Robust Early Stage Botnet Detection using Machine Learning

Ali Muhammad, Muhammad Asad and Abdul Rehman Javed

June 7, 2020

Robust Early Stage Botnet Detection using Machine Learning

Abstract—**Among the different types of malware, botnets are rising as the most genuine risk against cybersecurity as they give a stage to criminal operations, for example, launching distributed denial of service (DDOS) attacks against targets, malware scattering, phishing, and click fraud and identity theft. Existing botnet detection techniques work only on specific botnet command and control (C&C) protocols and lack in providing early botnet detection. In this paper, we propose an approach for the early-stage detection of Botnets. Our approach first selects the optimal features using PCA (Principal Component Analysis) and Information Gain (IG) feature selection and feed these features into machine learning methods to evaluate the performance of our proposed technique. Our approach efficiently classifies normal and malicious traffic from normal ones. Our approach achieves the accuracy of 99%, TPR of 0.99%, and FPR of 0.007% in comparison with the existing approach.**

Index Terms—**Botnet, Botnet Detection, Cybersecurity, Distributed Cyberattacks, Random Forest, PCA, C&C (Command and Control Channel)**

I. Introduction

Internet security & protection has become necessary as any type of attack or breach may result in loss of sensitive data [1]. Cyberattack is an attempt to steal, destroy, or gain unauthorized access to any computer system, network, or any other device [2], [3]. There are a few kinds of Cyberattacks including data fraud, extortion, malware, phishing, spamming, Trojans, denial of services attacks, and so forth [4]. Among the different kinds of cyber attacks, botnets are developing as the most genuine risk against cybersecurity. They give a dispersed stage to numerous felonious exercises like Denial of Service Attacks on Critical Targets, Phishing, and Malware.

Botnets are systematic computer networks infected with malware called “Bots” under the remote control of a human operator called “Botmaster” often used to conduct denial of service (DDOS) attacks, spreading electronic spam, distributing pirated media, software and identity theft. Internet bots are essentially programs used to do routine sorts of tasks consequently at some specific time [4]–[6].

Botmaster uses C&C channels to create and manage botnets that are, an army of bots. C&C channel can be considered the weakest link in the botnet. Botmaster gives instructions to bots via the same C&C channel. Besides, the detection of the C&C channel will expose the C&C servers and bots in the monitored network. So, understanding and detecting C&C has great importance in countering botnets. Many existing C&C botnets use Internet Relay Chat (IRC) protocol, HTTP protocol, and peer-to-peer (P2P) protocol. IRC protocol provides a centralized C&C (command and control) mechanism. The botmaster interacts with the bots to execute commands and receive real-time responses. This IRC-based C&C mechanism has been adopted by many botnets and rated as highly successful. Some botnet uses HTTP protocol for the C&C channel. The HTTP based C&C mechanism is also centralized, but the HTTP botmaster doesn’t use mechanisms like a chat to cooperate with the bots. Instead, the bots constantly contact the C&C server to receive commands and create a botnet. The attacker (Botmaster) attacks a C&C server to gain command and control and later to issue instructions related to attack against a target client. Finally, an attack is launched by the bots on to the victim(s). Now the most recent is P2P based protocol which utilizes a peer to peer communication to perform its operation. P2P is a decentralized architecture where each node can act as a server and also a client. No such centralized coordination point is available to target. Therefore, early detection of distributed attacks can prevent the upcoming attack.

This paper makes the following contributions:

- Propose a robust technique to detect botnets at the early stage of C&C communication.
- Extract the features from the CCC dataset.
- Select optimal features using PCA (Principal Component Analysis) and Information Gain (IG) feature selection.
- Feed these features into machine learning methods to evaluate the performance of our proposed technique.
- Efficiently classify normal and malicious traffic from the normal one.

The remaining paper is organized as follows: Section II discusses the related research on botnet detection. Section III demonstrates the botnet detection approach and Section IV present the experimental analysis and results. Finally, the conclusion is summarized in Section V.

II. Prior and related work

Botnet Detection is already been in the discussion for quite long. Guofei Gu [5] presented a botminer framework that identifies different types of botnets on normal traffic. Botminer is a sophisticated tool to detect botnet. Maryam Feily [7] discusses Mining-based Detection techniques in survey paper to identify botnet C&C traffic. Authors in [8] present a multi-facet technique to identify and mitigate botnet while [9] use machine learning methods to detect the attack against intrusive network systems. They design multiple designs of network intrusion systems using publicly available datasets to train the classifiers. Authors in [10] presented bi-step subspace clustering methods. The first cluster the various botnets and then classify their types according to each host. Maryam

Feily [7] discussed techniques in a survey paper which is Mining-based Detection: This technique focuses on botnet detection by identifying botnet C&C traffic. Authors in [11] focused on the early-stage detection so that distributed cyber-attacks can be mitigated. They developed a model which perform early detection technique during C&C communication of distributed attacks and found only 10 most top features. The network traffic behaviors approach is used by Hossein Rouhani Zeidanloo [12] for the identification of peer to peer Botnets. The main contribution of this method is to identify botnets before they launch an attack by distinguishing the normal traffic and botnet traffic by analyzing the traffic against some characteristics.

For effective DDOS prevention Yang-Seo Chai [13] proposed IDDI (integrated DDOS Defense Infrastructure) which analyses and combines all techniques. All information from security or network devices of all kinds must be collected, integrated, and analyzed in IDDI. To efficiently respond against DDoS attacks, the entire process for collecting, analysis and generating defense rules must be automated. Also, IDDI is present at the center of the network, in which all the information is collected. BotHunter [14], is a warning system where certain robot behaviors are detected by Snort [15]. The involvement of bots in different events and activities can also be identified by selecting the C&C sessions that take place before the distributed attacks. Various techniques have been presented for selecting Command and Control sessions using some of the functions of the secessions. For example, The Command and Control server uses small packets to send and issue commands. Several methods for recognizing C&C sessions have also been proposed based on various protocols. In work [16], the author presented a C&C traffic detection method based on network traffic analysis based on seven characteristics (e.g. access time and standard deviation of access time). The study claims that this approach makes it easy to capture commands and control traffic on the various protocols used. According to another study by D. Ashley [17], when communication between the command and control server occurs, there is a period between the associated operation and standard deviation and in general, the HTTP bots are mostly smaller than the more common communication. However, a large number of exceptions can occur due to network interception or a command that cannot be received for any reason. Although studies are working on botnet detection, they lack fast detection of botnet attack and low TPR which we address and improve in this paper.

III. Proposed Methodology

In our proposed method, our focus was on the C&C command channel which is considered to be the primary communication phase for bots, and it also provides the benefit of applying our approach at this stage. In the C&C channel, we just select the centralized architecture which is IRC and HTTP in this work. The C&C botnet architecture is shown in Figure 1.

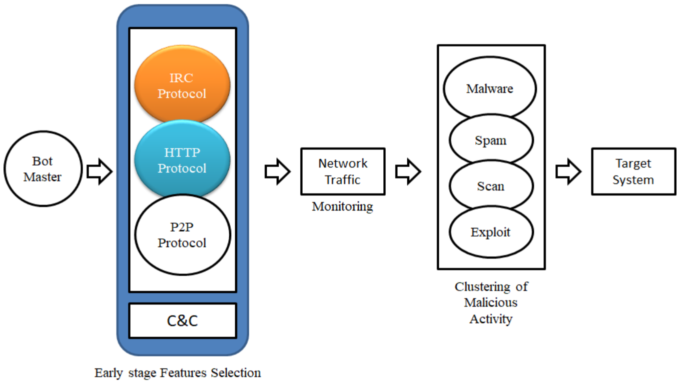

Fig. 1. Botnet C&C Architecture

After selecting the C&C channel our main focus was on features through which we can detect the attack at an early stage before the system gets compromised. Our target is to reduce a total number of features using efficient feature selection techniques while measuring the detection performance of different machine learning techniques as shown in Figure 2, feature selection has got a significant role in most machine learning-based detection techniques we have studied so far. Feature Selection can help speed up the whole process and help in the selection of the best machine learning approach that can ultimately produce good results.

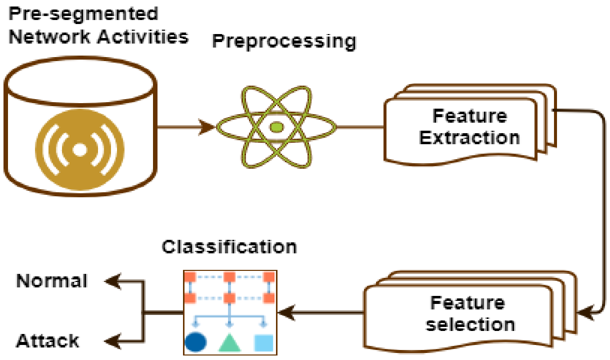

Fig. 2. Proposed Model Flow Diagram for Botnet Detection

Figure 2 represents the overall flow of the proposed model is shown in which the first stage is preprocessing. In the feature selection process, we use PCA and Information Gain to get the most important features. Features obtained are given as an input to different machine learning techniques to get results and evaluate the model. Results provide us a clear picture of which technique is useful. So we evaluated the performance of all techniques and chosen the better one among them which will be used for the detection of botnets at the starting phase of C&C.

A. DatasetOverview

In our experiment, we use the publicly available dataset CCC datasets that contain C08, C09, C10, and C13 datasets [11]. Cyber Clean Center has collected a set of CCC data. The collected data set consists of traffic packets on the port number. 6667 is used for IRC and port numbers. 80 used in the HTTP protocol. The bot must be connected to a command and control server.

B. Feature Selection

Feature selection has got a significant role in most of the machine learning-based botnet detection techniques. Feature Selection can help speed up the whole botnet detection process resulting in reduced time and space complexity and also help in selecting the best ones only that can ultimately yield good results. We apply PCA and Information Gain. Both approaches are discussed in detail below.

1) Principle Component Analysis(PCA): PCA is a well-known statistical technique used to identify different patterns and also efficient to reduce the dimension of a dataset without losing valuable information. PCA provides a set of correlated features as an output from which the best one and highly ranked features are mostly chosen [18]. Using PCA to select the feature, we selected a list of first consecutive 35 features out of 55.

2) Information Gain: Information Gain measures the relevance of a feature fi in class ci. To measure relevance, the entropy of each class ci is calculated using below Equation 1:

$$H(c_i) = -\sum_i P(c_i) \times log_2(P(c_i)) \qquad (1)$$

Where P (ci) is the probability of class ci. Entropy measures the degree of impurity and is maximum when the dataset is heterogeneous. We selected the 5 most important features using information gain which is then mixed with other features that we obtained from PCA.

C. Classifiers

After selecting these total 40 features the next step is to apply different classifier i.e. (Support Vector Machine, Logistic Regression, Multilayer Perceptron, and Random Forest) for better detection of the botnet at an early stage.

1) Random forest(RF): The random forest is an ensemble learning method for unpruned classification, regression, or other tasks that consists of building multiple decision trees. The main idea of the algorithm is to create multiple decision trees that will further yield independent results. Each tree in the forest gives an outcome about the class of a new sample data that needs to be classified. The class which gets the most votes for the object is chosen by the forest [19].

2) Support Vector Machine(SVM): Support Vector Machine (SVM) is a classification technique having an ability to deal with high dimensional data points. SVM classifier objective to find the hyperplane which has maximum margin length that separates the data samples into a distinct predefined number of classes [20].

3) Logistic Regression: Logistic Regression is used when the dependent variable can be categorized. Linear regression is unbounded, and here comes Logistic regression into the picture. The model assumes that all the predictors are linearly to log all the odds out. Mainly logistic regression selects only one feature out of highly correlated ones and assigns or reduces the coefficients of others to zero. Values of Logistic regression strictly are in range 0 or 1 [21].

4) Multilayer Perceptron: Multilayer Perceptron is an artificial neural network classifier inspired by the human brain that generates a set of outputs from a set of inputs. MLP contains 3 layers such as the input layer, hidden layers, and the output layer. Each layer contains nodes that are connected in the form of a directed graph. In the initial phase, these nodes have their own and a weight value at the input layer which is feed-forward into the hidden layer. Furthermore, these value sum of at certain points to decide based on some activation function which gives the result on the output layer [22].

IV. Experimental Analysis And Results

To evaluate the performance of the proposed technique, we measured the performance of botnet detection on an early stage upon the basis of evaluation metrics and feature evaluation. At the end of this section, we conducted experiments and presented an analysis of the experimental result.

A. Evaluation Measures

We measure the performance of botnet detection based on the following performance metrics which are accuracy, true-positive rate (TPR), and false-positive rate (FPR).

Accuracy: Accuracy is the percentage of correctly classified instances in a dataset and defined as follows

$$Accuracy = TP/(TP + TN + FP + FN) \qquad (2)$$

TPR represents number of samples that are correctly classified as normal.

$$TPR = TP/(TP + FN) \qquad (3)$$

FPR represents number of samples that are incorrectly classified as normal.

$$FPR = FP/(FP + TN) \qquad (4)$$

B. Feature Selection Evaluation

Table I shows the top 5 important features that are highly correlated selected by using PCA and IG. We evaluate these features using feature importance methods.

TABLE I
Top 5 important Features

Sr. #	Features	Description
1	Receive_ratio13	Proportion of the received packets in the size-intervals of 1200-1299 bytes
2	Receive_ratio15	Proportion of the received packets in the size-interval more than 1400bytes
3	FlagR_per	Proportion of packets with flag R of all packets in a session
4	FlagF_per	Proportion of packets with flag F of all packets in a session
5	interval_time_s_min	Minimum interval time in received package

To validate our feature selection results we cross-check our proposed method with the feature importance method. In Figure 3 and Table I, it can be seen the features selected by PCA and IG are also selected by the feature importance method.

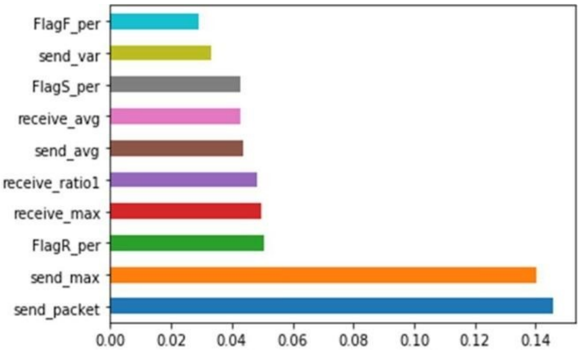

Fig. 3. Validation of the feature selected by PCA by feature importance method

C. Experiment 1

To validate the effectiveness of the proposed system, some performance tests were conducted in which we focus on feature selection and use 4 different classifiers such as (SVM, Logistic Regression, Multilayer Perceptron, and Random Forest). So we evaluated the tests based on accuracy, TPR, and FPR. Figure 4 represents several features on the x-axis and the y-axis, shows the TPR, so the below graph shows the correctly classified botnet attacks of all classifiers. In which SVM and logistic regression method show the lowest detection rate of botnets attacks while Multilayer Perceptron determines the average score, on the other hand, RF show highest detection rate of correctly classified attacks concerning other classifiers in all cases (concerning several features) and some gradual decrease is observed when the number of features was decreasing when we move from right to left.

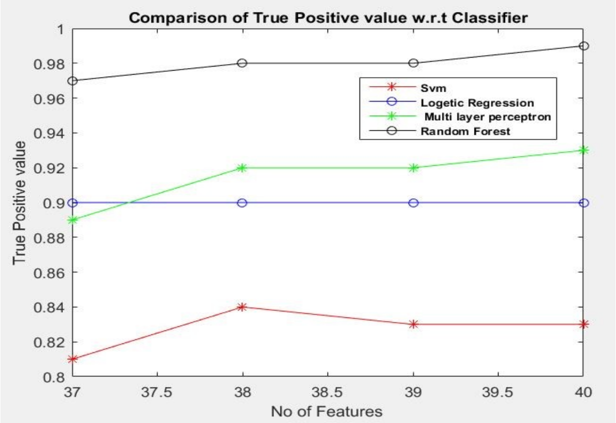

Fig. 4. Comparison of TP value w.r.t ML Techniques

D. Experiment 2

Figure 5 represents the evaluation score of the FPR which shows incorrectly classified attacks. In which SVM and Logistic Regression Classifier show the same results as a straight line in all cases (concerning several features) while Multilayer Perceptron illustrated better results than SVM and Logistic Regression depicted gradually increased in false classified classes of botnet attacks when the number of features was reducing, On the other hand, Random Forest has produced low FPR which corresponds to lower false classified attacks in all cases of features and also from all other classifiers.

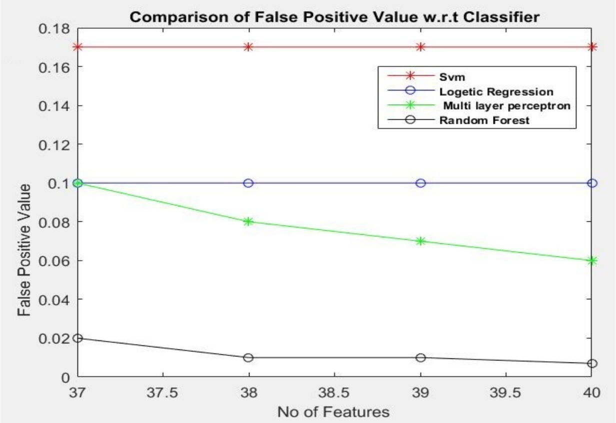

Fig. 5. Comparison of FP Value w.r.t ML techniques

The target of achieving the highest TPR and lowest False Positive was accomplished by using Random Forest and making use of 40 features. Results are better than the existing scheme model [11]. Details are shown in Figures 4 and 5 respectively. Random Forest also provides a good result for correctly classified botnet attacks if we use the least number of features which is 37 in that case, also TP value is better than the existing PCA technique [11]. So our proposed method performs well enough from the existing scheme and achieved a good result in all cases of RF classifier.

Summary of our experiments is that RF Classifier and Multilayer Perceptron efficiency were observed better but we selected Random forest which performed best in all cases as far as accuracy, TP Value, FP Value is concerned, but our main goal was to focus on important features which truly affect the detection performance. Our proposed model performed much better considering it while having 40 features and performed well in other cases as well where we had reduced the number of features from the existing method [11]. According to that, we can clearly state the model helped increase the performance of the system in case of time and space complexity. The results are shown in Table II.

TABLE II
PERFORMANCE COMPARISON OF PROPOSED METHOD WITH REDUCED NUMBER OF FEATURES. KEY: RF – RANDOM FOREST.

Techniques	Total Features	Classifiers	Accuracy	TPR	FPR
[11]	40	RF Classifier	97%	0.97	0.03
This Work	40	RF Classifier	99%	0.99	0.007
This Work	37	RF Classifier	97.8%	0.97	0.02

During initial stages when we had 40 features the proposed botnet detection system out-performed the existing model in Accuracy, TPR, and FPR. After that our focus was to reduce the number of features one by one without compromising on the accuracy and correct detection of botnet traces. We successfully managed to reduce feature till 37 that is a point, where we achieved a stable TPR and lower FPR which was still better than the existing scheme. The RF classifier provides promising results in comparison with all others. Therefore, we conclude that the proposed model performed better in the detection of botnet traces from the network traffic during the early stage of the C&C channel.

V. Conclusion

In this paper, we focused on the feature selection problem for the detection of a botnet in distributed cyber attacks. We applied our early-stage detection technique on the C&C channel which is a preliminary stage during the botnet life cycle. In our proposed approach, we not only adopted new techniques the way features are selected but also reduced the number of features. Experimental results show our approach gives better results. Our approach achieved an accuracy of 97.8%, TPR of 0.97%, and FPR of 0.02% using 37 features and accuracy of 99%, TPR of 0.99%, and FPR of 0.007% using 40 features in comparison with the existing approach. The proposed system focus on IRC and HTTP protocols during the C&C channel which are referred to as centralized architecture. In the future, further investigation is needed to discover the detection of botnet not just for centralized architecture but also to deal with decentralized architecture which is P2P based botnets.

References

[1] "Abi tyas tunggal, what is a cyber attack, upguardabi tyas tunggal, what is a cyber attack," https://www.upguard.com/blog/cyber-attack, accessed: 2020-04-21.

[2] A. R. Javed, M. O. Beg, M. Asim, T. Baker, and A. H. Al-Bayatti, "Alphalogger: detecting motion-based side-channel attack using smartphone keystrokes," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–14, 2020.

[3] A. R. Javed, M. U. Sarwar, S. Khan, C. Iwendi, M. Mittal, and N. Kumar, "Analyzing the effectiveness and contribution of each axis of tri-axial accelerometer sensor for accurate activity recognition," *Sensors*, vol. 20, no. 8, p. 2216, 2020.

[4] S. Baruah, "Botnet detection: analysis of various techniques," *International Journal of Computational Intelligence & IoT*, vol. 2, no. 2, 2019.

[5] G. Gu, R. Perdisci, J. Zhang, and W. Lee, "Botminer: Clustering analysis of network traffic for protocol-and structure-independent botnet detection," 2008.

[6] C. Iwendi, Z. Jalil, A. R. Javed, T. Reddy, R. Kaluri, G. Srivastava, and O. Jo, "Keysplitwatermark: Zero watermarking algorithm for software protection against cyber-attacks," *IEEE Access*, 2020.

[7] M. Feily, A. Shahrestani, and S. Ramadass, "A survey of botnet and botnet detection," in *2009 Third International Conference on Emerging Security Information, Systems and Technologies*. IEEE, 2009, pp. 268–273.

[8] M. Albanese, S. Jajodia, S. Venkatesan, G. Cybenko, and T. Nguyen, "Adaptive cyber defenses for botnet detection and mitigation," in *Adversarial and Uncertain Reasoning for Adaptive Cyber Defense*. Springer, 2019, pp. 156–205.

[9] G. Apruzzese, M. Colajanni, and M. Marchetti, "Evaluating the effectiveness of adversarial attacks against botnet detectors," in *2019 IEEE 18th International Symposium on Network Computing and Applications (NCA)*. IEEE, 2019, pp. 1–8.

[10] S. Araki, K. Takahashi, B. Hu, K. Kamiya, and M. Tanikawa, "Subspace clustering for interpretable botnet traffic analysis," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[11] Y. Feng, H. Akiyama, L. Lu, and K. Sakurai, "Feature selection for machine learning-based early detection of distributed cyber attacks," in *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)*. IEEE, 2018, pp. 173–180.

[12] E. B. Beigi, H. H. Jazi, N. Stakhanova, and A. A. Ghorbani, "Towards effective feature selection in machine learning-based botnet detection approaches," in *2014 IEEE Conference on Communications and Network Security*. IEEE, 2014, pp. 247–255.

[13] Y.-S. Choi, J.-T. Oh, J.-S. Jang, and J.-C. Ryou, "Integrated ddos attack defense infrastructure for effective attack prevention," in *2010 2nd International Conference on Information Technology Convergence and Services*. IEEE, 2010, pp. 1–6.

[14] G. Gu, P. A. Porras, V. Yegneswaran, M. W. Fong, and W. Lee, "Bothunter: Detecting malware infection through ids-driven dialog correlation." in *USENIX Security Symposium*, vol. 7, 2007, pp. 1–16.

[15] "Snort - rule doc search," https://www.snort.org/docs, accessed: 2020-04-21.

[16] K. Yamauchi, J. Kawamoto, and K. Sakurai, "Evaluation of machine learning techniques for c&c traffic classification," *IPSJ Journal*, vol. 56, no. 9, pp. 1745–1753, 2015.

[17] D. Ashley, "An algorithm for http bot detection," *University of Texas at Austin-Information Security Office*, 2011.

[18] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," in *2010 international conference on system science, engineering design and manufacturing informatization*, vol. 1. IEEE, 2010, pp. 27–30.

[19] M. Almseidin, M. Alzubi, S. Kovacs, and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," in *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 2017, pp. 000 277–000 282.

[20] "Support vector machine — introduction to machine learning algorithms," https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms, accessed: 2020-04-27.

[21] "Logistic regression — detailed overview," https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc?gi=32dc0c30c9b6, accessed: 2020-04-27.

[22] "Multilayer perceptron (mlp)," https://www.techopedia.com/definition/20879/multilayer-perceptron-mlp, accessed: 2020-04-27.