IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Browser Based Malicious Domain Detection through Extreme Learning Machine

Dr. G. Nanthakumar[1], Arunprakash. R[2], Balamurugan. G[3], Karthick. S[4]
Professor, Department of Computer Science Engineering[1]
Final Year Students, Department of Computer Science Engineering[2,3,4]
Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Tiruvarur, Tamil Nadu, India

Abstract: *The main aim of the project is to detect the malicious domains in the internet through Machine learning approaches and is helpful in preventing the cybercrime. For that Extreme Learning Machine(ELM) approach is used to find the malicious domains in the web. The Terms used for the project are Browser application, extreme learning machine, feature selection, malicious domain detection, machine learning, Real-time training. This approach will be helpful in reducing the cybercrime such as harmful websites, fake websites that stores the user information, malicious attack website domains.*

Keywords: DNS- Domain Name System, URL- Uniform Resource Locator, ELM –Extreme learning Machine .

I. INTRODUCTION

In recent years, the usage of bad domains has increased dramatically in cybercrime. An adversary may, for example, use malicious domains created by him or her to run command and control (C&C) servers or set up phishing sites. Preparing a deny list of these malicious domains is a common countermeasure against them. Nonetheless, because domain generation algorithms (DGAs), which produce new domains automatically, are often used, new domains appear on a regular basis. As a result, deny-list-based countermeasures are insufficient, and a framework that covers even unknown domains is essential. Based on the foregoing, machine learning-based domain discovery has gotten a lot of attention in recent years. Mostafa M. Fouda, on the other hand, approved it for publishing. To the best of our ability Although there is VT4Browsers1, a Google Chrome add-on that automatically detects dangerous sites on a browser, it is only a refuse list-based technique. Existing browser-based apps, for example, may be rendered ineffective against unknown sites. Existing (browser-independent) harmful domain identification approaches based on machine learning, on the other hand, frequently employ a complicated and large-scale architecture. Despite giving good inference accuracy, the training duration increases in proportion to the design complexity. In other words, they're difficult to implement in a browser context where a user may use them in real-time. , a machine learning-based domain detection tool for a browser, which is the user's closest interface, has never been proposed to the community before. The associate editor in charge of the project.

1.1 Machine Learning

A machine learning-based domain detection algorithm makes inferences to identify whether or not the specified domains are dangerous. Informally, the objective model is created when a machine learning model learns domain attributes and labels that classify domains as benign or dangerous. Following that, the model receives information from a target domain as inputs and infers the domain's label. In recent years, neural networks have become a popular tool for domain discovery.

1.2 Extreme Learning Machine

The model explains how the extreme learning machine (ELM), a fast machine learning algorithm for training single hidden layer feedforward neural networks, came to be. ELM, in general, can achieve efficient generalisation performance and, according to the research, can compute a global optimization that is equivalent to or better than support vector machine (SVM) ELM is also considerably faster and easier to install than most cutting-edge machine learning techniques. As a result, ELM has a wide range of applications in fields such as life science and computer

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

vision. Meanwhile, to the best of our knowledge, ELM has never been used in the context of malicious domain detection other than for process.Because concept drift is necessary for hostile domain detection, as mentioned in Section I, the real-time training covered in this research is a distinct problem than the previous works. In the context of supervised learning, the original ELM is utilised for both classification and regression problems. Although this study focuses on the original ELM, there are various modifications of ELM that can be used in the real world to cope with unbalanced data and errors. To deal with imbalanced classification data, Xiao et al. suggested a class-specific cost regulation extreme learning machine (CCR-ELM), which introduced a class-specific regulation cost into the classification. To deal with the prediction, Zhang et al. presented residual compensation ELM (RC-ELM).

1.3 Preparing the Dataset

This dataset was created in order to detect malicious domain. It contains

- Previously detected malicious domains.
- Collection of malicious Domain data.

II. LITERATURE REVIEW

Using seven distinct machine learning techniques, such as decision tree, Adaboost, K-Star, kNN (n = 3), Random Forest, and SMO, a phishing detection system was created. Various number/type features based on NLP functions, word vectors, and mixed functions, as well as Naive Bayes. Create useful features to improve the detection system's accuracy. Making lists is a crucial task. All of the algorithms have been put to the test, and the results have been compared. They have created a new hybrid model that combines NLP (natural language processing-based characteristics) with word vector. Language independence, real-time execution, detection of new websites, independence from third parties, use of feature-rich classifiers, and so on are all advantages. Two modules extract functional representations of URLs in parallel. The character level CNN module is the first. Another approach is to use an attention-based hierarchical RNN module to detect phishing URLs. They discovered that Random forest classifiers have the highest AUC and Bag-of-words trained SVM classifiers have the highest accuracy. The effectiveness of the phishing URL detection strategy has been demonstrated in this paper using a deep learning-based algorithm. Experiment with ablation Word-level temporal feature representation extracted from character-level spatial feature representation . From character-level CNNs to attention-based URLs, there's a lot to learn. The performance of hierarchical RNN modules is improved. This method has a good generalisation ability.

This model addresses the issue of detecting phishing URLs in the absence of strong supervision, requiring a lower amount of labelled data to begin the learning process.

III. PROPOSED SYSTEM

Fig.1 represents the working process of our system i.e., the methodology of our work. Where the algorithm goal is to find out the current URL website which is browsing is good or bad. By using the machine learning algorithm, we are going to classify which type of URL is a suspicious URL.

For that detection process, the model use for the processing and extract the feature and through the data set we used for the processing and a huge dataset which is classified over, The data set is split for training and testing data are in ratio 75/25. The data set is trained and after the train and accuracy obtain later that passing the test dataset and predict the result for the test dataset and final we use the new data for the prediction of unknown data and the result is obtained. Classification error which is reduced .

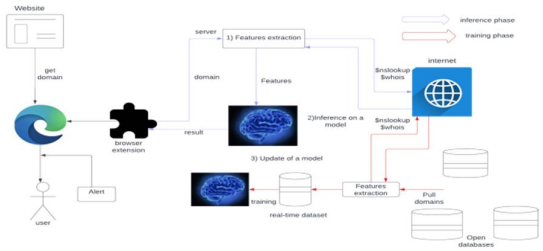


Fig.1 System Architecture of Proposed model

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

IV. MODULE DESCRIPTION

Proposed System Modules

- Feature Extraction
- Classification
- Real Time Training
- Malicious Domain Detection

4.1 Feature Extraction

A Domain itself has few features generally. Consequently, Extracting features based on DNS records for a domain is expected since the information itself is insufficient. However , in real-time domain detection through an add-on , the throughput of feature extraction should be considered because the above feature extraction is a time-consuming task in general.DNS based features represent information obtained from DNS records of their corresponding domains and discuss the difference of DNS records between malicious domains and benign domains.

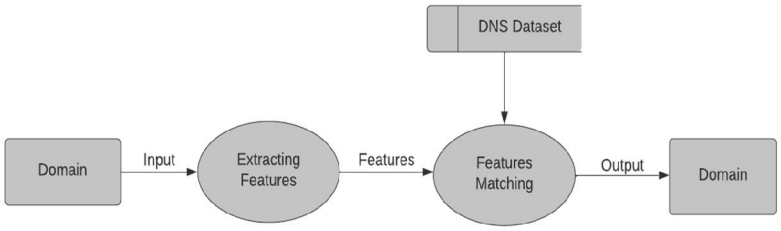


Fig. 2 Features Extraction

4.2 Classification

Classification can be useful to classify the domain from the trained dataset. The given domain is compared with the trained Dataset and then it classifies the domain which is malicious. With the help of above module , we can predict the domain is malicious or not.

A classification algorithm, in general, is a function that weighs the input features so that the output separates one class into positive values and the other into negative values.

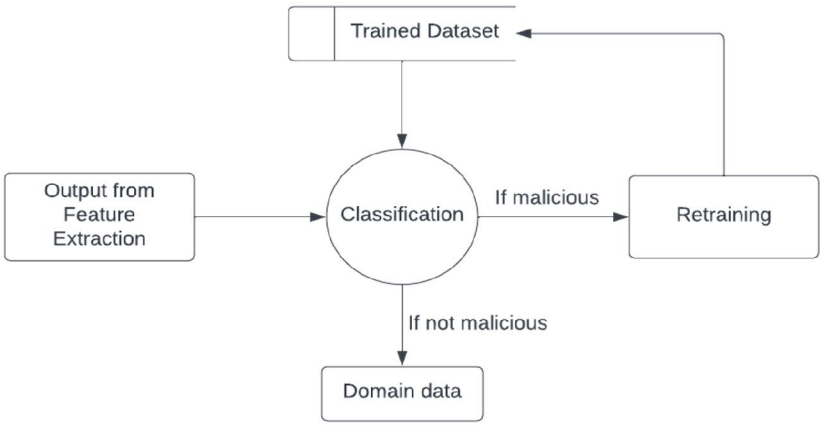


Fig. 3 Classification

4.3 Real Time Training

From Kaggle, the data are stored in the database. Based on that data , the input would be compared to show the output.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

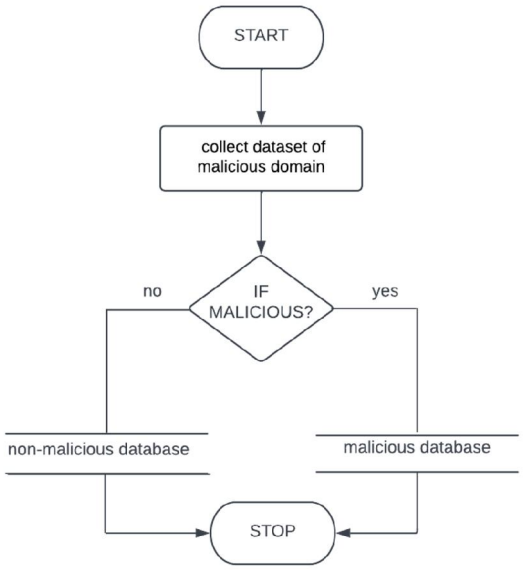


Fig. 4 Real Time Training

4.4 Malicious Domain Detection

ELM has been proposed for training single hidden layer feedforward neural networks. The training process of neural networks is roughly divided into three kinds of layers, i.e., an input layer, one or more hidden layers, and an output layer. The performance of neural networks is commonly improved by increasing the number of hidden layers and neurons in each layer.

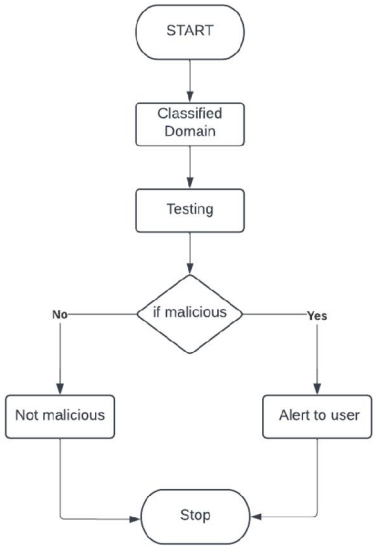


Fig. 5 Malicious Domain Detection

V. IMPLEMENTATION

The result shows that the website is malicious or not. If the domain is malicious , then it shows the user as 'Unsafe' , Otherwise it shows 'Safe'. It compares the user domain input with the database that is already retrained.

The large amount of data is trained in the database . Based on that url data it will shows the output.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Example 1:

Input

https://www.google.com

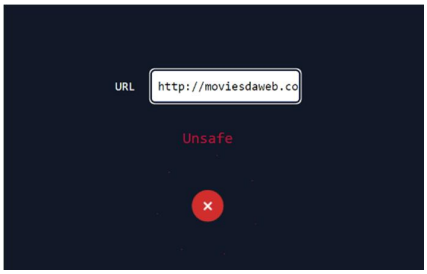
Output

Example 2:

Input

https://www.eqbqcguiwcymao.com

Output



VI. CONCLUSION

The Model presented is a browser-based application for malicious domain detection by leveraging extreme learning machine (ELM).In the real-time training, the retrained model could continuously detect unseen malicious domains while the accuracy of the normal model decreases because of missing a concept drift of malicious domains.
This project will be very useful in terms of detecting the malicious domain in the web and it reduces the cybercrime.

REFERENCES

[1]. Y. Zhauniarovich, I. Khalil, T. Yu, and M. Dacier, ‘‘A survey on malicious domains detection through DNS data analysis,’’ ACM Comput. Surv., vol. 51, no. 4, pp. 1–36, Sep. 2018.
[2]. J. Woodbridge, H. S. Anderson, A. Ahuja, and D. Grant, ‘‘Predicting domain generation algorithms with long short-term memory networks,’’ 2016, arXiv:1611.00791.
[3]. S. Vosoughi, P. Vijayaraghavan, and D. Roy, ‘‘Tweet2Vec: Learning tweet embeddings using character-level CNN-LSTM encoder-decoder,’’ in Proc. 39th Int. ACM SIGIR Conf., 2016, pp. 1041–1044.
[4]. D. S. Berman, ‘‘DGA CapsNet: 1D application of capsule networks to DGA detection,’’ Information, vol. 10, no. 5, p. 157, Apr. 2019.
[5]. Y. Qiao, B. Zhang, W. Zhang, A. K. Sangaiah, and H. Wu, ‘‘DGA domain name classification method based on long short-term memory with attention mechanism,’’ Appl. Sci., vol. 9, no. 20, p. 4205, Oct. 2019.
[6]. L. Yang, G. Liu, Y. Dai, J. Wang, and J. Zhai, ‘‘Detecting stealthy domain generation algorithms using heterogeneous deep neural network framework,’’ IEEE Access, vol. 8, pp. 82876–82889, 2020.
[7]. F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, and L. Cavallaro, ‘‘Tesseract: Eliminating experimental bias in malware classification across space and time,’’ in Proc. USENIX Secur., 2019, pp. 729–746.
[8]. Y. Shi, G. Chen, and J. Li, ‘‘Malicious domain name detection based on extreme machine learning,’’ Neural Process. Lett., vol. 48, no. 3, pp. 1347–1357, Dec. 2018.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

[9]. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, ''Extreme learning machine: Theory and applications,'' Neurocomputing, vol. 70, nos. 1–3, pp. 489–501, Dec. 2006.

[10]. W. Cao, X. Wang, Z. Ming, and J. Gao, ''A review on neural networks with random weights,'' Neurocomputing, vol. 275, pp. 278–287, Jan. 2018.

[11]. C.-J. Chien, N. Yanai, and S. Okamura. (2021). Design of Malicious Domain Detection Dataset for Network Security. [Online]. Available: http://www-infosec.ist.osaka-u.ac.jp/~yanai/dataset.