International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Using a Multi- Layered Framework for Botnet Detection Based on Machine Learning Algorithms

[1]Aditee Mattoo, [2]Soumya A K, [3]Vineet Saxena, [4]Manish Shrivastava

Submitted:20/04/2023 **Revised:**10/06/2023 **Accepted:**25/06/2023

Abstract: By allowing attackers to take control of a significant amount of infected devices for illegal purposes, botnets pose severe challenges to network security. Due to the dynamic nature of botnet infrastructures and the advanced tactics used by attackers, identifying and preventing attacks by botnets is a complex undertaking. In this study, we present a novel machine learning (ML)-based black hole optimized random forest (BHO-RF) for botnet detection. The BHO method enhances the performance of the RF classifier by modifying the hyperparameters, boosting its ability to recognize botnet traffic. It is inspired by the behaviour of black holes in space. We have extensive tests utilizing the CTU-13 dataset to assess the efficacy of our suggested framework. The outcomes show that our multi-layered strategy outperforms conventional techniques, delivering higher accuracy, f1-score, precision, and recall in botnet identification. The approach also demonstrates robustness over noise and changes in transmission characteristics.

Keywords:*Botnet, security, machine learning (ML), black hole optimized random forest (BHO-RF)*

1. Introduction

A botmaster, a hacker remotely commands infected devices, is called a botnet. The term "botnet" is a combination of the words "robot" and "network," and refers to the network of robots that act as the botmaster's goons. The botnet's responsibility is to carry out attacks as directed by the directives given to it by the botmaster. Information security is now significantly at risk from botnet attacks, a severe problem. Botmasters and researchers, who defend botnets, are engaged in an ongoing arms race (Khan et al., 2019). Each side keeps honing its abilities to prevail in combat. The vast number of bots in the botnet strengthens attacks because of their sheer volume. The power of botmasters to keep the bots hidden from security systems also plays a significant role in supporting the bots. One of the most popular botnets, the Mirai Botnet, shocked everyone worldwide by having so many infected devices (Ibrahim et al., 2021).

Web cams, closed-circuit television cameras (CCTV), and other IoT gadgets with lax security were all used by the Mirai botnet to spread itself using Trojans. With 100,000 IoT devices engaged, the most severe Mirai attack resulted in a 1.3 Tbps attack. The present methods for detecting botnets are signature-based algorithms that are effective at finding similar botnet kinds or well-known botnets instead fail to see a previously unknown or evolved botnet (Vinayakumar et al., 2020). Botnets are currently changing to evade detection by security measures. Making sure that no one can access the packet data is one of the ways, for instance, by employing a concealment method like encrypting it, confusion, or a personal online network. The disadvantage of detection based on signatures and network-based IDS is that they cannot detect malware when obfuscation techniques are in use because of the limitations of the present detection models. As a result, scientists are working to develop a malware detection algorithm without gaining access to the packet's content (Alqatawna et al., 2021). The restricted updated assaults dataset for research is due to the packet's content, which may hurt people aside from that. A technique for examining network traffic without having access to the material is behavior-based analysis. The packet header rather than the payload is used in research based on behavior to preserve the confidentiality of sensitive packet data. The benefit of behavior-based network traffic analysis is that it can identify malware that uses a VPN or other obfuscation or encryption techniques (Shinan et al., 2021). The issue is the requirement for a framework for reliable and effective botnet detection that can identify changing botnet strategies, circumvent evasion techniques, take into account various layers of information, and reduce false positives and false negatives while maximizing

1Assistant Professor & Dy. HoD, Department of Information Technology and M.Tech Integrated, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India, Email id: aditeemattoo@niet.co.in

2Assistant Professor, Department of Computer Science and IT, Jain(Deemed-to-be University), Bangalore-27, India, Email Id: soumya.k@jainuniversity.ac.in

3Assistant Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Email id: tmmit_cool@yahoo.co.in

4Professor, Department of Computer Science & Engineering, Vivekananda Global University, Jaipur, India, Email Id: manish.shrivastava@vgu.ac.in

network security.

In this paper, we introduce a unique black hole optimized random forest (BHO-RF) for botnet identification that is ML-based. By altering the hyperparameters, the BHO approach improves the RF classifier's functionality and increases its capacity to identify botnet traffic.

The remainder of the paper is divided into subsequent parts. Part 3 contains the proposed method explained. Part 4 contains the results and analysis.while Part 5 discusses the conclusions.

2. Related Works

Letteri et al., 2018 explained the task of detecting botnets is crucial since they are currently among the most pervasive and deadly types of malware on the internet. However, a lot of research in this area relies on outdated or biased traffic samples, exploits general malware detection methods, and produces conclusions that are not entirely trustworthy. Costa et al.,2019 described internet security as being threatened by mobile botnets. These botnets prey on less powerful, less secure devices while occasionally exploiting their unique capabilities, including SMS texts. In order to identify mobile botnets with attributes generated from system calls, they suggest a host-based strategy employing machine learning techniques. Applications with comparable actions often share the same patterns that are developed. As a result, various botnets probably share similar ways of system calls. Gadelrab et al.,2018 stated researchers provide a thorough explanation of the methodology to create a machine learning (ML)-based botnet detection system. Numerous research projects have focused on identifying botnet traffic or detecting botnet member hosts. First, the requirement for Deep Packet Inspection-DPI and the need to gather traffic from numerous infected hosts are two significant drawbacks of current botnet detection methods that this research tries to address.

Gaonkar et al.,2020 expressed security issues have increased along with the rate of internet usage. Botnets are a significant danger to network security. A botnet is described as a grouping of different bots that the botmaster manages via the Command and Control (C&C) channel. In recent years, several technologies and methods have been put out to monitor the identification of botnets. Soe et al., 2020 described Devices connected to the Internet of Things (IoT) are becoming the focus of a growing variety of cyberattacks because of their rapid expansion and widespread acceptance. According to one report, botnet-based attacks make up the bulk of incidents in IoT settings. There are still numerous security holes in IoT devices since most lack the storage and processing capacity required for efficient security measures. Moorthy et al., 2023 aimed that over 80 million records were compromised by cybercriminals in 2021. The majority of these attacks took the form of cyberattacks and ransomware attacks. Attackers utilize various techniques to target specific users, but they rely on Botnet forces to infiltrate an organization. A botnet is a network that has been infected with malware and is run by a single attacker known as the bot-herder. Wai et al., 2018 presented The control of connected devices as a botnet for criminal behavior is possible since they are more prone to malware infestations. preventing the spread of botnets and securing local networks and infrastructure, quick detection of infected workstations is necessary. Letteri et al.,2019 proposed as the Internet of Things (IoT) spreads, everyone's life is becoming more integrated with cyber-physical intelligent gadgets, but this also exposes them to malware made for traditional online applications, such as botnets. Since botnets are among the most pervasive and harmful malware, it is crucial to find them. Numerous research in this area rely on outdated or skewed traffic samples and use general malware detection algorithms, which reduces the reliability of their findings. Yang et al., 2022 suggested the incredibly dangerous botnet is a brand-new attack technique created and built on the foundation of conventional harmful code like network viruses and backdoor tools. In order to identify and categorize botnets, this course uses deep learning and neural networking methods. Koroniotis et al., 2018 said that the Internet of Things (IoT) is a system of linked, commonly encountered "things" that have been marginally enhanced with computing power. The Internet of Things (IoT) has recently been impacted by numerous botnet-related activities. Existing Network Forensic approaches cannot identify and monitor the most sophisticated botnet strategies today since botnets have been responsible for significant security issues and financial harm over the years.

3. Methodology

3.1. Black Hole Optimisation Algorithm (BHO)

The Black Hole Optimisation Algorithm (BHO) is a search approach that draws inspiration from the behaviour of black holes in actual space. A black hole's powerful gravitational pull creates a tremendous density in a virtual area. More stars are drawn into it due to the enormous gravitational forces it adds. When the lead in the galaxy collapses, a black hole is created in that area. It is referred to as black since it doesn't absorb anything. BHA has adopted this behavior. An event horizon is a sphere-shaped structure that forms space-based black holes. Schwarzschild radius is the term used to describe this event horizon radius. A star is sucked into the black hole as it approaches the event horizon, which has an enormous mass and a powerful gravitational pull and would then disappear. Since there is no escape from inside an event horizon due to the escape velocity being equal to the quantity of light speed in the area, The Euclidean distance between a star and a black hole is calculated in a BHA.

Replaces the existing star in any location within the search space when the distance is less than the Schwarzschild radius. In the BHA, it is suitable to update their places when a new star reaches the black hole at the lowest cost. With the Hyperbolic Tangent function, stars can be moved around. In the search area, stars are positioned at random. The solutions of a BHOA are affected by the stars. The fitness evaluation function calculates an individual fitness value for each star. The classifier's accuracy score is taken into account while calculating fitness. The best ideal solution is referred to as a black hole. The star with the highest fitness value is selected as the black hole. When a star has the bare minimum of features, and those traits have the same fitness value as those found in a black hole, the star's location should be adjusted. As soon as accuracy improves noticeably, increase the number of iterations. The categorization accuracy is much enhanced by this strategy.A flowchart for the Black Hole Optimisation Algorithm (BHO) is shown in Figure 1.

3.2. Popular stars

One hundred stars are thought to exist. There have been 500 iterations, according to this estimate. The initial stage in BHA is to initialize the population of stars. Choose a random set of qualities, and then fill in the stars. Add the property to the star's data collection if the value generated at random is 1. The amount of stars is regarded as performance equal to or better.

3.3. Updated star position

Using the tangent function with a threshold value of 0.6, the position of the stars was updated. In comparison to the sigmoid function, the hyperbolic tangent function has better performance. The hyperbolic tangent function, which has a return value of either zero or one, is used to choose a feature. As a result, the stars' positions are updated according to equations (1) and (2).

$$U(h_{jt}(e+1)) = epg\left(tanh\left(h_{jt}(e+1)\right)\right) \qquad (1)$$

$$h_{jt}(e+1) = \begin{cases} 1 \; if \; U\left(h_{jt}(e+1)\right) > rand \\ 0 \qquad\qquad\qquad otherwise \end{cases} \qquad (2)$$

The string's length corresponds to the total number of characteristics. Consideration of the classifier's precision as a fitness function is used to calculate each star's optimal value. In the BHO approach, a subset of the best features is provided; this shortens training time and increases classification model precision. Thallium scan, kind of pain in the chest, and exercise-induced angina are the most significant and pertinent elements in the dataset according to the BHO algorithm.

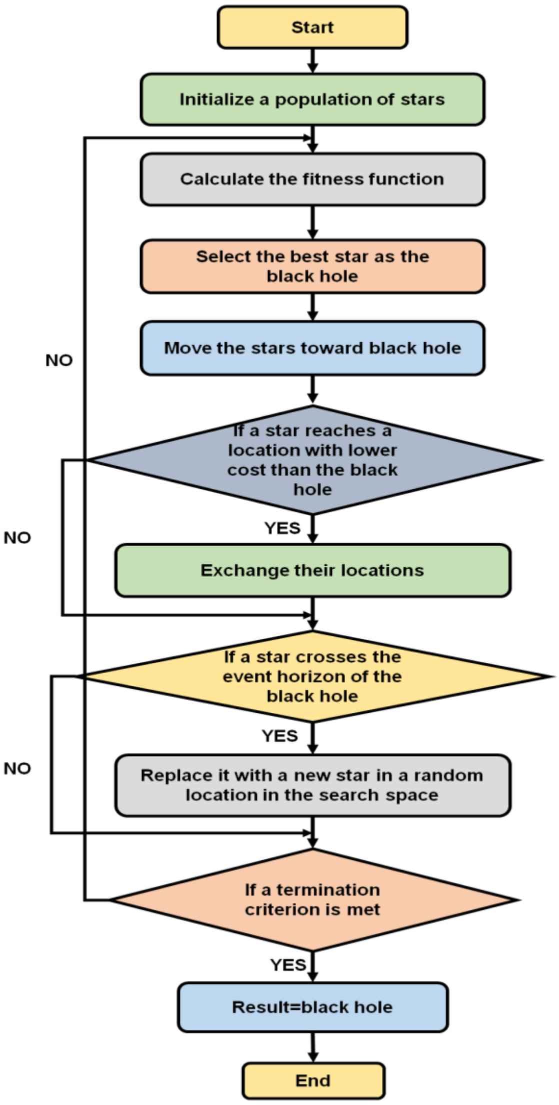
Start
Initialize a population of stars
Calculate the fitness function
Select the best star as the black hole
Move the stars toward black hole
If a star reaches a location with lower cost than the black hole
YES
Exchange their locations
If a star crosses the event horizon of the black hole
YES
Replace it with a new star in a random location in the search space
If a termination criterion is met
YES
Result=black hole
End
NO
NO
NO

Fig.1. BHO flowchart

3.4. Random Forest (RF)

Using the concept of randomization, the ensemble classifier known as random forest builds a collection of distinct and non-identical decision trees. Every decision tree uses a vector of randomness as a parameter, selects characteristics from a sample at random, and selects a subset of the training set from the sample set of data at random. This is the algorithm used to create the random forest. Each decision tree's training data set corresponds to a certain number of samples, indicated by the variables k and n in the random forest model. When segmenting one of the nodes of a decision tree, the term "M" refers to the sample's feature number, which is expressed as the number of features.

1. Make a k group exercise set. (bootstrap sampling) selecting samples from each training sample set N times using a repeated sampling technique. Construct a decision tree using models for each training set.

Taken from "out of the bag" (also known as "OOB") data from k groups;

2. Based on the m attributes that are randomly selected for each node of the node in the decision tree, determine the best segmentation characteristics for each node.;
3. Each decision tree continues to grow unabated;
4. Use a random forest model, which is created by combining many decision trees, to locate and categorize the unknown data.

4. Result and Discussion

Python and Scikit-learn were used throughout the study. Anaconda, which incorporates the open-source Jupyter Notebook, was used.CTU-13 dataset is used in this research. Data is divided into two categories: training (70%) and testing (30%). The data model has 769 entries, each representing a link, an accumulation of a few packet flows, mixed with 335 regular and 434 malicious entries of images.

Table 1. Confusion matrix

	Attack	Non Attack
Attack	TP = 430	FP = 20
Non Attack	FN = 4	TN = 315

A typical method of determining accuracy is to compare the total number of occurrences (both positive and negative) in the dataset to the number of cases that were correctly classified (true positives and true negatives). It evaluates the detection system's performance in accurately differentiating routine and botnet-related activities. the accuracy of the proposed method is high than existing methods in botnet detection (Figure 2).

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \qquad (3)$$

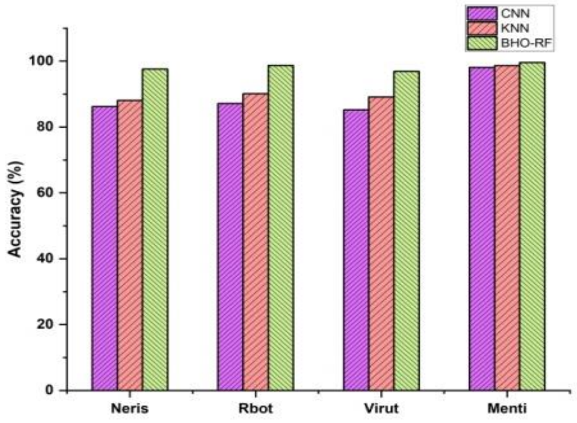

Fig.2. Accuracy of the suggested and current approaches.

Using a precision performance indicator, one may assess how well a detection system has identified instances as botnets. It determines the ratio of genuine positives (occurrences that were successfully identified as botnets) to the sum of true positives and negatives (occurrences that were mistakenly classed as botnets). True positives plus false positives are multiplied to get precision. Compared to other approaches to botnet detection, the suggested method has higher precision (Figure 3).

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

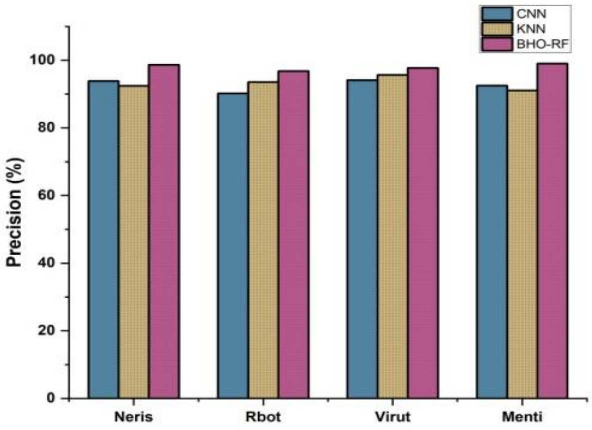

Fig.3. Precision of the existing and proposed methodologies.

Recall, often referred to as sensitivity or actual positive rate, is a performance indicator that assesses how well a detection system can accurately identify every instance of a botnet. It calculates the proportion of true positives (botnets that were accurately recognized) to the sum of fake negatives and genuine positives (botnet instances that were not discovered). Compared to the ones already in use, the proposed method has a greater recall rate for botnet identification (Figure 4).

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

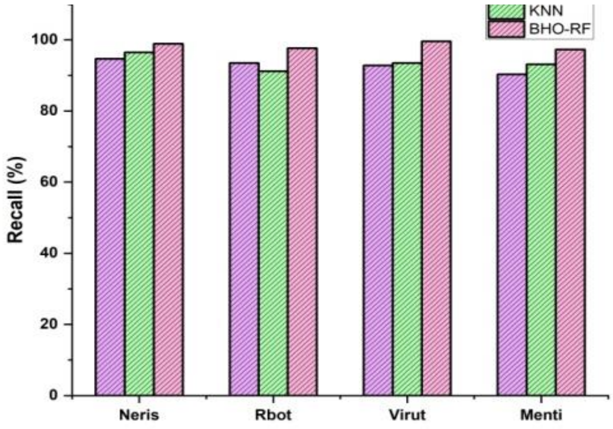

Fig.4. Recall the suggested and current approaches.

The F1 score is generated using the harmonic mean of both recall and precision, giving both measures equal weight. By taking both false positives and false negatives into account, it offers a fair evaluation of the model's performance. Calculating the F1 score uses the following formula: The suggested method outperforms existing botnet detection techniques regarding the f1 score (Figure 5).

$$F1 = \frac{2\times precision\times recall}{precision+recall} \quad (6)$$

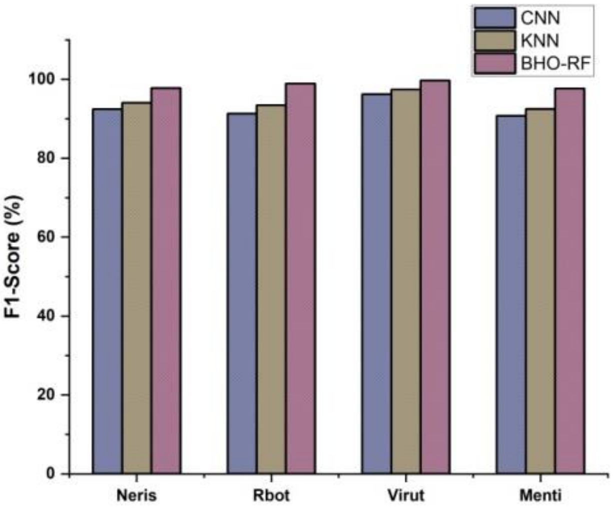

Fig.5. The F1-score with existing and suggested techniques

5. Conclusion

The Multi-Layered Framework for Botnet Detection offers a thorough method for locating and reducing the threat posed by botnets. This framework provides a higher level of security and accuracy in botnet identification by combining numerous detection approaches and a layered structure. In this paper, we proposed a novel machine learning (ML)-based black hole optimized random forest (BHO-RF) for botnet detection. The BHO method enhances the performance of the RF classifier by modifying the hyperparameters, boosting its ability to recognize botnet traffic. The outcomes show that our multi-layered strategy outperforms conventional techniques, delivering higher accuracy, f1-score, precision, and recall in botnet identification. In a subsequent investigation, a few difficulties still need to be addressed. As we can see, clustering the decentralized botnet reduced performance. Plans call for expanding our approach to test various unique botnet kinds and rank them according to efficiency and execution time. Using numerous benchmark botnet datasets, we want to build an evolving structure to forecast future botnet behavior.

References

[1] Khan, R.U., Zhang, X., Kumar, R., Sharif, A., Golilarz, N.A. and Alazab, M., 2019. An adaptive multi-layer botnet detection technique using machine learning classifiers. Applied Sciences, 9(11), p.2375.

[2] Ibrahim, W.N.H., Anuar, S., Selamat, A., Krejcar, O., Crespo, R.G., Herrera-Viedma, E. and Fujita, H., 2021. Multilayer framework for botnet detection using machine learning algorithms. IEEE Access, 9, pp.48753-48768.

[3] Vinayakumar, R., Alazab, M., Srinivasan, S., Pham, Q.V., Padannayil, S.K. and Simran, K., 2020. A visualized botnet detection system based on deep learning for the internet of things networks of smart cities. IEEE Transactions on Industry Applications, 56(4), pp.4436-4456.

[4] Alqatawna, J.F., Ala'M, A.Z., Hassonah, M.A. and Faris, H., 2021. Android botnet detection using machine learning models based on a comprehensive static analysis approach. Journal of Information Security and Applications, 58, p.102735.

[5] Shinan, K., Alsubhi, K., Alzahrani, A. and Ashraf, M.U., 2021. Machine learning-based botnet detection in software-defined network: a systematic review. Symmetry, 13(5), p.866.

[6] Letteri, I., Della Penna, G. and De Gasperis, G., 2018. Botnet detection in software defined networks by deep learning techniques. In Cyberspace Safety and Security: 10th International Symposium, CSS 2018, Amalfi, Italy, October 29–31, 2018, Proceedings 10 (pp. 49-62). Springer International Publishing.

[7] Costa, V.G.T.D., Barbon, S., Miani, R.S., Rodrigues, J.J. and Zarpelão, B.B., 2019. Mobile botnets detection based on machine learning over system calls. International Journal of Security and Networks, 14(2), pp.103-118.

[8] Gadelrab, M.S., ElSheikh, M., Ghoneim, M.A. and Rashwan, M., 2018. BotCap: Machine learning approach for botnet detection based on statistical features. Int. J. Commun. Netw. Inf. Secur, 10(3), p.563.

[9] Stojanovic, N. . (2020). Deep Learning Technique-Based 3d Lung Image-Based Tumor Detection Using segmentation and Classification. Research Journal of Computer Systems and Engineering, 1(2), 13:19. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/6

[10] Shende, P. ., Vishal Ashok, W. ., Limkar, S. ., D. Kokate, M. ., Lavate, S. ., & Khedkar, G. . (2023).

Assessment of Seismic Hazards in Underground Mine Operations using Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 237–243. https://doi.org/10.17762/ijritcc.v11i2s.6142

[11] Gaonkar, S., Dessai, N.F., Costa, J., Borkar, A., Aswale, S. and Shetgaonkar, P., 2020, February. A survey on botnet detection techniques. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-6). IEEE.

[12] Soe, Y.N., Feng, Y., Santosa, P.I., Hartanto, R. and Sakurai, K., 2020. Machine learning-based IoT-botnet attack detection with sequential architecture. Sensors, 20(16), p.4372.

[13] Moorthy, R.S.S. and Nathiya, N., 2023. Botnet Detection Using Artificial Intelligence. Procedia Computer Science, 218, pp.1405-1413.

[14] Wai, F.K., Lilei, Z., Wai, W.K., Le, S. and Thing, V.L., 2018, October. Automated botnet traffic detection via machine learning. In TENCON 2018-2018 IEEE Region 10 Conference (pp. 0038-0043). IEEE.

[15] Letteri, I., Penna, G.D. and Gasperis, G.D., 2019. Security in the internet of things: botnet detection in software-defined networks by deep learning techniques. International Journal of High Performance Computing and Networking, 15(3-4), pp.170-182.

[16] Yang, X., Guo, Z. and Mai, Z., 2022, July. Botnet Detection Based on Machine Learning. In 2022 International Conference on Blockchain Technology and Information Security (ICBCTIS) (pp. 213-217). IEEE.

[17] Koroniotis, N., Moustafa, N., Sitnikova, E. and Slay, J., 2018. Towards developing network forensic mechanism for botnet activities in the IoT based on machine learning techniques. In Mobile Networks and Management: 9th International Conference, MONAMI 2017, Melbourne, Australia, December 13-15, 2017, Proceedings 9 (pp. 30-44). Springer International Publishing.