Turkish Journal of Computer and Mathematics Education Vol.12 No. 7 (2021), 2648-2653

Research Article

Phishing Website Detection Using Novel Features And Machine Learning Approach

[1]S.T.Deepa, [2]Dr.K.K. Thanammal, [3]Dr.S.S. Sujatha

[1]MS University,,
Research Scholar, Department of Computer Applications, S.T. Hindu College,
Abishakapatti, Tirunelveli-627012, Tamilnadu, India.
deepapramod1972@gmail.com
[2]MS University,
Associate Professor, Department of Computer Science and Applications, S.T. Hindu College,
Abishakapatti, Tirunelveli-627012, Tamilnadu, India.
[3]MS University,
Associate Professor, Department of Computer Science and Applications, S.T. Hindu College
Abishakapatti, Tirunelveli-627012, Tamilnadu, India.

Abstract: Phishing, a type of digital assault adversely affects individuals where the client is coordinated to counterfeit sites and tricked to uncover their delicate and individual data which incorporates passwords of records, bank subtleties, ATM pin-card subtleties and so forth Subsequently shielding touchy data from malwares or web real; irregular woods phishing is troublesome. Inferable from the restrictions of existing advances in identifying a phishing site, anticipating that the users should notice and can decide if a URL is phishing or genuine would be unreasonable, wasteful and mistaken. Consequently, in tending to these difficulties, a robotized approach should be considered for phishing site recognition. This research aims at detecting phishing website using novel features and machine learning algorithm. The input URL websites are first feature extracted using Convolutional auto encoder. Then those features are sent to deep neural network classifier for better classification of Phishing and legitimate URL's. The system is tested for its accuracy and detection rate. It shows that the implemented system is best in detecting phishing websites with 89% accuracy.

Keywords: Phishing, novel features, Convolutional auto encoder, Deep Neural Network, legitimate URL.

1. Introduction

In this mechanical time, the Internet has advanced toward become an unavoidable piece of our lives. It prompts numerous advantageous encounters in our lives in regards to correspondence, amusement, instruction, shopping, etc. As we progress into online life, lawbreakers see the Internet as a chance to move their actual violations into a virtual climate. The Internet gives accommodation in different perspectives as well as has its disadvantages, for instance, the obscurity that the Internet gives to its clients. As of now, numerous sorts of wrongdoings have been directed on the web [1]. [2] However, the Internet is likewise portrayed by some unavoidable security issues, for example, phishing, vindictive programming and protection divulgence, which have just carried genuine dangers to the economy of clients. Subsequently, the fundamental focal point of our examination is phishing. Phishing has gotten one of the greatest security dangers in the Internet.

Phishing sites come in two structures: parody and created. Satire locales mirror existing, by and large notable sites to participate in data fraud or malware spread. Created locales are anecdotal sites intended to direct social designing, fake internet promoting, or dark cap website improvement based assaults for financial increases or malware spreads [3][4]. The two classes of phishing sites have genuine ramifications for Internet clients and associations, for example, harming brand value and expanding client agitate rates. Created sites additionally often show up in highest level indexed lists and regularly spread malware to clueless site guests. Phishing-site identification devices ensure clients against such locales.

Machine learning is a multidisciplinary approach at first utilized in administered figuring out how to frame insightful models. It plays a significant viewpoint in a wide extent of genuine applications, for example, picture acknowledgment, information mining, gifted frameworks and picture acknowledgment [5]. This methodology seems appropriate to address phishing page location, since this issue can be changed over into an undertaking of grouping. ML strategies can be utilized to create models to distinguish phishing exercises dependent on ordering old pages and afterward these models can be incorporated into the program.

2. Literature Survey

[6] focused on applying a profound learning structure to recognize phishing sites. They have planned two sorts of highlights for web phishing: unique highlights and communication highlights. A recognition model dependent on Deep Belief Networks (DBN) was then introduced. Te test utilizing genuine IP streams from ISP (Internet Service

Provider) showed that the distinguishing model dependent on DBN accomplished a roughly 90% genuine positive rate and 0.6% false positive rate.

[7] proposed a novel phishing identification model dependent on a novel neural organization arrangement technique. The location model accomplished high exactness and had great speculation capacity by configuration hazard minimization standard. Moreover, the preparation interaction of the novel location model was straightforward and stable by Monte Carlo calculation. In view of testing of a bunch of phishing and kind sites, they have noticed that phishing location model accomplishes the best Accuracy, True positive rate (TPR), False-positive rate (FPR), Precision, Recall, F-measure and Matthews Correlation Coefficient (MCC) tantamount to different models as Naive Bayes (NB), Logistic Regression(LR), K-Nearest Neighbor (KNN), Decision Tree (DT), Linear Support Vector Machine (LSVM), Radial-Basis Support Vector Machine (RSVM) and Linear Discriminant Analysis (LDA).

[8] propelled to perform ELM got from various 30 primary parts that are sorted utilizing the ML approach. The majority of the phishing URLs use HTTPS to try not to get identified. There are three different ways for the location of site phishing. The crude methodology assessed various things of URL, the subsequent methodology dissected the authority of a site and computing if the site was presented and it additionally examined who was overseeing it, the third methodology checking the validity of the site

[9] investigated AI strategies and assessed their exhibitions when prepared to perform against datasets comprising of highlights that can separate between a Phishing Website and a protected one. The capacity of distinguishing those destinations from each other was imperative in the current web surfing. As increasingly more of assets move on the web, one weakness and a hole of delicate data by somebody could bring everything down in an associated network. Their significant target was to feature the best strategy for distinguishing perhaps the most generally happening digital assaults and consequently permit quicker ID and boycotting of such locales, thusly prompting a more secure and safer web riding experience for everybody. To accomplish that, they depicted every one of those methods into in incredible detail and utilized distinctive assessment strategies to depict their presentation outwardly. In the wake of setting those strategies in opposition to one another, they have closed with a clarification that Random Forest Classifier accomplishes to be sure turn out best for Phishing Website Detection.

[10] proposed an improved boycott technique that utilizations key segregate highlights removed from the source code of the site for the location of phishing sites. The primary center was to identify the phishing destinations that are copies of existing sites with controlled substance. Each phishing site was related to a novel unique finger impression that was created from the arrangement of proposed highlights. They utilized Simhash calculation to produce unique finger impression for every site. The highlights utilized for computing unique mark are filenames of the solicitation URLs (js, img, CSS, favicon), pathnames of solicitation URLs (CSS, contents, img, anchor connections), and characteristic estimations of labels (H1, H2, div, body, structure). The experimentation distinguished 84.36% of phishing locales as copies of other phishing sites with controlled substance while keeping zero bogus positive rate. The technique was like that of conventional boycott with a bit of leeway that it can distinguish reproduced and controlled phishing destinations proficiently

3. Proposed Methodology

The work stream of the proposed approach is portrayed in figure 1. The data used for the implementation is retrieved from PhishTank.com dataset and is processed as follows:

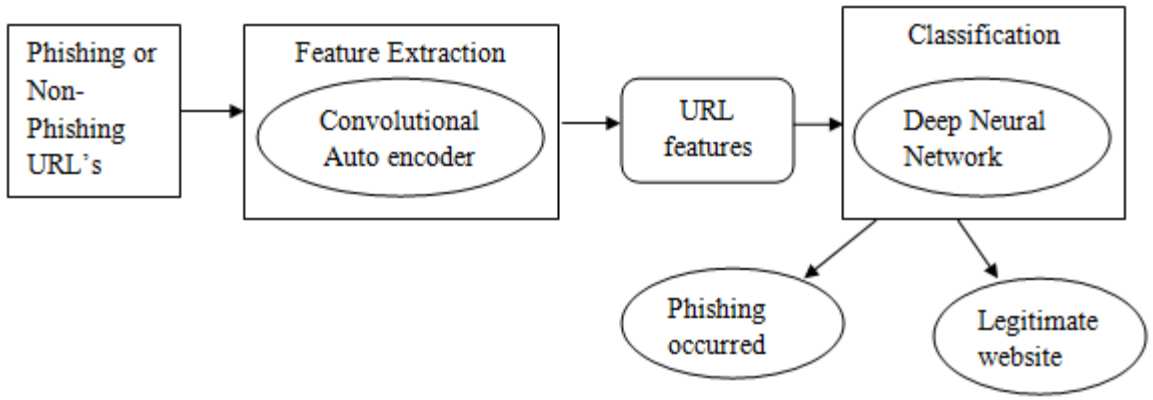

Figure 1: Work flow of the proposed methodology

Dataset:

We gather 16000 of phishing and real URLs. The phishing sites comprise of 12000 phishing URLs that has been gathered from PhishTank. In the other hand, the genuine sites comprise of 4000 real URLs that have been gathered by an everyday use from 10 picked clients. In any case, the final dataset subsequent to dealing with missing information and eliminating the copy is size of 6116 [11]. The phishing sites have certain qualities and examples that can be considered as highlights.

3.1 Feature Extraction:

At first, all the phishing and non-phishing URL's are fed into the feature collector. The feature collector then extracts features based on Convolutional autoencoder. The extracted features such as address bar based features, domain based features and HTML, javascript based features.

3.1.1 Convolutional auto encoder

The element extraction step comprises of a Python shape recognition calculation that empowers pin following. The picked highlights are in this way the 254 pin organizes (x and y directions of the 127 pins). Here, an unaided component determination is led on all the acquired URL's decreased to 28 × 28 pixels. The manual decision of sufficient portrayal factors is skirted utilizing the de-noising auto encoder neural organization engineering. Basically, auto encoder network attempts to estimated the character work, in order to yield, that is like info. The conventional autoencoder neural organization engineering comprises of an encoder and a decoder part, associated at the tightest purpose of the organization [12]. The two can be characterized as planning

$$\left.\begin{array}{l}\Gamma : A \to F \\ \Phi : F \to A\end{array}\right\} \in \arg\min \left\| X - (\Phi \circ \Gamma) X \right\| \qquad (1.1)$$

The encoder some portion of the organization diminishes the quantity of highlights starting with one concealed layer then onto the next, ideally bringing about a bunch of (pseudo-) symmetrical highlights with insignificant data misfortune, so the decoder part can recreate the info. The back spread calculation changes the auto encoder neuron loads toward the misfortune work negative inclination, limiting the distinction between the information and remade yield picture. When the auto encoder is prepared, just the encoder part is conveyed in applications, giving a more modest arrangement of highlights, while the decoder part is just utilized as a way to prepare the encoder part, and is dismissed in the wake of preparing. Autoencoder design can be assembled utilizing any sort of neurons. As an option in contrast to PC vision procedures, convolutional bits were utilized as the fundamental structure squares of the auto encoder neural organization, because of their capacity to effectively deal with 2D information. The decision of this arrangement was roused by the 2D organization of the sensor yield for speed and effortlessness, yet with the essential explanation being the capacity to store and accentuate the neighborhood spatial relations. Utilizing CNNs sidesteps PC vision calculations that can fizzle in mass location under enormous disfigurements. Another favorable position of utilizing CNNs is the capacity to safeguard and use the spatial data naturally encoded in the URL. When utilizing linearised portrayals, for example, a variety of places of pins, some spatial data can be lost relying upon the pin requesting inside the exhibit. The highlights recovered are of three classes. These are as per the following:

- Address Bar-based features (URL Based)
- Domain Based Features.
- HTML and Javascript Based features.

All these features have a unique property that distinguishes it from other features and thus helps in the detection of phishing web pages. Based on these features a specific URL can be categorized either as legitimate or phishing. The feature clearly states that if the website contains an IP address, it can be considered as an illegitimate one. If the URL string is too long then it can be a malicious one. Based on URL length string, URL can be categorized into legitimate, and phishing. If the URL length is too short, there is a chance of a website phishing attack on the computer user. The rule depicted is used to differentiate between a normal URL and an abnormal one. The use of "@" symbol redirects the browser to ignore the address preceding this symbol. If "//" exist within the URL, it will redirect to another webpage, typically a malicious one. The location of "//" appears to be in the sixth position, in case of HTTP, and at seventh position in case of HTTPS. Use of "–" symbol is often avoidable in URL. The attacker used this symbol to add as a prefix or suffix which tend to look a legitimate webpage. Similarly, if identity is not the part of the URL, then that URL is categorized into an illegitimate one.

3.2 Classification:

The URL highlights are arranged into phishing or authentic dependent on profound neural organization calculation. The order structure dependent on profound neural organization chiefly incorporates three sections. The info layer, shrouded layers and the yield layer. The contrast between various order models is the organization structure

chose by the profound neural organization layer [13]. Dissimilar to conventional AI models, profound learning models can encode the attributes of a sentence. This portrayal of highlight reflection decreases measure of manual element extraction. The profound underlying model of profound neural organizations takes into account a more extravagant level learning of highlights. The yield from the element extraction is in worked to the contribution of profound neural organization. This thusly lessens the mistake rate and creates the outcome as genuine or phishing URL.

4. Result and Discussion

This section clearly depicts the results obtained from the implementation of phishing website detection using novel features and machine learning techniques. The results are evaluated based on the performance metrics such as computational time, accuracy and detection time. The results obtained are as follows:

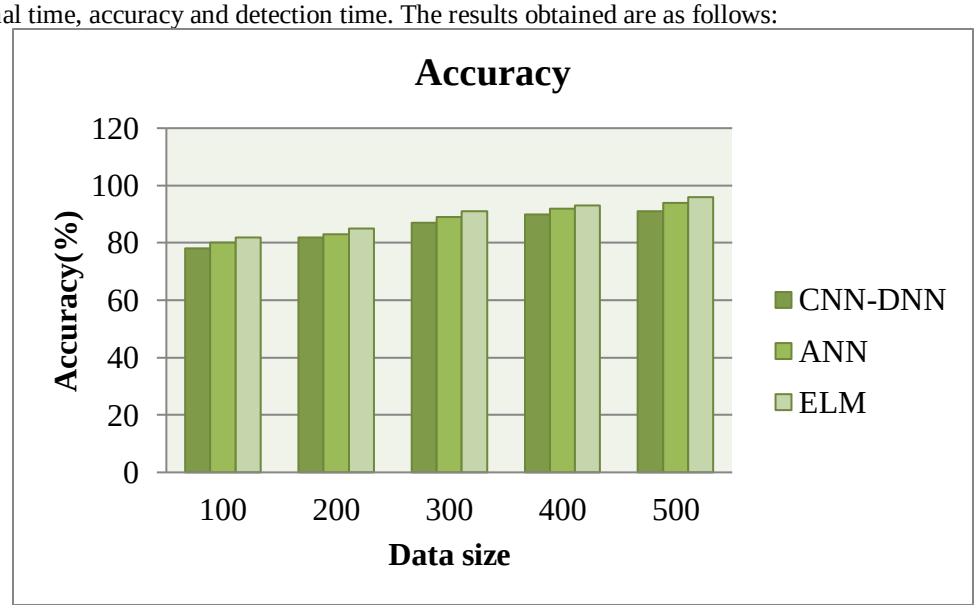

Figure 2: Accuracy rate obtained for the proposed methodology

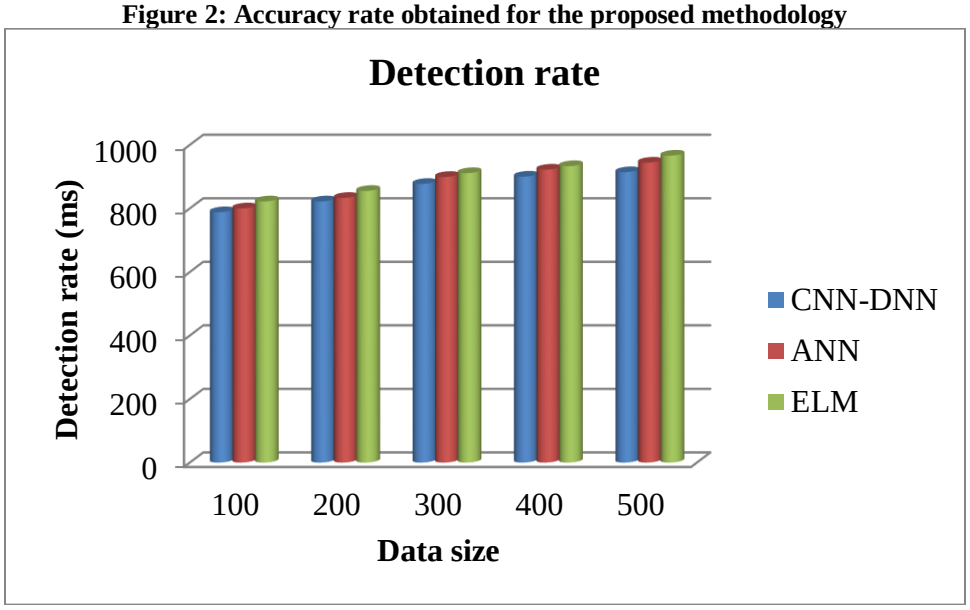

Figure 3: Detection rate obtained for the proposed methodology

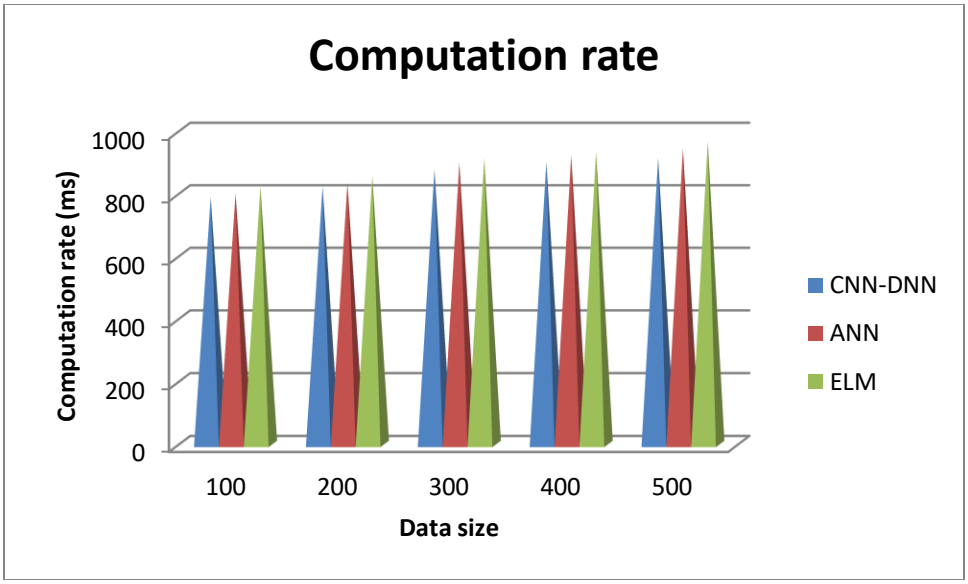

Figure 4: Computational time obtained for the proposed methodology

5. Conclusion

In this paper, we analyzed the features of phishing websites and presented three types of features for web phishing detection. At first, the URL websites were loaded for feature extraction. Then the extracted features were classified based on Deep Neural Network algorithm. The implemented results showed that the accuracy, computational time and detection rate for the proposed methodology is much better for the varied data sizes. The accuracy rate achieved is 86% for the proposed methodology.

References

1. Sahingoz, Ozgur Koray, Ebubekir Buber, Onder Demir, And Banu Diri. "Machine Learning Based Phishing Detection From Urls." Expert Systems With Applications 117 (2019): 345-357.
2. El Aassal, Ayman, Shahryar Baki, Avisha Das, And Rakesh M. Verma. "An In-Depth Benchmarking And Evaluation Of Phishing Detection Research For Security Needs." IEEE Access 8 (2020): 22170-22192.
3. Jain, Ankit Kumar, And Brij B. Gupta. "A Machine Learning Based Approach For Phishing Detection Using Hyperlinks Information." Journal Of Ambient Intelligence And Humanized Computing 10, No. 5 (2019): 2015-2028.
4. Rao, Routhu Srinivasa, And Alwyn Roshan Pais. "Jail-Phish: An Improved Search Engine Based Phishing Detection System." Computers & Security 83 (2019): 246-267.
5. Mao, Jian, Wenqian Tian, Pei Li, Tao Wei, And Zhenkai Liang. "Phishing-Alarm: Robust And Efficient Phishing Detection Via Page Component Similarity." IEEE Access 5 (2017): 17020-17030.
6. Chin, Tommy, Kaiqi Xiong, And Chengbin Hu. "Phishlimiter: A Phishing Detection And Mitigation Approach Using Software-Defined Networking." IEEE Access 6 (2018): 42516-42531.
7. Feng, Fang, Qingguo Zhou, Zebang Shen, Xuhui Yang, Lihong Han, And Jinqiang Wang. "The Application Of A Novel Neural Network In The Detection Of Phishing Websites." Journal Of Ambient Intelligence And Humanized Computing (2018): 1-15.
8. Smadi, Sami, Nauman Aslam, And Li Zhang. "Detection Of Online Phishing Email Using Dynamic Evolving Neural Network Based On Reinforcement Learning." Decision Support Systems 107 (2018): 88-102.
9. Bahnsen, Alejandro Correa, Eduardo Contreras Bohorquez, Sergio Villegas, Javier Vargas, And Fabio A. González. "Classifying Phishing Urls Using Recurrent Neural Networks." In 2017 APWG Symposium On Electronic Crime Research (Ecrime), Pp. 1-8. IEEE, 2017.
10. Smadi, Sami, Nauman Aslam, And Li Zhang. "Detection Of Online Phishing Email Using Dynamic Evolving Neural Network Based On Reinforcement Learning." Decision Support Systems 107 (2018): 88-102.

11. Basit, Abdul, Maham Zafar, Abdul Rehman Javed, And Zunera Jalil. "A Novel Ensemble Machine Learning Method To Detect Phishing Attack." In 2020 IEEE 23rd International Multitopic Conference (INMIC), Pp. 1-5. IEEE, 2020.
12. Ryu, Seunghyoung, Hyungeun Choi, Hyoseop Lee, And Hongseok Kim. "Convolutional Autoencoder Based Feature Extraction And Clustering For Customer Load Analysis." IEEE Transactions On Power Systems 35, No. 2 (2019): 1048-1060.
13. Cao, Xiaoyu, And Neil Zhenqiang Gong. "Mitigating Evasion Attacks To Deep Neural Networks Via Region-Based Classification." In Proceedings Of The 33rd Annual Computer Security Applications Conference, Pp. 278-287. 2017.