International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

Machine Learning Algorithms to Detect Suspicious Domain Names in Internet Security

P. Maragathavalli[1], B. Tamilarasi[2], R. Nivetha[3], S. Anjali[4]

[1,2,3,4]Department of Information Technology, Pondicherry Engineering College, Puducherry, India.
[1]marapriya@pec.edu, [2]tamilarasiboopathy2@gmail.com

Abstract: Internet security is an advanced protection solution against unknown threats in application based networks. Attackers can use a server of commands and controls to exploit communication. So malware detection of domain name is a serious issue in internet security. The signature based detection technique has been widely used as the main method of detecting malware, but with obfuscation techniques, it has failed to detect modern malware. Blacklisting is the basic method in detecting the malicious URLs. The other current Heuristic classification method is an update to the Black Category. In this process the signatures are matched and checked to find the correlation between the new URL and current malicious URL signature. Although the malignant and benign URLs can be effectively categorized by both Black-Listing and Heuristic classification. We cannot cope with the emerging methods of attacking. To overcome these issues, machine learning techniques for Malicious URL Detection is applied and use a set of URLs as training data, and depending on the statistical properties, learn a prediction function to recognize a URL as malicious or benign. The machine learning technique is to train a learning-based prediction mechanism, based on data is used for current machine-learning methods may be categorized as Supervised learning Unsupervised learning, Semi-supervised learning, Support Vector Machine. The Probabilistic Neural Networks in Machine Learning gives good performance than the primitive technologies like black listing, heuristic approach.

Keywords: Blacklisting, Internet Security, Machine Learning, Malicious Domain Names, URL Features.

1. INTRODUCTION

Domain Name System is being attacked by the attackers for the purpose of malicious activities. Many researchers have introduced so many methods for the detection of malicious domain names. However, some of them failed in their experiment. Mock apetris initiates a reputation mechanism that dynamically assigns. For day to day tasks, people are dependent on the Internet. The malicious activities cause a major loss in the economy. The URL is the main components in the Internet, that helps users to type in names of websites and resolve them to addresses of the Internet. Malware attackers try to infiltrate protective layers and defensive solutions that result in threats to a computer network and its properties. Usually attackers use a Command and Control server to exploit the interaction. So, finding malware is a significant challenge. Anti- software has long been widely used by enterprises in offering some level of security on computer networks and systems to detect and prevent malware attacks.

Moreover, many anti-malware solutions usually use static string matching methods, hashing schemes, or white listing for network communication. These solutions are too easy to overcome sophisticated malware attacks that can hide communication channels by intentionally incorporating evasive techniques to bypass most detection schemes. The problem has posed a significant threat to an enterprise's protection and it's also a huge obstacle that required to be addressed. Many of the advanced malware attackers use either a static or a dynamic technique to interact with a centralized server for a Command and Control service. Everything is fixed at a static method. For instance, the malware has both a fixed IP address and a fixed domain name permanently (i.e., it does not change its domain name during its lifetime). So, as long as this malware is recognized as a threat, a basic rule can be put in place to handle the malware issue.

The signature based detection technique has been widely used as the main method of detecting malware, but with obfuscation techniques, it has failed to detect modern malware. Obfuscation is used to hide the details, so that the true meaning is not found by others. Software vendors may use obfuscation strategies to make it impossible for the engineer to reverse the software. Signature-based detection is an anti-malware technique that detects the

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

existence of malware infection or example by matching at least one of the software's byte code patterns with the signature database of known malware programs, also known as blacklists. This identification scheme is based on the premise that patterns (also called signatures) can be used to identify malware. The most widely used technique for anti-malware systems is the signature-based detection. These features can be used in constructing the signature of the particular malware. Signature-based detection, therefore, uses the information of what is considered malicious to find out the program's maliciousness under inspection. Many researchers have tried to improve intrusion based detection methods. This technique for detecting the malicious domain names from the normal domain names. The character for determining the normal and abnormal websites should be analyzed. This detection is useful for detecting the unknown attackers. This technique has a major advantage capability of identifying the new characters or unknown characters of the domain names.

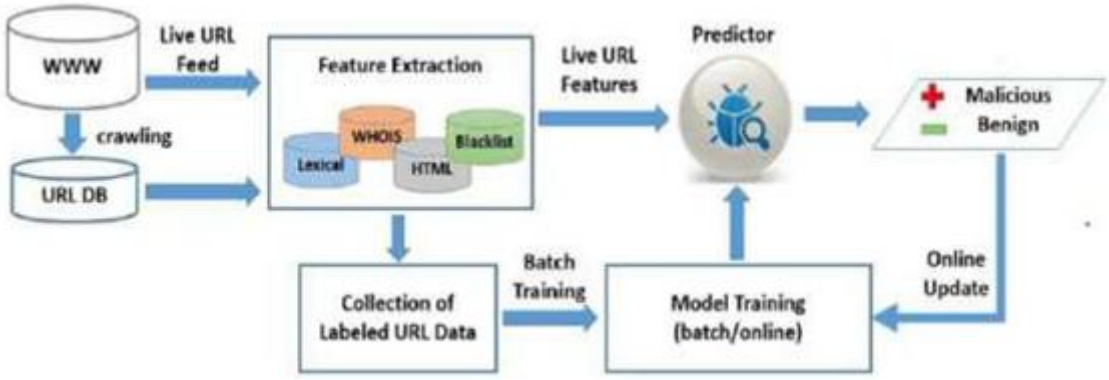

Figure 1. Detection of Malicious Domain Names

In the above Figure 1, The URL is feed into the feature extraction. It has lexical, WHOIS, HTML, blacklist. From this feature it collects the labeled URL data and it will train using batch and online learning into the predictor it will gives the feedback whether it is a malicious or not.

Infoblox DNS Firewall created by Infoblox Inc., is the most strong firewall in the domain name system to protect malicious activities. It will prevent the malicious by automatically updating the URL. It will block the connection by domain name system communication. This system will determine the infected devices in a short amount of time. The set up process is complicated ,it is the major disadvantage in the domain name system.

Domain-Flux and Fast-Flux (or IP-Flux). The former refers to the technique of having associated multiple FQDNs with one IP address. Using a Domain Generation Algorithm (DGA), a malware can generate new domain names dynamically (see DGA Taxonomy Reference), usually depending on the date and time. This method makes it very difficult to block the domain names used by a given botnets, short of having reverse engineered the DGA, since these domains have a very short lifetime. Active DNS data collection, a data collector would deliberately send DNS queries and record the corresponding Figure 1 in order to actively obtain the DNS data. The list of queried domain URLs is to be built using multiple sources, typical ones contain popular domain lists such as the Alexa Top Sites, domains appearing in different blacklist techniques, or those from authoritative server zone files. Passive DNS data collection, passive DNS data collection is accomplished by installing sensors in front of DNS servers, or by accessing DNS server logs to get specific DNS queries and replies.

Hence, passively collected DNS data is much more representative and more "revealing" in the sense of a rich set of characteristics and statistics that can be derived to detect malicious activities.

The technology intrusion detection systems (IDSs) provide security experts with powerful and multifunctional and against cyber threats. Attackers also take advanced methods in the network to perform the steal and major attacks. Advanced Persistent Threats (APTs) are considered to be the difficult and atrocious cyber threats, also known as targeted attacks. APT as the victim of the pre-selected organization or enterprise that, over time, suffers from long-term penetrative attacks and steals them. APT attacks are multiple attack methods, like phishing, social engineering, malware and backdoor program as well. The kinds of attack methods, the difficulty of malware tools and the well-organized campaign behind an APT make it difficult to detect and threaten. However, if security specialists may recognize communications with malware command and control

(C&C), which play a role as the bridge between attacks and computers, malicious operations will never remain covert and undetectable, it needs DNS as its backbone for malicious activities. Therefore, an analysis of the domain name system was proposed as an important and promising detection method on BlackHat 2014 USA. This malicious detection paper suggests a tool to identify malicious domains that can be used as a supplement to detect malicious attacks in the domain name system.

2. LITERATURE SURVEY

In this survey, review about the various machine learning techniques for malicious domain names detection and the medical solutions for the heart disease, diabetes, Lung Pancreatic Tumor characterization are also solved by using supervised and unsupervised machine learning algorithms in literature. In presenting the formal formulation of malicious URL detection as a function of machine learning, and categorizing and evaluating the contributions of literature studies that discuss various dimensions of the issue (feature representation, algorithm design ,parameters and dataset) in Table 1.

The comparison of the Detection of Malicious Domain Name with various Techniques and Classifiers using Machine Learning Algorithms.

Table 1. Comparative study on various Existing Domain Name Techniques for Internet Security

s.no	Journal Name , Year	Paper Title	Techniques/ Methodology	Parameters	Dataset	Advantages	Disadvantages
1	IEEE Transactions and Journals 2019	Machine Learning Framework for Domain Generation Algorithm (DGA)-Based Malware Detection	Hidden Markov Model (HMM)	Accuracy, Precision	Real Life Traffic Data	Better Accuracy	Detection time is high.
2	International journal of Innovative Technology and Exploring Engineering 2019	Detection of Malicious URLs using Machine Learning Technique	Support vector machine	Rank Host, Path Token count	Input-URL	Better Performance for newly generated URL	Difficult to process for large input data size
3	Expert System With application-An international Journal 12451 2019	Detection of Algorithmically Generated Malicious Domain Names using Masked N-Grams	Masked N-Gram Method	Variance & Standard Deviation	Alexa's 1M	Better Performance in Accuracy	Own dataset for experimentation may not be the correct choice
4	IEEE Transactions And Journals 2019	Lung and Pancreatic Tumor characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches	Convolutional Neural Network	Lung Nodules	Lung Image Database Consortium	Improve risk satisfaction of lung nodules	cost is high
5	Springer National Natural Science Foundation of China 2019	Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-sectional Retrospective Stud in Chinese Adults	Support Vector Machine, Random Forest	Accuracy, Precision	Nanjing Drum Hospital-people with non T2DM and with T2DM	High Stability	Experimentation done for Small dataset

6	IEEE Xplore (ICIICT) 2019	Prediction of Heart Disease using Machine Learning Algorithms	Decision Tree Algorithm, Naïve Bayes Algorithm	Data mining	Heart disease data, UCI	Better Accuracy	Difficult to extend for automation
7	ACM DIGITAL LIBRARY (ICFNDS) 2018	Malware Detection using DNS Records and Domain Name Features.	Black-listing	F-measure and Matthews correlation coefficient	Alexa and Google Search Engine	Better accuracy	Only for Limited Domains
8	ACM Computing Surveys (CSUR) 2018	A Survey on Malicious Domains Detection through DNS Data Analysis	Knowledge Based &Hybrid Approaches	Precision, Accuracy	DNS data collected in Architecture, DNS Server &ISP	Easy to study various approaches	Lack of public reference dataset
9	Springer US Neural Processing Letters 2017	Malicious Domain Name Detection Based on Extreme Machine Learning	Extreme Learning Machine (ELM)	Detection Rate, Accuracy Rate	5 DNS servers in Network Information Center	Better Accuracy and Precision	Less Efficient.
10	Arxiv:1701.07179v2[cs.LG] 2017	Malicious Url Detection using Machine Learning .A survey	Learning-First and Second Order Algorithm & Unsupervised Machine Learning	Collection Time and Processing Time	Input-URL	Better Detection rate, Provides Scalability	Difficult to extend Large dataset

Although blacklisting is commonly used, it isn't enough and can't be used for new domains like botnets, fast flux networks, drive-by-downloads, phishing, spam, and malicious advertising. Blacklisting is the protection, and to ensure more knowledge and using it to protect users, more precise security measures must be in place. The main blacklisting challenge is the low rate for the newer malicious domains that keep changing domain names that hosting services. The introduction of detailed, precise and up-to-date reputation lists of the tens of thousands of domains registered daily becomes a major challenge.

Signature-based detection and anomaly-based detections are the systems to detecting malware activities but not related to Domain Name System. Signature base technology is detects the malwares based on the existing or stored signature's database. And it is capable of using pattern matching to recognize malware in contact traffic. Moreover, a major drawback surrounds this technology; it is unable to identify new malicious domain names if the signature of the new malicious domain names does not reside in the already developed signature database.

What we want to know in the end is whether a domain is malicious or not. The word malicious, however, can be understood differently. For example, some domains may include spamming or phishing, serving communications with C&C, or simply acting as proxies for many other types of campaigns. Among many proposed methods, some are capable of recognizing a specific types of "maliciousness" while others are unable to clarify whether they adjudicate a particular domain is malicious or not. Hence, in this article, we divide techniques between those who detect particular malicious activity and those who are agnostic to malicious behavior according to the outcome of their operation.

Characterization of tumors through these tools can also allow the staging, prognosis, and fostering of personalized care planning as part of precision medicine. It is based on the machine learning techniques that are both supervised and unsupervised.

In supervised machine learning, especially through the use of a 3D conventional neural network and transference learning. A common problem in medical applications, in the unsupervised learning algorithm to resolve the limited availability of labeled training knowledge. The heart disease is a leading cause of death worldwide. The program calculates risks resulting from heart disease. The result of this method contains the percentage percentage chances of heart disease occurring. They used two key machine learning algorithms, namely Decision Tree and Naive Bayes Algorithm, which shows the best heart disease level algorithm among these two. Type 2 Diabetes mellitus (T2DM) has become a prevalent health issue, particularly in urban areas, in order to assess the risk of developing T2DM in an urban Chinese adult population by combining rules and different machine learning techniques.

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

The machine learning techniques are Multilayer Perception (MLP), AdaBoost (AD), Trees Random Forest(TRF), Support Vector Machine(SVM) and Gradient Tree Boosting (GTB), which was used to predict the risk of T2DM growth with the proposed model. The result shows that the combination of machine learning models could provide an effective model for the prediction of T2DM risks.

A machine learning technique is used for the study and classification of domain names. Neural network use, however, is rarely seen in previous studies because the slow learning speed limits efficiency in detection issues. Huang et al's proposed Extreme Learning Machine (ELM) is a new learning scheme for Single-- Feed for Neural Networks (SLFNs) with fast learning speed. One of the most common supervised methods of learning is the support vector machine.

Using a maximum margin learning method, it uses the systemic risk minimization principle, which can basically be viewed as an instance of the regularized loss minimization structure. Additionally, using kernels, SVM can learn nonlinear classifiers; SVM is likely one of the most widely used classifiers for malicious URL detection. Naive Bayes is a classification generative model which is "naive" in the sense that his model assumes that all of X's features are independent of each other. Decision trees is one of the most common inductive inference methods and have a significant advantage of their highly interpretable decision tree classification models which can also be translated into a human readability rule. Decision trees were used to identify malicious URLs / web. Extreme Learning Machines (ELM) for classifying phishing websites use ELM by integrating hybrid features with a spherical classification method that allows for the adaptation of batch learning models to a large number of cases. Set malicious domains as positive instances, and benign domains as negative to assess the performance of our detection system.

3. FEATURE EXTRACTION

In Preprocessing stage, input dataset / database in the format of excel / csv file type, which is imported by using MATLAB. After data import, separate the data like numbers & string/cell structure (i.e. predicted class label) and also eliminate the “Null” & “NaN – Not an number” from the data. When data is applied to wavelet transform, it decomposes into 4 parts – approximation, horizontal, vertical, diagonal coefficients.

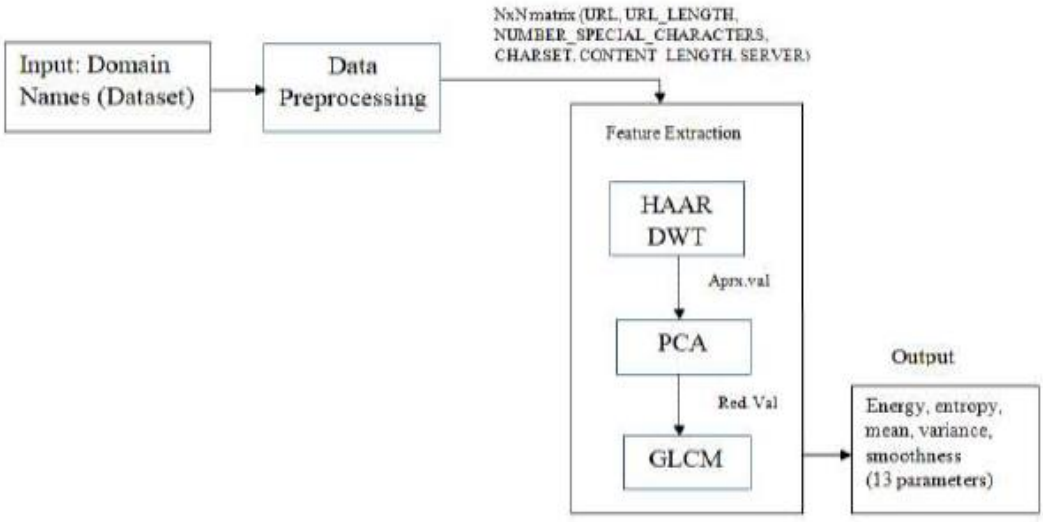

Figure 2. Feature Extraction of Malicious Domains

In that approximation coefficients will have a complete information about the data. So that approximation coefficients are taken for the next level. After that, the output of approximation coefficients is applied to PCA. In order to handle curse of dimensionality and avoid issues like over-fitting in high dimensional space, and it also used to reduce the no of variables in the data by extracting important one from a large pool, like in DWT, it’s dimension greater than 2 means, PCA will reduce the dimension level. So that PCA is used. The proposed Malicious Domain Name Detection scenario is shown in Figure 2. And then PCA output is applied to GLCM, from the gathered features like energy, homogeneity, correlation, skewness, standard deviation, smoothness etc. Like that calculate & obtain the 12 to 13 parameters. These parameters will be used for classification purpose.

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

Input and Output Screenshots

The Kaggle Dataset of the Malicious Domain Names properties as URL Length, Server details, DIST_Remote_TCP_port, Source_App_Bytes, Remote_App_Bytes used for the detection of malicious domain names using the machine learning algorithm shown in figure 3.

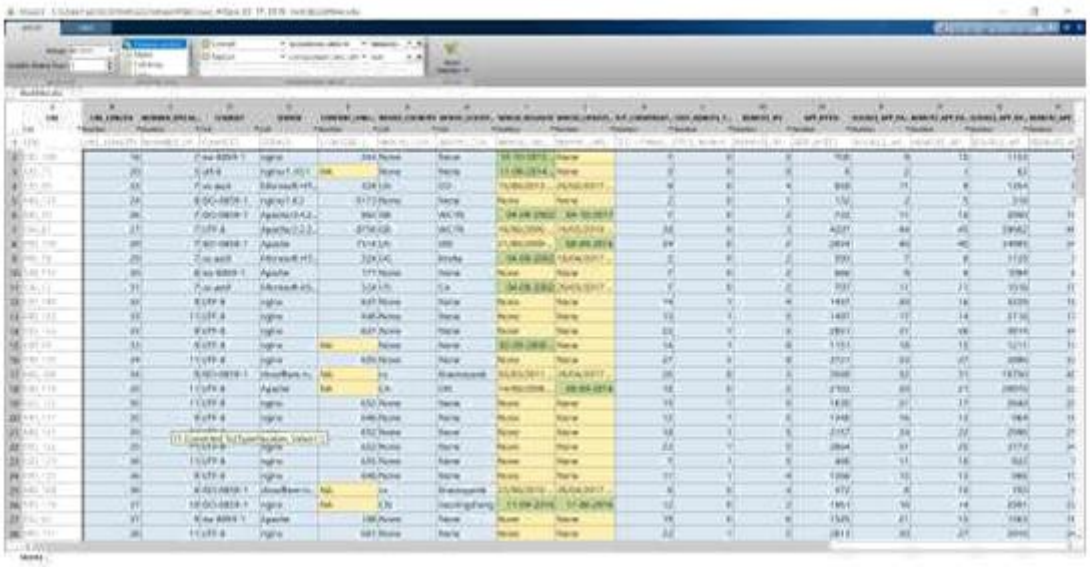

Figure 3. Input Domain Names Properties

The results of the Feature Extraction are shown in Figure 4. Includes the parameters of Contrast, Correlation, Energy, Homogeneity, Mean, Standard Deviation, Entropy, RMS, Variance, Smoothness, Kurtosis, and Skewness for each Domain Names in the Dataset.

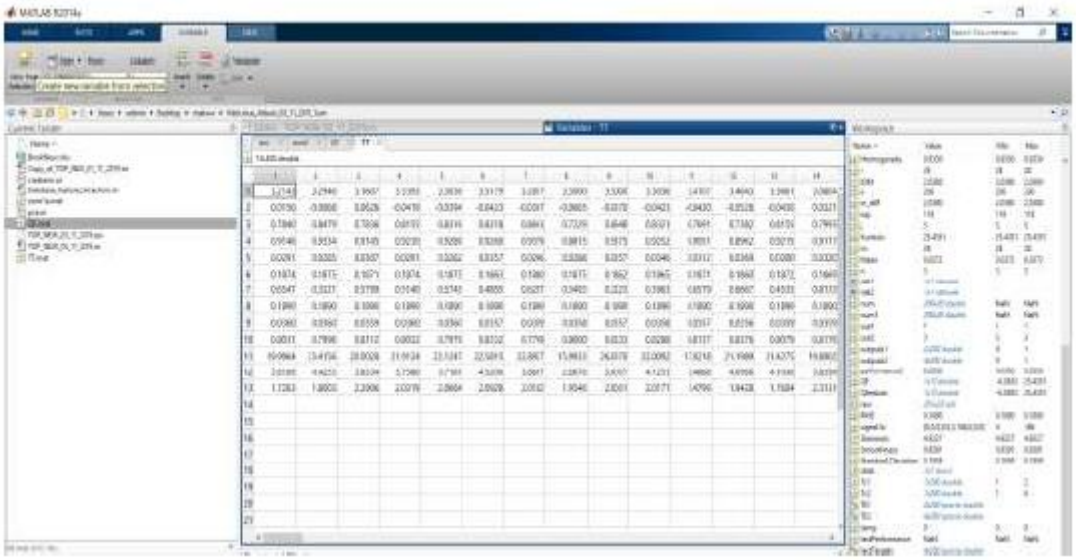

Figure 4. Feature Extraction Results

Malicious Detection and its Attacks

After the feature extraction, all features are extracted from each data and combined them into a single matrix. While on preprocessing Predicted label/class are already separated, based on the class we represent the class to Probabilistic neural networks in number format like 1 & 2 i.e. “1” for Not Malicious & “2” for Malicious.

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

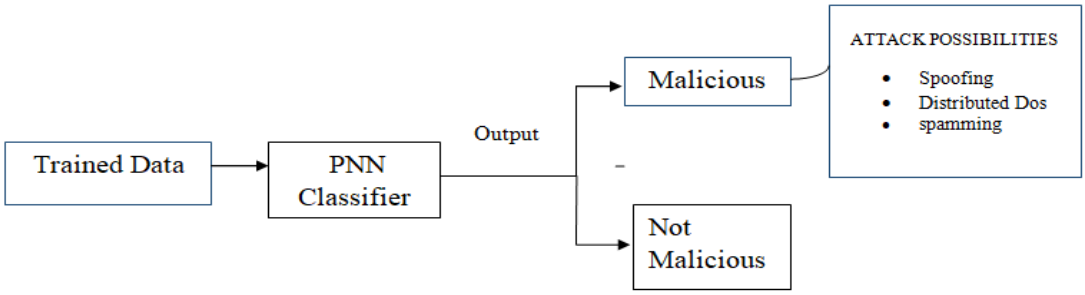

Figure 5. Malicious Detection and its Attacks

Based on that feature extracted data & predicted class/ label is applied to Probabilistic neural networks, will analyze the data & predicted the corresponding result/class whether it is malicious or not based on feature extract data. After the detection of malicious domain names by using the PNN classifier and For instance some domain names may be involved in DNS attacks, which are capable to recognize specifically is shown in Fig 5. Malicious URLs have been widely used to mount various DNS attacks including spoofing(1), Spamming(2), Distributed Denial of Service(3).

Architectural Diagram

In this Architectural Diagram Figure 6, it defines the input layer, hidden layer and output layer of the Probabilistic neural networks and its process of detecting the malicious domain names and in Figure 7 shows the Detection of Attacks which is induced by the malicious domains.

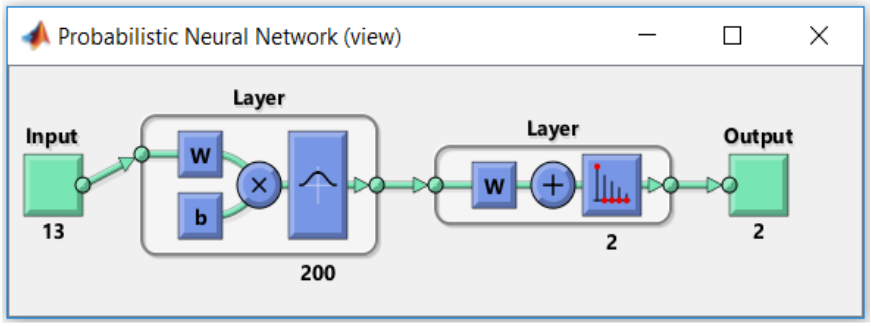

Figure 6. Architectural diagram for Malicious Detection using PNN

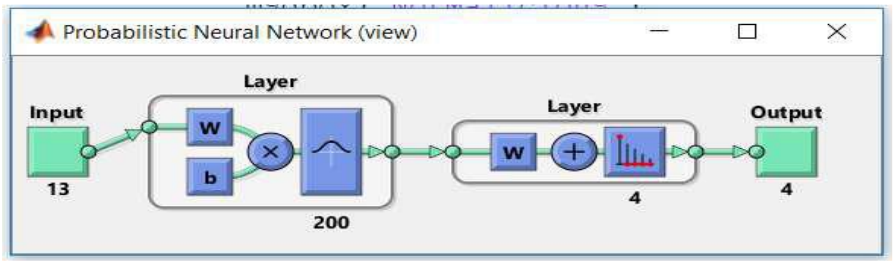

Figure 7. Architectural diagram for Attacks Detection using m-PNN

4. RESULTS OF THE MALICIOUS DETECTION

The detected malicious domain name shown in figure 8. And the specific Attack which is induced by the malicious Domains.

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

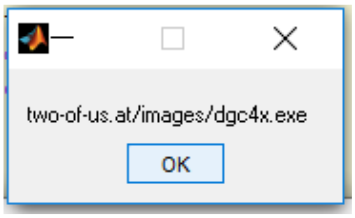

Figure 8. Detection of Malicious Domain Name

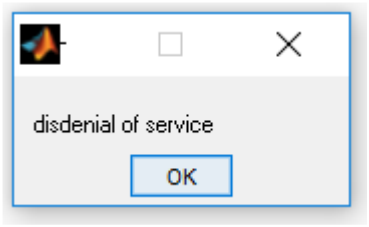

Figure 9. Detection of Malicious Attack

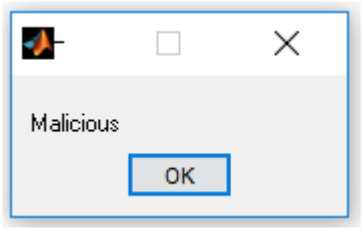

Figure 10. Output of Testing Domain Name (Malicious or Not Malicious)

Performance Analysis

The confusion matrix in figure 11 shows the accuracy of 98.5% malicious domain names detection which defines the target class and output class of PNN and in Figure 12 shows the accuracy of detecting the Attacks with 93%.

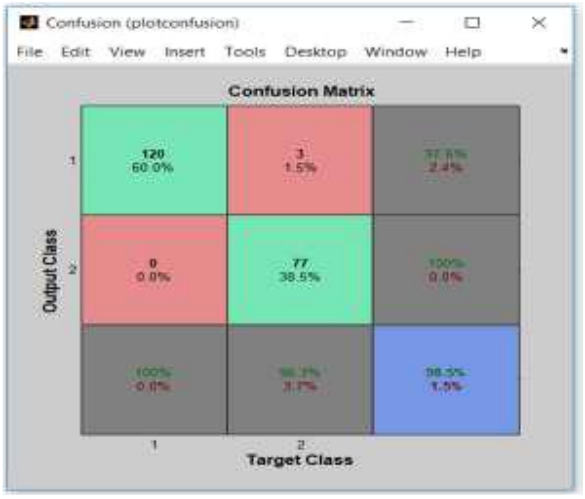

Figure 11. Accuracy for Malicious Detection

Figure 12. Accuracy for Attack Detection

The overall detection time of the Domain Names and its Attacks is shown in seconds as elapsed time in Figure 13.

Figure 13. Detection time for Malicious Detection

5. CONCLUSION

Malicious Domain Name Detection plays a critical role for many cyber security applications and clearly machine learning approaches are a promising direction. In this Project we have considered the problem of malicious Domains in the internet. Specifically, the feature sets and an approach for classifying the given feature sets of the Malicious Domain Names detection. When traditional method fall short in detecting the new malicious domain names and not support for large datasets, the proposed method can be augmented with it and provides the improved results.

REFERENCES

[1] Yi Li,KaiqiXiong, Tommy Chin and Chengbin Hu, "A Machine Learning Framework for Domain Generation Algorithm(DGA)-Based Malware Detection", IEEE Transaction and Journals, DOI:10.1109/ACCESS.2019.28915 88,Jan 2019, pp.1-16.

[2] Ashwini Mujumdar, GayatriMasiwal and Dr.B.B.Meshram, "Analysis of Signature-Based and Behavior –Based Anti-Malware Approaches", international journal of advanced research in computer engineering and technology(IJARCET) vol. 2,Jun 2013,pp. 121-132.

[3] Xiao-luXiong, Rong-xin Zhang, Yan Bi, Wei-hong Zhou, Yun Yu, Da-long Zhu,"Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross –sectional Retrospective Study in Chinese Adults", Vol.39 Issue:4, DOI: https://doi.org/10.1007/s11596-019-2077-4, Aug 2019, pp. 582-588.

[4] DaHuang KaiXu JianPei, "Malicious URL detection by dynamically mining patterns without pre-defined elements", Received: 22 Jun 2012 / Revised: 4 Mar 2013 / Accepted: 24 Jul 2013, Springer Science, 2013,DOI 10.1007/s11280-013-0250-4,pp. 1-23.

[5] Sandeep Yadav, Student Member, IEEE, Ashwath Kumar Krishna Reddy, A.L. NarasimhaReddy , Fellow, IEEE Member, ACM and Supranamaya Ranjan, Member, IEEE ,"Detecting Algorithmically Generated Domain-Flux Attacks With DNS Traffic Analysis ",IEEE/ACM Transactions on Networking,Vol.20, Issue.5,Oct 2012,pp. 1-12.

[6] Schiavoni.S, Maggi.F, Cavallaro.L and Zanero.S, "Phoenix: DGA-based botnet tracking and intelligence," in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, 2014, pp. 192–211.

[7] YuryZhauniarovich, Issa Khalil, Ting Yu and MarcDacier, "A Survey on Malicious Domains Detection through DNS Data Analysis", ACM Computing Surveys, Vol. 1, Feb 2018,pp.1-28.

[8] Yong Shi, Gong Chen and Juntao Li, "Malicious Domain Name Detection Based on Extreme Machine Learning", Springer US Neural Processing Letters, DOI: 10.1007/s11063-017-9666-7, Jul 2017,pp.1-10.

[9] Sarfaraz Hussein, PujanKandel, Candice W.Bolan, Michael B. Wallace and UlasBagci, "Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches", IEEE Transacions, DOI:10.1109/TML.2019.2894349, Vol.38, Iss. 8, Aug 2019,pp. 1-10.

[10] Immadisetti Naga VenkataDurga Naveen, Manamohaa K and RohitVerma, "Detection of Malicious URL'S using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering, ISSN:2278-3075, Vol.8, Iss. 4S2, Mar 2019,pp.1-4.

[11] Jose Selvi, Ricardo J.Rodriguez and Emilio Soria "Detection of Algorithmically Generated Malicious Domain Names using Masked N-Grams",DOI: https://doi.org/10.1016/j.eswa. 2019.01.050,Reference: Expert Systems with Applications 12451, Jan 2019,pp.1-24.

[12] Hong Zhao, Zhaobin Chang, Guangbin Bao and Xiangyan Zeng, "Malicious Domain Names Detection Algorithm based on N-Gram", Hindawi Journel of Computer Networks and Communications, Vol.56, 2019, Article ID4612474, DOI: https://doi.org /10.1155/2019/4612474, pp. 1-8.

[13] Khulood Al Messabi, MontherAldwairi, Ayesha Al Yousif, Anoud Thoban and FatnaBelqasmi, "Malware Detection using DNS Records and Domain Name Features", in Proceedings of the 20th International Conference on Future Networks and Distributed Systems",ACM,26-27th Jun-2018, New York, USA, ISBN 978-1-4503-6428,
DOI: 10.1145/323/053.3231082,pp. 1-6.

[14] Ryan R.Curtin, Andrew B. Gardner and SlawomirGrezonkowshi, "Detecting DGA domains with recurrent neural networks and side information", arxiv:1810.02023v1[Cs.Cr] ,Oct 2018, pp. 1-11.

[15] Doyen Sahoo, Chenghao Liu and Steven C.H.Hoi, "Malicious URL detection using Machine Learning: A Survey", arXiv:1701.07179, Mar 2017, pp.1-18.

[16] Santhana Krishnan.J and Geetha.S, "Prediction of Heart Disease Using Machine Learning Algorithms", in Proceedings of the 2019 1st International Conference on Innovations in Information and Communication Technology", IEEE Xplore,25-26th Apr -2019,Chennai , DOI:10.119/ ICIICT1 .2019.8741465, Jun 2019, pp. 1-4.

AUTHOR BIOGRAPHY

Dr.P.Maragathavalli

She received her B.E. degree in CSE from Bharathidasan University, M.Tech. and Ph.D. degree in CSE from Pondicherry University. She joined Pondicherry Engineering College in 2006 and currently working as Assistant Professor in the Department of Information Technology. She has published several research papers in various referred journals and international conferences. Her area of Interest includes Security Testing, optimization Techniques, Genetic Algorithms and Information Security. She is a Life member of ISTE.

B.Tamilarasi

She is pursuing her B.Tech degree in the Department of Information Technology, Pondicherry Engineering College, Pondicherry.

R.Nivetha

She is pursuing her B.Tech degree in the Department of Information Technology, Pondicherry Engineering College, Pondicherry.

S.Anjali

She is pursuing her B.Tech degree in the Department of Information Technology, Pondicherry Engineering College, Pondicherry.