Phishing Website Detection using Machine Learning Techniques and CNN

Asst. Prof. Deepa Mary Vargheese
Department of Electronics
TIST
Kochi, India

Sreelakshmi N R
Department of Electronics
TIST
Kochi, India

***Abstract*—Today, everyone is highly dependent on the internet. Everyone performed online shopping and online activities such as online Bank, online booking, online recharge and more on internet. Phishing is a type of social engineering attack often used to steal user data including login credentials and credit card numbers. As the Internet grows in size, automatic URL detection becomes more important to provide end users with timely protection. The proposed system is developed to provide an effective and versatile malicious URL detection system with a comprehensive set of attributes that reflect many aspects of phishing webpages and their hosting platforms, including features that are difficult to fabricate by a criminal.This system will helps to identify the number of suspicious emails and bringing a new level of security in the insecure world.The proposed technique exhibits optimistic results when bench marking with a range of standard phishing datasets.**

Keywords- Phishing, web, machine learning, principal component analysis, Random Forest classifier, support vector machine, convolutional neural network

I. INTRODUCTION

The internet has evolved into a platform for many illegal enterprises such as spam, financial fraud, and malware distribution. The precise commercial reasons for this strategy may differ, but one common thread is that users are not required to visit their website. This visit should be available by email, web query items, or links from other site pages; however, the client must be able to make a quick decision, such as indicating the optimal URL (Uniform Resource Locator) and obtaining important information. To combat this, the security community developed a blacklist service that is packed in toolbars, devices, and search engines and provides accurate warnings or alerts. Many hazardous sites are not banned because the site is too new, unclassified, or misclassified. The internet has evolved into a resource for a variety of purposes. Phishing is a type of cyber-attack that uses websites to obtain valuable buyer information such as store card numbers, accounts, login credentials, and more.

The anti-phishing solution is most extensively deployed on a blacklist warning system, which is existing in common web browsers like Chrome, Internet Explorer and Mozilla Firefox. The blacklisting interrogative gadget has a central database of regarded phishing URLs, and consequently can't discover newly launched phishing web sites. Machine learning based phishes detection gadget relies upon efficiently on the aspects of accuracy. The most of anti-phishers researchers center of attention on optimizing new feature proposals or classification algorithms, where developing proper features analysis and selection techniques is not the important plan.There are several methods for detecting software, including blacklists, machine learning, and hybrid approaches. In general, two primary strategies for picking characteristics were used: filter size and wrapper. On the other hand, filter measurements are metrics. It is derived from statistical and informative theories that can reflect the merits of any character function without requiring the use of a precise classifier. The wrapper procedure is repeated, with each execution resulting in the production of a subset of elements and their classification.

II. RELATED WORKS

PILFER is a technique presented by Sadeh et al. [2] for classifying phishing URLs. They extracted a collection of ten traits that are aimed to reveal fraudulent techniques used to deceive people. There are approximately 860 phishing emails and 6950 non-phishing emails in the data set. They used 10-fold cross validation to train and test the classifier and got 92 percent accuracy.Ma et al. [3] treated URL classification as a binary classification problem and developed a URL classification system that processes a live feed of labelled URLs. It also collects URL characteristics from a large Web mail provider in real time.Both lexical and host-based features were used. They were able to train an online classifier using a Confidence Weighted (CW) method using the acquired characteristics and labels.After examining 358 research papers in the area of phishing countermeasures and their effectiveness, Parkait et al. [4] give a complete literature evaluation. They divided anti-phishing techniques into eight categories and emphasised advanced anti-phishing techniques.Multi-label Classifier based on Associative classification was developed by Abdelhamid et al. [5] for detecting phishing URLs (MCAC). They divided URLs into three categories based on sixteen features: phishing, legitimate, and suspect. The MCAC is a rule-based system that extracts several label rules from phishing data.In their assessment on malicious webpage detection systems, Patil [6] presented a quick outline of several types of web-page attacks.

Hadi et al. [7] classified phishing URLs using the Fast-Associative Classification Algorithm (FACA). FACA works by identifying all common rule item sets and creating a classification model. They looked at 11,055 websites and divided them into two categories: authentic and phishing.

There were thirty features in the data collection. They selected two percent as the minimal support criterion and fifty percent as the minimum confidence threshold, respectively.Nepali and Wang introduced a novel method for

detecting fraudulent URLs that relied solely on public social network features.Using supervised learning using features points derived from WHOIS and DNS metadata, Kuyama et al [5] devised a method for identifying the Command and Control server (C&C server). They used domain names and email addresses from WHOIS as machine learning input values. Ouyang et al introduced a machine learning-based multi-stage pipelined spam email detection system that includes a large number of network features. They examined their methods using email data acquired over the course of two years,consisting of over 1.4 million messages, and reported a true positive rate between 12% to 77% using the Decision Tree algorithm.Xiang et al.19 introduced CANTINA+, a multi-layer machine learning system that uses features from URLs, HTML DOM, search engines, and third-party services like Page Rank to detect phishing websites. Of sum, 10 of the 15 characteristics in their architecture are derived from HTML or URL textual patterns and forms.But this resulting model is hard to interpret due to the massive number of algorithmically generated lexical features.

III. PROPOSED SYSTEM

The proposed system uses machine learning and Deep Learning Algorithm, to predict the legitimate and phishing websites.

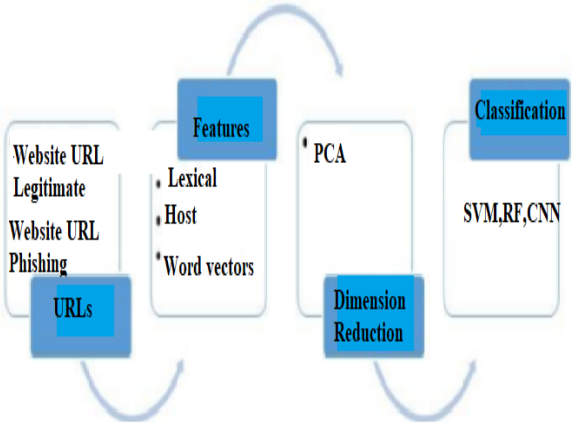


Fig 1: Proposed System

The block diagram of the phishing website detection technique is shown in figure 1.

First of all,using the GNU Wget and Python scripts, collect web pages automatically.Also download relevant resources (e.g. images, CSS, JavaScript) in addition to the whole HTML text so that it can provide a browser to all the web pages that are downloaded. All web sites screenshots are also saved for later inspection and filtering.Download datasets that are further processed to load phishing and legitimate data sets, or to fix web page problems that have resulted in a "Error 404" page. The web page's duplicate instance is likewise erased. The function is extracted once the sample has been filtered. In classic phishing site detection investigations, there are two categories of features: Internal functions; External features.The internal functions are obtained from the webpage's URL and HTML source code, which may both be viewed straight from the web page. External functions, on the other hand, are focused on benchmarking and are obtained from requests from third-party services such as domain registries, search engines, and WHOIS records.Then choose 5000 phishing web pages in particular, and all of them are more stable, notably in terms of URLs. The fish tank is totally constructed from Alexa URL and Common Crawl archives.To extract feature vectors from the input URL employed vocabulary, host, and word. The vocabulary feature is a feature of text URLs that includes things like host-name length, URL tokens, and so on. For excessive classification of machine learning vocabulary features, a simple calculation, security, and precision are required.The vocabulary function is the property's URL text, not the page's content. The host-name length, the entire URL length, the number of dots in the URL, the hostname (separated by '), and the binary function (separated string) for each symbol are among these properties. (. '/',,'=','?' "-'East' ') in the URL path. This is also known as "pocket","where dangerous sites are housed," "have," and "control" are all examples of host-based characteristics.The hostname as part of the URL identifies the following host properties.Words in vectors are especially useful for performing crucial activities such as the URL of a web page. It primarily comprises of a text with numerous words. The automatic vectorization method is preferable to modifying the text in this manual. A Weka function called "String to Word Vector" is used to transform each URL into carrier-specific words. For features, the Principal Component Analysis (PCA) was used to prevent high dimensionality.

The goal of PCA is to condense a huge number of variables into a manageable number of variables. It's a well-known statistical strategy for explaining the covariance shape of data with a small number of variables.These elements are linear mixtures of the original variables that frequently allow interpretation and a better understanding of the various sources of version. To get the final result, employ a classifier in this phase. The classifier is just a machine learning system that has been taught to predict and classify results. Because no single classifier is complete and precise. Classifiers were chosen primarily because they have already been utilized for Google-related concerns like spam detection, phishing emails, phishing websites, and malicious URLs. The system merely tries to use this system for final categorization and prediction.For classification, support vector machine,Random forest classifier and CNN are utilized.SVM works with instances of training and changes that have been made, such as saving a sample of URLs from two classes with a hyperplane in the feature space modified, and making maps of feature set to build a feature room changed.

IV. RESULTS

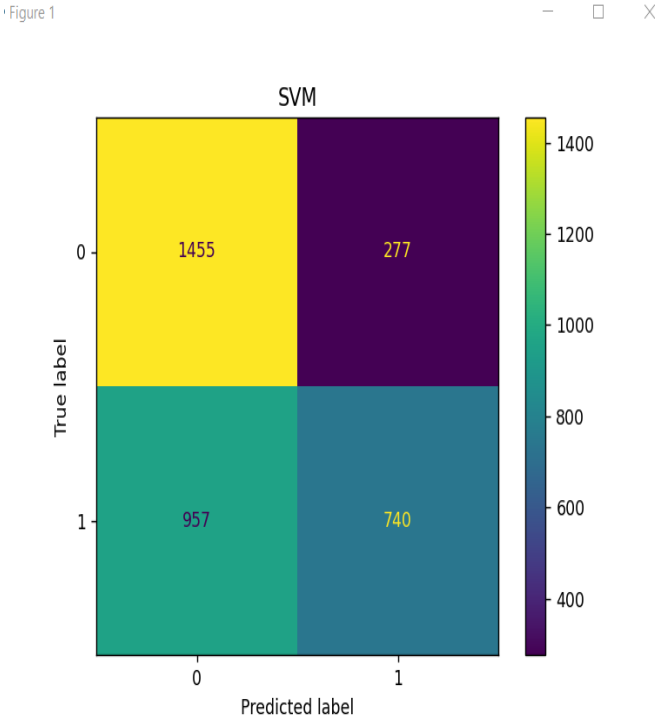
A comparison is made between the suggested machine learning-based technique and the existing technique. The split test is trained using a comparable classification algorithm. 70% of the statistics used for training are retained for testing purposes in each partition. This equation is used to calculate accuracy. Positive denotes correct, while TN denotes true negative, implying FN false negative and false positive FP capabilities. We calculate the accuracy and results for machine learning algorithms using this equation.

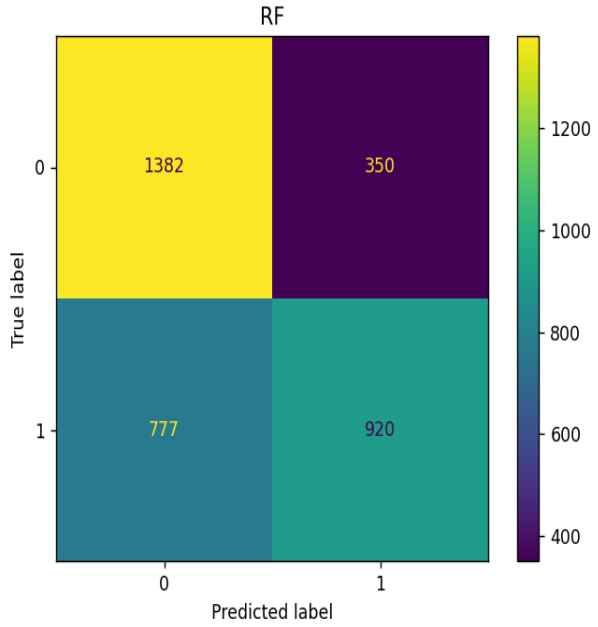
The proposed technique outperforms earlier techniques in terms of total performance. Table 1 compares the overall performance of the alternative strategies..

	SVM	RF
Precision 0-Legitimate 1-Phishing	 0.60 0.73	 0.64 0.72
Recall 0-Legitimate 1-Phishing	 0.84 0.44	 0.79 0.55
f1-score 0-Legitimate 1-Phishing	 0.70 0.55	 0.71 0.62
Accuracy	64	67

Table 1:Performance comparison of SVM & RF

The overall accuracy of SVM and RF after training the dataset is about 64% and 67% respectively.So inorder to save humans from large phishing attacks,a convolutional neural network (CNN) is required to safe ourselves from attacks.





The diagrams showing the overall performance and accuracy of SVM,RF and CNN.There are 2 types of feature set.
1) 0 for legitimate
2) 1 for phishing
In SVM,1455 datasets are correct and 277 are wrong,means that 277 datasets which are taken as legitimate will shows the result as phishing and 957 datasets will show the result as legitimate.Same thing is happened in RF,that is 350 legitimate sites after training will give a result as phishing and 777 data's of phishing sites will show result as legitimate.

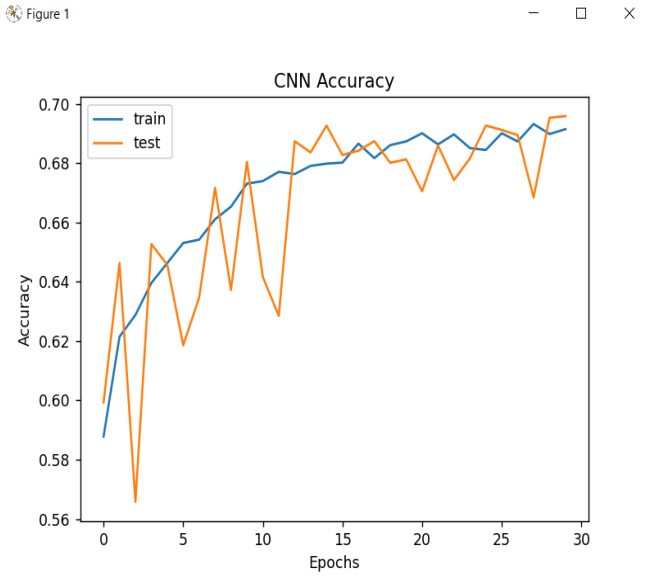


Fig 2:Graph showing the accuracy of CNN

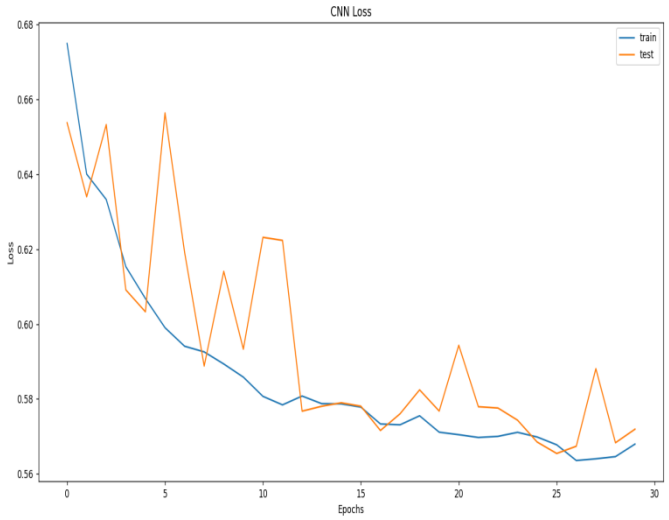


Fig 3:Graph showing the loss of CNN

Compared to accuracy of SVM and RF,CNN is having better accuracy and performance of about 70% and can detect the URL of the phishing website without relying on third-party data and search engines,with a highest classification accuracy.

V. CONCLUSION

The proposed phishing website detection technique uses both machine learning and deep learning technologies to detect phishing attacks .CNN was employed in the proposed technique, which had a 70 percent accuracy and an extremely low false-positive rate.This method can detect new temporary phishing sites and lessen phishing attack impact.The feature extraction and the models used on the dataset helped to uniquely identify phishing URLs and also the performance accuracy of the models used. It is also surprisingly accurate at detecting the genuineness of a URL link. The proposed machine learning and CNN-based technique outperforms existing phishing detection solutions in terms of effectiveness.

VI. REFERENCES

[1] Higashino, M., et al. An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage. in 2019 5th International Conference on Information Management (ICIM). 2019.
[2] H. Bleau, Global Fraud and Cybercrime Forecast,. 2017.
[3] Michel Lange, V., et al., Planning and production of grammatical and lexical verbs in multi-word messages. PloS one, 2017. 12(11): p. e0186685-e0186685.
[4] Rahman, S.S.M.M., et al. Performance Assessment of Multiple Machine Learning Classifiers for Detecting the Phishing URLs. 2020. Singapore: Springer Singapore.
[5] He, M., et al., An efficient phishing webpage detector. Expert Systems with Applications, 2011. 38(10): p. 12018-12027.
[6] Mohammad, R.M., F. Thabtah, and L. McCluskey. An assessment of features related to phishing websites using an automated technique. in 2012 International Conference for Internet Technology and Secured Transactions. 2012.
[7] Abdelhamid, N., A. Ayesh, and F. Thabtah, Phishing detection based Associative Classification data mining. Expert Systems with Applications, 2014. 41(13): p. 5948-5959.
[8] Toolan, F. and J. Carthy. Feature selection for Spam and Phishing detection. in 2010 eCrime Researchers Summit. 2010.
[9] Jain, A.K. and B.B. Gupta, Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems, 2018. 68(4): p. 687-700.
[10] 1PhishTank, Phishing dataset. 2018, Verified phishing URL.
[11] 1Chiew, K.L.,et al.,Utilisation of website logo for phishing detection.Computers & Security, 2015. 54: p.16-26.
[12] Benavides, E., et al. Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review. 2020. Singapore: Springer Singapore.
[13] Zhang, W., et al., Two-stage ELM for phishing Web pages detection using hybrid features. World Wide Web, 2017. 20(4): p. 797-813.