING MACHINE LEARNING TECHNIQUES

Mrs RATNAKUMARI JOGI[1], VANKAYALAPATI JYOTHI SAIPRIYA[2], SYED TEHZEEBA[3], THOTA SRINIVASA KRISHNA[4], SHAIKKHADAR BASHA[5]

[1]Department of CSE & AI, Chalapathi Institute Of Engineering And Technology, LAM, Guntur, Andhra Pradesh, India.

[2]Department of CSE & AI, Chalapathi Institute Of Engineering And Technology, LAM, Guntur, Andhra Pradesh, India.

[3]Department of CSE & AI, Chalapathi Institute Of Engineering And Technology, LAM, Guntur, Andhra Pradesh, India.

[4]Department of CSE & AI, Chalapathi Institute Of Engineering And Technology, LAM, Guntur, Andhra Pradesh, India.

[5]Department of CSE & AI, Chalapathi Institute Of Engineering And Technology, LAM, Guntur, Andhra Pradesh, India.

Abstract: A reliable Cyber Attack Detection Model (CADM) is a system that works as safeguard for the users of modern technological devices and assistant for the operators of networks. The research paper aims to develop a CADM for analyzing the network data patterns to classify cyber-attacks. CADM finds out attack wise detection accuracy using ensemble classification method. LASSO has been used to extract important features. It can work with large datasets, and it has more visualization capability. Gradient Boosting and Random Forest algorithms have been used for classification of network traffic data to build an ensemble method. Gradient Boosting algorithm trains weak learning models and select the best decision trees to deliver more improved prediction accuracy and Random Forest algorithm trains each tree in parallel manner. In this research work, Jive datasets such as NSL-KDD, KDD Cup 99, UNSWNB15, URL 2016 and CICIDS 2017 are also applied to check the efficiency of the proposed model.

Keywords: Cyber Attack, DBSCAN, Embedded Method, LASSO, Ensemble method, Gradient Boosting, Random Forest.

1. INTRODUCTION

Today, we are living in the society where many things are going to be automated and digitalize. Technology is now involving in our daily life and there are many simple examples for that such as mobile phones, personal computers etc. Converting things to smart devices and making these processes automated, IoT is one of the technology which plays an important role for that purpose. So we can say that it is one of the most important technologies for businesses as well as for our daily life. But, it is important to remember that as the technology increases there are also a number of issues increases related with that technology. Similarly, as the number of devices connected it means the more information is sharing between these devices and if there is any type of bug in the sharing system, there is a chance that each connected device could corrupt, and confidential information could steal by the hacker.

There should be an international standard for compatibility of IoT here which is not yet, therefore it is very difficult for devices which are manufactured from different companies to communicate with each other. Also there are many IoT devices which requires and ask to input user personal information such as name, location and contact as well as data which are important to hackers such as social media information. Therefore, the information sharing between IoT devices needs to be secured. Also, IoT privacy and security are cited as major concerns. There are number of attacks on IoT including malware. Malware can be defined as malicious software or bug which is designed to gain access and damage your device, device could be computer or IoT device.

2. LITERATURE SURVEY

Ashu Bansal et al. [11], has showed that the big data has increased malicious activities such as MITM, DDos and Spoofing. The data dimensionality reduction scheme was proposed to minimize the dimensionality of data to get better detection rate. XGBoost and CTree as ensemble approach and SVM and NNet as standalone classifiers were used. The biggest challenge for using this scheme is to reduce the elapsed time for classification. S. Sandosh et al. [12], proposed a model to mitigate the problems of high accuracy with low complexity and time efficiency. Modified k-means clustering was applied for data segmentation. Then KNN classification algorithm is used to classify the traffic flow as known or unknown attack. Rim Ben Fekih et al. [13], has asserted that cloud computing solves the problems to store large and heterogeneous datasets. A distributed IDS model was presented to handle a large scale of alert data. The Spark tool was used to join and analyze large datasets. The machine learning pipeline developed by sci- kit learns was followed in this work. The model was trained by Naive Bayes classification algorithm. Sumaiya Thaseen et al. [14], has aimed tofind the critical features to build an intrusion detection model. Chi-Square feature selection technique has been applied to select features. Supervised classification algorithms (support vector machine (SVM), modified Naive Bayes (MNB) and LPBoost) were used in ensemble method.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

The existing system focuses on addressing the security and privacy concerns in IoT networks, recognizing the absence of international standards for compatibility in the IoT landscape. The project utilizes the Aposemat IoT-23 dataset, a labeled dataset created in the Avast laboratory, designed specifically to provide real-world examples of IoT attacks. The primary objective is to leverage artificial intelligence techniques to detect and classify unknown network behaviors based on historical data patterns. The machine learning algorithms employed include Decision Tree, Random Forest, and Naive Bayes. Through a comparative analysis, the results indicate that Random Forest proves to be the most efficient algorithm for detecting and classifying IoT network attacks on the Aposemat IoT-23 dataset.

LIMITATIONS OF EXISTING SYSTEM

- **Static Machine Learning Models:** The use of Decision Tree, Random Forest, and Naive Bayes machine learning algorithms implies a static approach to network attack detection. These models might struggle to adapt to dynamic and evolving attack strategies, potentially leading to a decreased accuracy in detecting novel threats.
- **Limited Generalization:** The effectiveness of the system may be constrained by its ability to generalize across diverse IoT network environments. Factors such as network scale, device types, and communication protocols may vary, affecting the system's performance in real-world scenarios that differ from the Aposemat IoT-23 dataset.

3.2 PROPOSED SYSTEM

The proposed system aims to overcome the limitations of the existing approach by introducing several enhancements to strengthen IoT network attack detection using artificial intelligence. Firstly, the system proposes the incorporation of a more diverse set of labelled datasets, beyond the Aposemat IoT-23 dataset, to ensure a comprehensive understanding of evolving attack patterns. This expansion enables the system to generalize better and recognize novel threats that may not be covered by a single dataset.

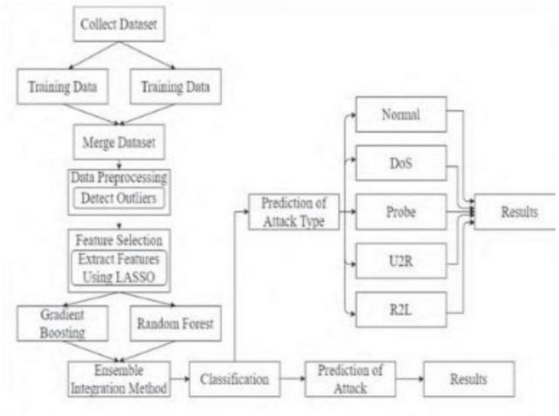
Secondly, the proposed system advocates for the integration of dynamic and adaptive machine learning models that can evolve with the changing nature of IoT attacks. This may involve exploring deep learning techniques or other advanced algorithms capable of capturing intricate patterns and adapting to emerging threats in real-time.

Thirdly, the proposed system emphasizes the development of a more scalable architecture, considering the increasing scale and complexity of IoT networks. This could involve the implementation of distributed computing techniques or lightweight algorithms suitable for resource-constrained IoT devices, ensuring efficient and effective network attack detection at scale.

Additionally, the proposed system suggests leveraging anomaly detection methods alongside traditional classification approaches to enhance the detection of previously unseen attacks. Anomaly detection can identify deviations from normal network behaviour, providing a proactive defense against emerging threats not explicitly defined in the training dataset.

Lastly, the proposed system aims to optimize feature engineering processes to minimize computational overhead. This involves refining pre-processing techniques to strike a balance between improving detection accuracy and ensuring efficient resource utilization, particularly in the context of IoT devices with limited computational capabilities. Overall, the proposed system seeks to advance the state-of-the-art in IoT network attack detection by addressing existing limitations and embracing more dynamic, scalable, and adaptable approaches.

4. SYSTEM ARCHITECTURE



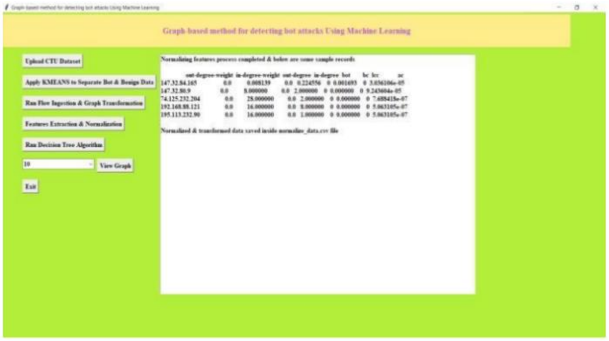
5. MODULES

Data Preprocessing Module: This module focuses on preparing and refining the input data for the machine learning algorithms. It involves tasks such as data cleaning, handling missing values, normalization, and transforming the raw data from the Aposemat IoT-23 dataset into a format suitable for effective learning by the algorithms.

Machine Learning Algorithm Implementation Module: This module encompasses the implementation of machine learning algorithms, including Decision Tree, Random Forest, and Naive Bayes. Each algorithm is configured and trained on the preprocessed data to learn and recognize patterns indicative of IoT network attacks.

Dataset Integration Module: The system integrates multiple labeled datasets beyond the Aposemat IoT-23 dataset to ensure a more comprehensive understanding of diverse attack patterns. This module facilitates the combination of various datasets, promoting a broader and more adaptable detection capability.

6. RESULT


Graph-based method for detecting bot attacks Using Machine Learning
Graph-based method for detecting bot attacks Using Machine Learning
Upload CTU Dataset
Apply KMEANS to Separate Bot & Benign Data
Run Flow Ingestion & Graph Transformation
Features Extraction & Normalization
Run Decision Tree Algorithm
10
View Graph
Exit

```
Normalizing features process completed & below are some sample records

               out-degree-weight  in-degree-weight  out-degree  in-degree  bot     bc  lcc           ec
147.32.84.165              0.0          0.008139         0.0   0.224556    0  0.001693   0  3.036106e-05
147.32.80.9               0.0          8.000000         0.0   2.000000    0  0.000000   0  9.243604e-05
74.125.232.204             0.0         28.000000         0.0   2.000000    0  0.000000   0  7.688415e-07
192.168.88.121             0.0         16.000000         0.0   8.000000    0  0.000000   0  5.063105e-07
195.113.232.90             0.0         16.000000         0.0   1.000000    0  0.000000   0  5.063105e-07

Normalized & transformed data saved inside normalize_data.csv file
```

7. CONCLUSION

The best effort has been put to produce a better predictive model with better accuracy and low false positives and false negative. In this study, the model has worked with 19 attacks in five datasets. The model gained better accuracy as well as better precision and recall with lower false positive rate for the five datasets with different dimensions than the other existing models. Our extreme effort has been made not to make the dimensions of the dataset shorter since higher dimensions make the model less efficient. For two attack types (Normal and R2L), 100% accuracies were obtained and for the other two (DoS and Probe), 99.96% accuracies were obtained and for U2R, we got 99.969% for NSL-KDD dataset. For all of four attacks better accuracy has been gained.

FUTURE SCOPE: In further research, it can be tried to improve the accuracy for this attack. But one advantage that have been obtained is understood that this is one of the attacks whose pattern is complex, it is needed to work separately for the separate attack's dataset. The field of this study can be increased by working with more attacks in other publicly available datasets. The prevention system can also be developed in further research.

REFERENCES

[1]VibekanandaDutta ,MichałChora´s, Marek Pawlicki and RafałKozik, "A Deep Learning Ensemble for Network Anomaly and Cyber-Attack Detection", Sensors, August 2020.

[2] Quoc-Dung Ngo, Huy-Trung Nguyen, Van-Hoang Le, Doan-Hieu Nguyen, "A survey of IoT malware and detection methods based on static features", ICT Express, December 2020.

[3] B. Ahmad, W. Jian and Z. Anwar Ali, "Role of Machine Learning and Data Mining in Internet Security: Standing State with Future Directions," Journal of Computer Networks and Communications, vol. 2018, pp. 1-10, 2018. doi: 10.1155/2018/6383145 [Accessed 2 October 2020].

[4] A. Gupta, G. Prasad and S. Nayak, "A New and Secure Intrusion Detecting System for Detection of Anomalies Within the Big Data," Studies in Big Data, pp. 177-190, 2018. doi: 10.1007/978-3-030-03359- 0_8 [Accessed 30 August 2020].

[5] T. Tang, D. McLernon, L. Mhamdi, S. Zaidi and M. Ghogho, "Intrusion Detection in SDN-Based Networks: Deep Recurrent Neural Network Approach," Deep Learning Applications for Cyber Security, pp. 175-195, 2019. doi: 10.1007/978-3-030-13057-2_8' [Accessed 30 August 2020].

[6] C. Gayathri Harshitha, M. Kameswara Rao and P. Neelesh Kumar, "A Novel Mechanism for Host-Based Intrusion Detection System," In Proc. First International Conference on Sustainable Technologies for Computational Intelligence, pp. 527-536, 2019. doi: 10.1007/978-981-15- 0029-9 42 [Accessed 21 June 2020].

[7] A. Ahmim, M. Ferrag, L. Maglaras, M. Derdour and H. Janicke, "A Detailed Analysis of Using Supervised Machine Learning for Intrusion Detection," Strategic Innovative Marketing and Tourism, pp. 629-639, 2020. doi: 10.1007/978-3-030-36126-6 70 [Accessed 7 August 2020].

[8] R. Jaiswal and S. Lokhande, "Analysis of Early Traffic Processing and Comparison of Machine Learning Algorithms for Real Time Internet Traffic Identification Using Statistical Approach," Advanced Computing, Networking and Informatics, vol. 2, Smart Innovation, Systems and Technologies, vol 28, pp. 577-587, 2014. doi: 10.1007/978-3-319-07350-7 64 [Accessed 24 September 2020].

[9] W. Zong, Y. Chow and W. Susilo, "Interactive three-dimensional visualization of network intrusion detection data for machine learning," Future Generation Com puter Systems, vol. 102, pp. 292-306, 2020. doi: 10.1016/j.future.2019.07.045 [Accessed

[10] H. Liu and A. Gegov, "Collaborative Decision Making by Ensemble Rule Based Classification Systems," Studies in Big Data, pp. 245-264, 2015. doi: 10.1007/978-3-319-16829-6_10 [Accessed 20 September 2020].

[11] A. Bansal and S. Kaur, "Data Dimensionality Reduction (DDR) Scheme for Intrusion Detection System Using Ensemble and Standalone Classifiers," In Proc. International Conference on Advances in Computing and Data Sciences, vol. 1045, pp. 436-451, 2019. doi: 10.1007/978-981-13-9939-8 39 [Accessed 15 July 2020].

[12] S. Sandosh, V. Govindasamy and G. Akila, "Enhanced intrusion detection system via agent clustering and classification based on outlier detection," Peer-to-Peer Networking and Applications, vol. 13, no. 3, pp. 1038-1045, 2020. doi: 10.1007/s12083-019-00822-3 [Accessed 15 July 2020].