National Conference on Advancements in Computer Science and Engineering
In association with International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

A P2P Botnet Detection Technique Using Machine Learning Classifiers

Yash Patwa[1], Tulika Kotian[1], Ralin Tuscano[1], Ms. Alvina Alphonso[1], Dr. Nazneen Ansari[2]
[1]Department of Information Technology, University of Mumbai, Mumbai, Maharashtra, India
[2]Department of Computer Engineering, University of Mumbai, Mumbai, Maharashtra, India

ABSTRACT

Today, botnets prove to be one among many scandalous perils to security in networks. While Client-Server botnets employ a centralized communication architecture, Peer-to-Peer(P2P) botnets acquire a decentralized structure for trafficking commands and controlling data, hence making them more difficult to be identified in a network. In this paper, the authors propose an effective system to detect Peer-to-Peer botnets by applying machine learning algorithms to network traffic parameters. The data from the CTU-13 dataset is input into the system. The proposed system has 3 phases. In the first stage, feature reduction was performed on the network traffic to recognize which of the features affected the classification considerably. In the second stage, the detection model was developed, which classified the traffic into Botnet(malign) traffic and Legitimate(benign) traffic in the last phase. The output of the system generates the classification of the network traffic with visualizations to gain insights into the network activity. The five machine learning algorithms employed are Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Logistic Regression, and Naive Bayes. On performing comparative analysis, the Decision Tree algorithm successfully detected Peer-to-Peer botnet traffic by demonstrating an accuracy of 99.90%.

Keywords — botnet detection, decision tree algorithm, machine learning, network security, P2P botnets, ReactJS

I. INTRODUCTION

In today's world of a billion Internet-connected devices, security issues are increasing as the network environments have become more intricate. Conficker was one of the largest botnets so far that affected 10.5 million computers. The damage caused on such a scale is huge and for the hackers, more damage means more profit. This had a huge impact on various sectors including government organizations, large institutions, and almost every social networking website like Facebook, Twitter, Instagram, etc., e-commerce websites Amazon, Flipkart, etc., in short, every firm on the internet was compromised by this malware.

A Botnet is a network of robots used for committing a cybercrime on the internet. A botnet mainly consists of three elements that are mandatory for performing malicious activity successfully: attackers, bots, and handlers [5]. The cybercriminals controlling the botnets are called Botmasters. Centralized botnets can be detected and destroyed. However, P2P botnets are difficult to detect and destroy because of their cagey nature [4]. Thus, P2P botnets are becoming popular among the attackers.

Today more and more businesses are coming online, thus securing businesses against physical and data threats is becoming very important. Now, with many devices communicating with each other over wired, wireless, or cellular networks, more complex and distributed networks are into the picture. Today, mobile botnets have become very popular due to the increased use of smartphones [2]. Mobile botnets pose serious threats to mobile security. Hence, network security becomes a very important concept.

Classification in machine learning is a supervised machine learning approach in which the model is trained with labeled data which helps the model to classify new observations based on what it has learned from the data.

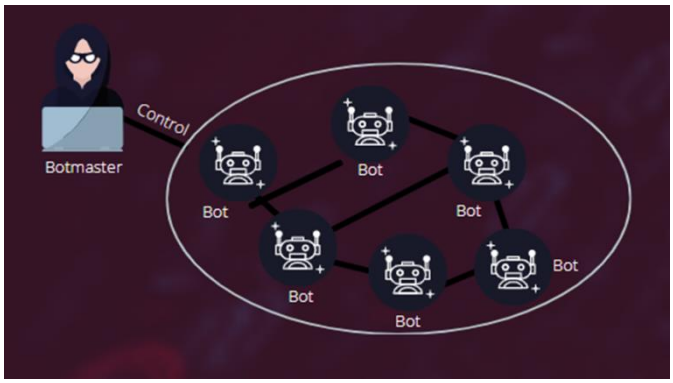

Fig. 1 P2P Botnet Structure

Due to the increase of IoT devices in networks, network security is getting more convoluted. Botnets will continue to be one of the major reasons for jeopardy in cybersecurity. Moreover, these kinds of malicious software are available in the market for free leases. Hence, a model to protect systems from one of the popular malware - botnets, has been proposed. This system proposes the development of a multi-phase detection of botnets that will make use of machine learning classifiers like Decision tree, SVM, Logistic Regression, etc.

II. RELATED WORK

Paper [1] proposed the implementation of a host-based intrusion detection system for botnet detection. The authors have used a variation of a genetic algorithm to detect the peculiarity of these attacks. The experimental results concluded that a fitness level below 65% and a fitness level above 85% should be avoided as it leads to false-positive results.

Paper [2] provides a review of botnet trends, their evolution, and several ways to mitigate them. The authors have discussed the threats posed by the botnets viz. DDoS attack, Data theft, Spam, and more. The authors provide preventive measures viz. update the device to the latest operating system, avoid clicking on suspicious links, etc. The authors suggest detective measures like using the signature-based detection method, monitoring the anomalies in the network through DNS monitoring, Wireshark, and use of commercial network traffic monitors. The authors have also discussed the methods based on neural networks, logistic regression models, and machine learning with their respective accuracy of each method.

Paper [3] proposed "The Gunner System" which is a filtering approach and involves detecting DNS-based botnets. Through this approach, the authors aim to enhance the accuracy of DNS-based botnet detection. The authors conclude that the approach could successfully identify abnormal DNS queries and abnormal DNS responses.

Paper [4] uses feedforward artificial neural networks on convolutional features to detect P2P botnets using the idea of CNN. They achieved a detection accuracy of 94.7% and a false-positive rate of 2.2%. They have provided additional confidence testing for increasing the accuracy using the decision tree algorithm. After applying the confidence testing, the detection accuracy increases to 98.6% with a decreased false positive rate.

Paper [5] presents a thorough review of botnets, their lifecycle, and their types. Various botnet detection techniques like Signal Processing Technique, Entelecheia, PeerFox, Malicious Fast Flux Network Identification, Resource Sharing, and Online-Offline Detection were analyzed. The authors finally

conclude that these techniques provide high detection accuracy with negligible false positives. However, the authors also state that these techniques have a limited scope and cannot address all the problems of P2P botnets.

Paper [6] focuses on the research achievements of botnet detection that have been possible due to machine learning technology. The authors have explained the application process of machine learning in botnet detection. The security aspects of the existing solutions and the commonly used machine learning algorithms like SVM, KNN, etc. are analyzed and summarized.

Paper [7] focuses on the detection of social botnets in an online social network like Twitter. The proposed system uses a semi-supervised technique of machine learning for classification. The results of the experiments showed that they could accurately detect social bots with a low false-positive rate and an admissible detection time. One class SVM algorithm was used for the proposed system. The accuracy obtained was 99.95% with no false alarm rates.

Paper [8] makes use of the algorithms of artificial immune systems for detecting botnets in campus area networks. The proposed approach presents improvements in the BotGRABBER system. The experimental results indicated an accuracy rate of 95% with a false-positive rate of 3-5%. The model could successfully detect the HTTP, IRC, DNS, and P2P botnets using the clonal selection algorithm.

Paper [9] proposes the use of Bidirectional Long Short Term Memory based Recurrent Neural Network (BLSTM-RNN) along with Word Embedding for botnet detection. This model was also compared to a unidirectional LSTM-RNN. The experimental results showed that both the models returned high accuracy for the four attack vectors. The four attack metrics include the - Mirai, UDP, ACK, and DNS with corresponding accuracy rates 99%, 98%, 93%, and 98% respectively.

Paper [10] proposed a complete overview of existing botnet methods and along with a comparative analysis on them. The authors discuss botnet detection techniques like honeynets, signature-based detection, anomaly-based detection. They have provided a comparative study of various botnet detection approaches with their respective design details, detection strategy, and limitations.

Paper [11] proposed an algorithm for botnet detection based on the idea of statistics using random walks and then validated it on real-world data of unstructured P2P botnets. The authors imitated malware spread on large network graphs with actual botnet data. Through the experimental analysis, the authors could conclude that their algorithm yielded a high-precision rate of 90%.

Paper [12] proposes a novel method to enhance IDS performance in botnet detection. Their technique makes use of dual statistical ways - low variance filter and Pearson correlation filter, during the process of feature selection. The feature reduction stage reduced attributes to be processed by the IDS system from 77 to 15, hence lowering the computational time and giving an accuracy of 97%. The experimental results show that, despite less number of features, the accuracy does not vary largely.

Paper [13] discusses shortcomings of merging smaller datasets to form large datasets, about how they degrade the prediction performance of the machine learning models They have made use of the PCA, TrAdaBoost algorithms for the detection. The authors suggest using transfer learning rather than traditional machine learning ways to intensify the performance of systems used for Botnet identification and detection.

Paper [14] uses behavior-based techniques for TCP/HTTP botnet detection because of their variant nature. The proposed system first reduces the traffic to remove the irrelevant traffic flows and reduce the traffic workload. After this, feature extraction is performed that uses comparative features like one-way connection density, the ratio of TCP packets, etc to efficiently identify DDoS attacks. The author has used PSO and SVM algorithms for classification.

Paper [15] presents a comprehensive examination of all network connection features in botnet creation and working. The authors have inspected protocols, botnet topology, and examined a set of highly sophisticated botnets existing today. Based on the analysis, they introduce a novel classification of generalized patterns for botnet communication using modeling diagrams.

Paper [16] is inspired by IP tracking technology. The authors have proposed an inventive botnet detection method that studies data packets using graph clustering. This technique analyzes the content of the packets with the timestamps of the traffic flow. This is facilitated by refining the HEMST clustering algorithm. On analysis, results showed that the correct rate in clustering could reach a high of 97%.

Paper [17] provides a summary of various machine learning techniques and their role in botnet detection. The main aim of the paper was to clearly define the contribution of machine learning algorithms in botnet detection. The authors have discussed anomaly and DNS-based detection techniques. However, they are of the view that ML-based techniques are the most effective ones.

Paper [18] introduces an original flow-based detection system that employs supervised machine learning to identify botnet traffic. On experimental analysis, their results indicated that the system could accurately detect botnet traffic using purely flow-based analysis on traffic with supervised techniques. Along with this, they also concluded that to achieve better results, the packet flow needs to be observed for a fixed time period and fixed packet rate.

Paper [19] proposes a scalable system that is capable of detecting sneaky P2P botnets. The system recognizes all nodes that are likely to be a part of P2P communications. It then gains statistical insights to identify P2P traffic and further differentiate between P2P botnet and non-botnet P2P traffic. Substantial gauging had shown high accuracy in detecting these botnets and impressive scalability.

Paper [20] reviews the ongoing research on HTTP-based botnet detection along with its pros and cons. They also propose an approach to better the HTTP-based botnet detection, with detecting HTTP bots with random traces. They successfully showed that their proposed method led to a favorable result of reducing false alarm rates in HTTP botnet detection.

III. PROPOSED SYSTEM

In the initial age of botnet activity, various approaches to expose botnets have been suggested based on the behavior botnets depict [10]. McDermott et al. stated that contemporary techniques of botnet detection such as flow-based or signature anomaly intrusion detections have been manifested unsuccessful in fending off the growth of botnets in IoT networks. This has been mainly because of the modifications in simple code hence providing outdated attack signatures or a lack of support from protocols (Sflow, NetFlow) in networks[9]. This paper is directed towards inspecting strategies and attributes to recognize distinct botnet behavior leading to better detection of botnets. The authors' aimed to use a powerful method to classify P2P network traffic and recognize botnet activity by performing analysis on distinct network features, followed by feature selection to take out irrelevant characteristics, and then a machine learning classification algorithm to classify network traffic into legitimate and botnet traffic.

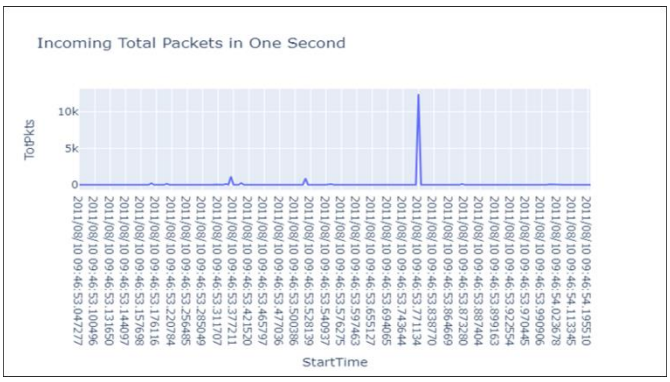

Fig. 2 Frequency and number of the influx of packets in a second

The evaluation of the system is limited to five algorithms for the comparison of experimental results namely Decision Tree, K-Nearest Neighbours, Support Vector Machines, Logistic Regression, and Naive Bayes. The accuracy of the finalized decision tree model is 99.9%. The presumption while applying machine learning approaches is that botnets create divisible patterns in network flow [18].

After comparison of the algorithms, the decision tree algorithm gave us optimum results, which is a supervised learning algorithm used in classification problems. Decision trees use a tree rendition approach to resolve problems where leaf nodes refer to a class label and features are portrayed on internal nodes. Decision trees are said to require less effort for data preparation during pre-processing.

A. System Architecture

The input to the model is a pre-labeled dataset i.e. CTU-13 dataset, that is publicly available and was captured in the CTU University. This dataset contains captures from different botnet scenarios. The model classifies the connection as legitimate or botnet. Firstly, the dataset is fed to the system as input. This input dataset contains various network attributes like the duration of the transmission, the number of bytes or packets transmitted, the protocol, the direction of the flow, and the label of the transmission. The training phase involves feeding this dataset i.e. Botnet Traffic, Normal Traffic, & Background Traffic. Next, the flows with fine-grained features are extracted and reduced to the effective features of the dataset. Feature selection was performed manually by checking how the features are related to one another by generating a correlation matrix with a heatmap between the features. Finally, the optimum classifier model (using a Decision Tree) is generated. The classifier model then classifies this traffic into two classes/labels of non-botnet and botnet connections. The output of the system generates the class of the network flow(malign/benign). The application also generates visualizations based on the input network data flow which can give deep insights and provide an easier understanding to naive users on their network activity.

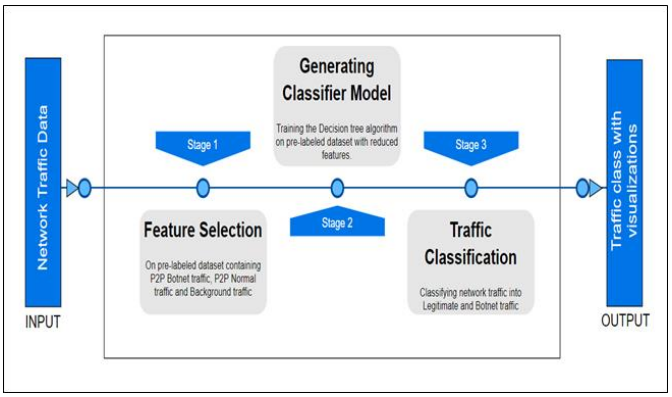

Fig. 4 System architecture

B. Algorithms Used

The classifiers first perform feature reduction and selection by selecting the most appropriate and useful features for the classification. At the first stage, feature reduction is performed on the network traffic to recognize which of the features affected the classification considerably. At the second stage, the detection model was developed, which classified the traffic into Botnet traffic and Legitimate traffic in the last phase. The proposed method gives an average accuracy of 93% considering all algorithms.

The algorithms used to provide a comparative analysis were Decision Tree, Support Vector Machine, K-Nearest Neighbour, Logistic Regression, and Naive Bayes algorithms. On the performance of these algorithms on this dataset and target prediction, Decision Tree gave exemplary results with a 99.90% accuracy among others. Also, decision tree algorithm models can handle sizable datasets, which is vital in this use case as substantial sizes of data packets flow in a network[12]. Naive Bayes gave the least accuracy of 72.25%.

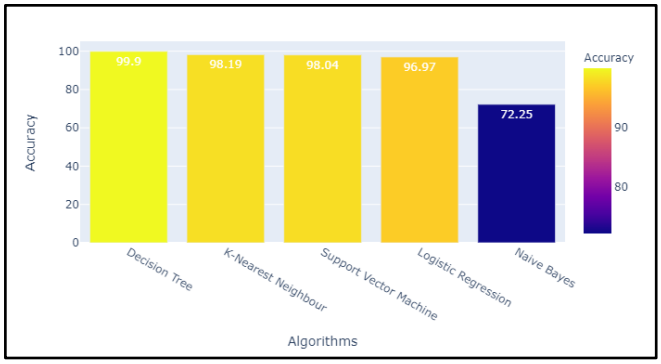

Fig. 5 Accuracies obtained by algorithms in detection model

IV. IMPLEMENTATION OF PROPOSED SYSTEM

The proposed model was trained and tested on the CTU-13 dataset. This dataset contains a large capture of mixed data containing botnet traffic, legitimate traffic, and background traffic from 13 scenarios like click fraud, fast flux, port scan, Distributed Denial of Service (DDoS), etc.

The CTU-13 dataset contains various features like the start time of the flow, duration of the flow, source and destination addresses, total packets, and bytes per transmission, with the corresponding labels of botnet activity (Normal, Background, or Botnet). The dataset was first analyzed using Wireshark and was then exported as an Excel file, to be analyzed using pandas, seaborn, matplotlib, etc.

During the analysis of features in the dataset, it was noticed that some features/attributes of network flow contribute distinctly to the nature of the traffic being botnet or non-botnet. A few of these features were the protocol of the network flow(TCP/UDP/ICMP), the ID that uniquely recognizes each transmission(since ID number ranges close to each other indicated similar nature of the network flows), etc.

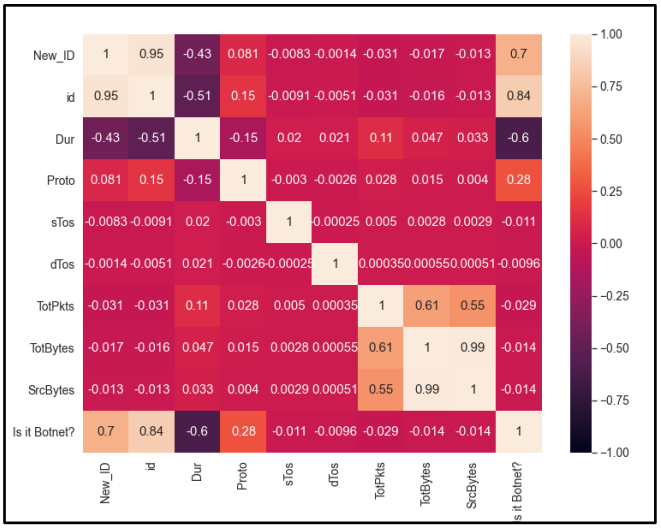

Fig. 3 Correlation of features of the dataset

The model was trained by splitting the data into training and testing sets, and five algorithms were tested on the system, where Decision Tree gave the most optimum results with minimum false rates.

The system was then deployed as a web application, using ReactJS technology. In this application, the user is first prompted to drop their network file (.binetflow) file which will contain the network flows and transactions of a fixed period. On processing this file, the system generates a dashboard for the user, wherein the risks to the user's network are mentioned by classifying the amount of botnet activity, and certain visualizations are generated via Tableau that produces statistics of the network flow.

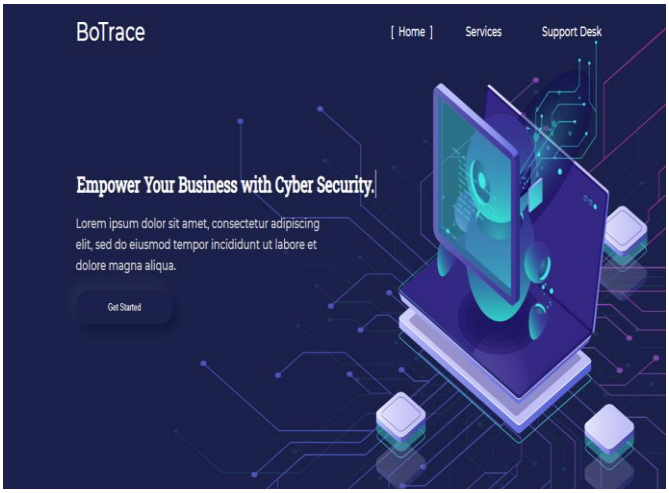

Fig. 6 Model deployed as a web application

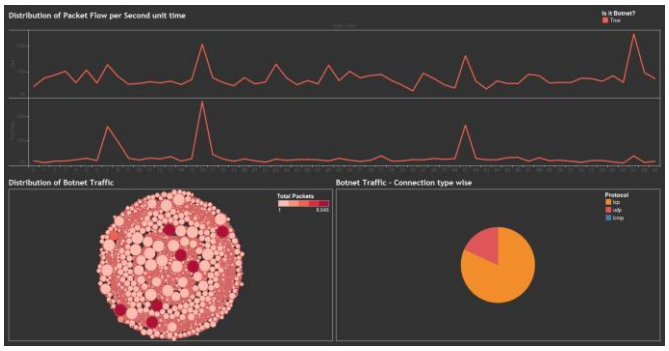

Fig. 7 Real-time statistics of network traffic generated from the system

V. RESULTS AND CONCLUSIONS

Bots use very minimal computing power to avoid disrupting or harming the device and thus alarming the user. Botnet designs continue to evolve which would make it even harder to detect.

The proposed system was successful in detecting the P2P botnet traffic. The multilayer approach overcomes the class imbalance problem of the single-layer botnet detection methods. The internet traffic is first reduced wherein only the TCP packets are filtered. Then the traffic is sent to the P2P and the Non-P2P traffic classifier where the traffic is filtered based on data packets, data stream, and session layer. The next step involves reducing the features that would marginally affect the classification. Finally, the traffic is fed to the machine learning model which then classifies it as legitimate P2P traffic and botnet P2P traffic.

The results of the proposed framework were compared with five different classifiers out of which Decision Tree gave an accuracy of 99.90% with minimum false alarm rates.

Hence, this system can be a useful tool for a network administrator or anyone who wishes to keep a check on their network flow, to track the presence of any malicious activity. In the future, the scope of this project can be further extended to detect other malware threats to a network like spyware, adware, worms, etc.

TABLE I
PERFORMANCE OF THE SYSTEM WITH RESPECT TO THE VARIOUS ALGORITHMS

Algorithm	Accuracy
Decision Tree	99.90%
K-Nearest Neighbour	98.19%
Support Vector Machine	98.04%
Logistic Regression	96.97%
Naive Bayes	72.25%

VI. REFERENCES

[1]. Y. ALEKSIEVA, H. VALCHANOV, and V. ALEKSIEVA, "An approach for host-based botnet detection system," 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), Varna, Bulgaria, 2019, pp. 1-4.

[2]. T. Lange and H. Kettani, "On Security Threats of Botnets to Cyber Systems," 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2019, pp. 176-183.

[3]. K. Alieyan, M. Anbar, A. Almomani, R. Abdullah and M. Alauthman, "Botnets Detecting Attack Based on DNS Features," 2018 International Arab Conference on Information Technology (ACIT), Werdanye, Lebanon, 2018, pp. 1-4.

[4]. S. Chen, Y. Chen and W. Tzeng, "Effective Botnet Detection Through Neural Networks on Convolutional Features," 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, 2018, pp. 372-378.

[5]. H. Dhayal and J. Kumar, "Botnet and P2P Botnet Detection Strategies: A Review," 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 1077-1082.

[6]. X. Dong, J. Hu and Y. Cui, "Overview of Botnet Detection Based on Machine Learning," 2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Huhhot, 2018, pp. 476-479.

[7]. A. Dorri, M. Abadi, and M. Dadfarnia, "SocialBotHunter: Botnet Detection in Twitter-Like Social Networking Services Using Semi-Supervised Collective Classification," 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech) , Athens, 2018, pp. 496-503.

[8]. S. Lysenko, K. Bobrovnikova and O. Savenko, "A botnet detection approach based on the clonal selection algorithm," 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), Kiev, 2018, pp. 424-428.

[9]. C. D. McDermott, F. Majdani and A. V. Petrovski, "Botnet Detection in the Internet of Things using Deep Learning Approaches," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-8.

[10]. G. Khehra and S. Sofat, "Botnet Detection Techniques: A Review," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 1319-1326.

[11]. D. Muhs, S. Haas, T. Strufe and M. Fischer, "On the Robustness of Random Walk Algorithms for the Detection of Unstructured P2P Botnets," 2018 11th International Conference on IT Security Incident Management & IT Forensics (IMF), Hamburg, 2018, pp. 3-14.

[12]. F. A. Saputra, M. F. Masputra, I. Syarif, and K. Ramli, "Botnet Detection in Network System Through Hybrid Low Variance Filter, Correlation Filter and Supervised Mining Process," 2018 Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, 2018, pp. 112-117.

[13]. B. Alothman and P. Rattadilok, "Towards using transfer learning for Botnet Detection," 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), Cambridge, 2017, pp. 281-282.

[14]. A. Kapre and B. Padmavathi, "Behavior-based botnet detection with traffic analysis and flow intervals using PSO and SVM," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 718-722.

[15]. G. Vormayr, T. Zseby and J. Fabini, "Botnet Communication Patterns," in IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2768-2796, Fourth quarter 2017.

[16]. X. Kong, Y. Chen, H. Tian, T. Wang, Y. Cai, and X. Chen, "A Novel Botnet Detection Method Based on Preprocessing Data Packet by Graph Structure Clustering," 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chengdu, 2016, pp. 42-45.

[17]. S. Miller and C. Busby-Earle, "The role of machine learning in botnet detection," 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST), Barcelona, 2016, pp. 359-364.

[18]. M. Stevanovic and J. M. Pedersen, "An efficient flow-based botnet detection using supervised machine learning," 2014 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, 2014, pp. 797-801.

[19]. J. Zhang, R. Perdisci, W. Lee, X. Luo, and U. Sarfraz, "Building a Scalable System for Stealthy P2P-Botnet Detection," in IEEE Transactions on Information Forensics and Security, vol. 9, no. 1, pp. 27-38, Jan. 2014.

[20]. M. Eslahi, H. Hashim and N. M. Tahir, "An efficient false alarm reduction approach in HTTP-based botnet detection," 2013 IEEE Symposium on Computers & Informatics (ISCI), Langkawi, 2013, pp. 201-205.