Sparse Decomposition Feature Selection Technique for Phishing Website Identification

Aswathy.S [1], Prameela.S [2], Suresh Kumar N [3]
[1]*Mtech Scholar*, Dept. of CSE, College of Engineering and Management, Punnapra, Kerala, India
[2] *Senior Lecturer*, Dept. of IT, College of Engineering and Management, Punnapra, Kerala, India
[3] *Senior Lecturer*, Dept. of CSE, College of Engineering and Management, Punnapra, Kerala, India
(E-mail: aswathyshreyas32@gmail.com)

Abstract- Phishing attacks are one of the major problems facing the cyberworld. It should be the common security challenges that peoples and organizations face in keeping their data secure. The Phishing website tries to steal passwords or other confidential information of users. This paper proposes the sparse decomposition feature selection method. Comparing to the existing schemes whose search time is very faster. Autoencoder machine learning technique is used for the identification of phishing website. It also compares the accuracy with the PCA feature selection method.

Keywords - *Phishing, Machine Learning, Feature Selection, Sparse Decomposition, Autoencoder (AE), Principal Component Analysis (PCA).*

I. INTRODUCTION

Along with the development of internet technology, machine learning research has greater importance and that grows fastest. Machine learning creates a model for predicting test data. Here the computer is learned from experience. Machine learning is the field of computer science and also the branch of artificial intelligence. The algorithm composed of mainly three phases. Training validation and testing. Machine learning has attracted increasing attention in many disciplines such as Speech Recognition, Medical Diagnosis, Automatic Language Translator, and online Phishing Website Identifications.

“Phishing may be a fraudulent attempt, usually made through email, to steal your personal information". The definition is taken according to Phish Tank [1]. The motives of phishing attackers are financial gain, Identity Hiding, etc. The phishing detection approach mainly is of two types. The user training approach and software classification approach depicted in figure 1. Phishing is a crime in which the attackers send fake emails to the company or organization. The phishing website looks very similar to the original website to attract a large no of users.

Phishing is a very popular method used in network attacks and leads to privacy leaks, identity theft, and property damage. The current phishing detection method based on machine learning mainly uses a supervised classification algorithm to detect the legitimacy of websites. The classification model introduces the marked website dataset, trains the prevailing classification model with a training dataset, and predicts the legitimacy of internet sites through the trained classifier. Phishing emails may also include attachments that will install malware on your computer when you open them. Figure 2 represents the original and fake web page of the popular website www.ebay.com. The fake page is similar to the original site but it spread phishing activities. The user can enter the identity details on the fake page the attacker can steal the private data and use it for fraud activities.

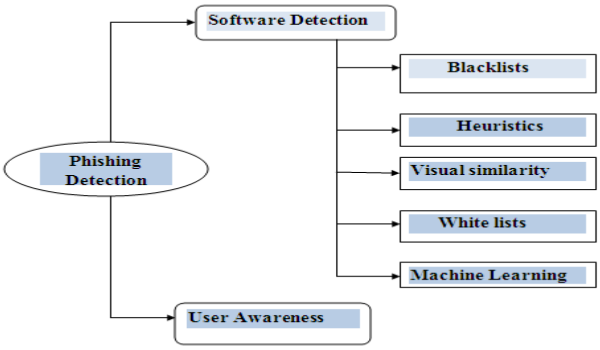

Fig. 1: Overview of phishing detection approaches.

Nowadays, a variety of phishing attacks. Spoofing, Instant spam spoofing, Host file poisoning, malware-based phishing, an-in-the middle, Session hijacking, DNS based phishing, Deceptive phishing, Key loggers/loggers, Web Trojans, Data theft, Content-injection-phishing.[2] Search-engine-phishing, Email/spam, Web-based delivery, Link Manipulation, System reconfiguration, Phone phishing, Data shoplifting, Blacklisting.

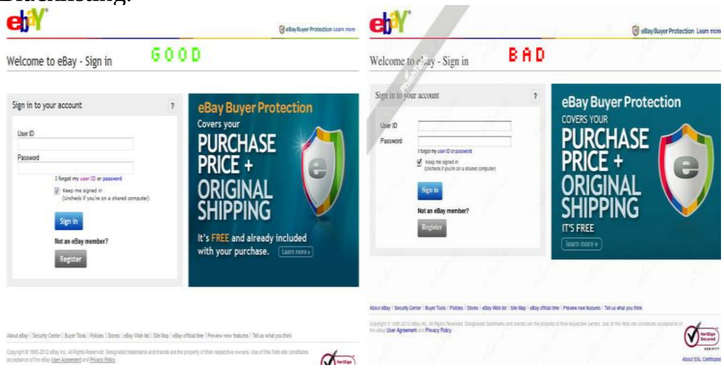

Fig. 2: Original and fake web pages of eBay.

[https://www.google.com/search?q=different+type+of]

This paper explores the idea of automatic Phishing Website Identification. There are different machine learning algorithms [13] that were developed for phishing website detection, but these algorithms use Principal component analysis for feature extraction. In this paper, we use the Autoencoder Machine learning technique for phishing identification. This technique uses the Sparse Decomposition method for feature selection. The sparse decomposition technique is very fast compared to existing feature selection schemes. Here the dataset is a collection of URLs.The data of URL contains a large no of features. Autoencoder is a powerful dimensionality reduction technique. It is an unsupervised label and does not require external labels. Learned automatically from the data examples. It generally consists of two parts first an encoder and seconds a decoder. In machine learning, we often have to deal with structural data of the table of rows and columns of a matrix. So, sparse decomposition is very essential for feature selection. Most of the elements of the matrix have zero value is called a sparse matrix. Our phishing dataset is ternary and it contains binary values. The dataset consists of 32 input attributes and 1 output attribute. The input attribute takes 3 different values 1, 0, and -1. The output attribute can take two values which are 1 and -1.

Detecting phishing domains is considered as a classification problem. Therefore labeled data that have samples as phishing domains and legitimate domains in the training phase are needed. The dataset used in the training phase is one of the crucial points to build a successful detection mechanism. Detection Systems should use samples whose classes are precisely known. So, the samples which are labeled as phishing must be detected as a phish. Likewise, the samples which are labeled as legitimate must be detected as legitimate. Otherwise, the system cannot work correctly if we use samples that we are not sure about the class information.

The proposed model consists of the following steps as Collection of the dataset, preprocessing, feature selection, Autoencoder training, and classification, and performance evaluation. Each step is described in further sections. This paper is organized as follows: Section II presents related works in connection phishing website identification. Section III describes the proposed methodology. Section IV gives details regarding the conclusion and future work.

II. RELATED WORKS

The Uniform Resource Locator (URL) of the phishing site and legal site will look similar and the user will be misguided. Criminals, who want to be accessed sensitive data, first build unauthorized replicas of a real website and email, typically from a financial institution or another organization dealing with financial data. One of the Phishing identification methods is the Blacklist approach is described in the paper [5]. It should be the updated list of previously detected phishing URLs. Comparing to the machine learning blacklist has lower false positive (FP) rates and cannot protect for zero hours phishing attack. It can detect only 20 phishing attacks.

In paper [6] describes a visual similarity approach, the user extracting images of the legitimate site. The limitation of this approach is image comparison takes more space and time. It produces a high false-negative rate. Visual similarity-based detection technique utilizes the feature set like text content, text format, HTML tags, Cascading Style Sheet (CSS). The attacker does not copy the visual appearance of a website well. The attacker fools the user by visual appearance, address bar, embedded objects, favicon similarities. Visual similarity uses a signature to identify phishing WebPages. The signature created by taking the features of a whole website or a single webpage. Therefore one signature is sufficient to detect various WebPages of a single website.

In paper [7] describes a Heuristic-based approach. It should be the extension of blacklists and can detect the new attack by using features. But it's limitation is it cannot find an all-new attack. If the attacker knows the algorithm and features used. Then not identify the phishing website. The heuristic approach identifies a zero-hour phishing attack. FP rates are greater than a blacklist. In paper [5][9] describes white listing is the practice procedure. It is the reverse of blacklisting. It externally allowing some entities to access a particular service. In paper [10] describes a content-based approach and is extracted features from a website by using HTML code or by using the content of email.his algorithm explains that WebPages containing more external field. Then the content-based approach is essential. Paper [10] describes the Machine learning technique from a given training set is to learn legitimate or phishing emails. His paper explains the insights into the effectiveness of using different ML techniques. The study is compared with a few learning approach including SVM, Random forest, and Naïve Bayes

III. PROPOSED METHODOLOGY

The proposed paper focuses on a feature selection method by using sparse decomposition and identifies a phishing website by using Autoencoder. It composed of five stages: Data collection, Preprocessing, Feature selection, Autoencoder Building, Performance analysis, and comparison with PCA. Figure 3 shows a flowchart of the proposed architecture.

A. Dataset

The dataset used in this study is taken from Kaggle (www.kaggle.com). This data is used for classification and it should be the collection of URLs. It contains 32 attributes and 11054 instances. Most of the instances in the data are binary values.

B. Preprocessing

Preprocessing is a method for extracting useful information from the data through various operations. The raw data will be preprocessed to remove duplicate or null values.

C. *Feature Selection*

Features (dominant) are selected using a sparse decomposition technique based on feature alone decomposition and target-based decomposition. The collected features are categorized into two types: lexical features and host-based features. The proposed feature selection reduces the no of features from high dimensional data. It transfers original features from different modalities to obtain correlation analysis.

1) *Sparse Decomposition Technique*

Sparse Decomposition Technique used for class discriminative feature selection. It is also helpful for finding the relationship between features and response variables. It naturally suffers from a high dimensional problem. Proposed feature selection of different kinds of relations inherent in data to select attributes. Features are dependent on a real application. Proposed feature transfer original features from different models to obtain correlation analysis. This method simplifies the redundant complex structure. It represents sparse solutions for a system of linear equations. The test sample is a linear combination of the training sample.

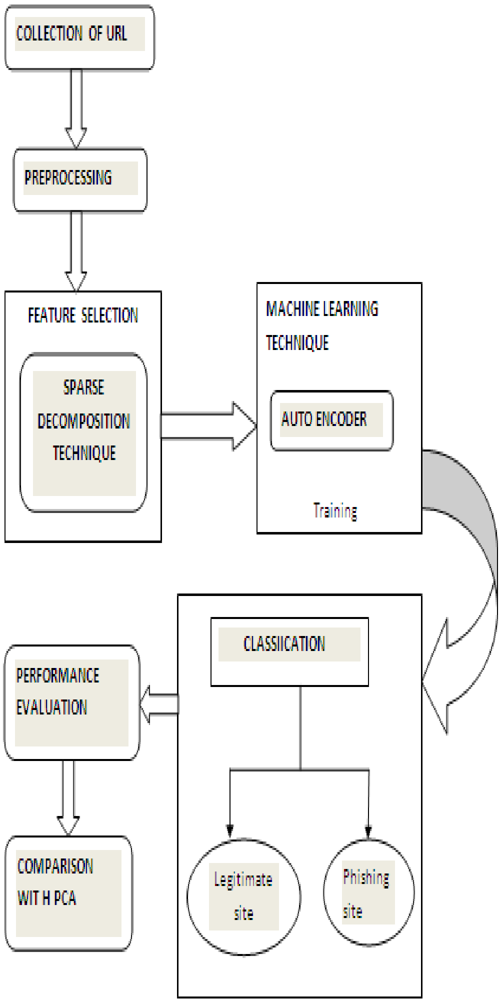

Fig .3: Proposed Architecture

D. *Autoencoder Building*

Autoencoder [9] is an unsupervised deep learning method. The structure of AE consists of an input layer, a hidden layer, and output layer. The no of neurons in the input and output layer is the same. The encoder part of the network is used for encoding and decreasing the no of hidden units in each layer. The decoder part increases the no of hidden units in each layer. It can be used for dimensionality reduction. The main properties of Autoencoder are Data specific, lossy, learned automatically from examples. The hyper parameters of the Autoencoder are code size, no of layers, no of nodes per layer, loss function. To building an Autoencoder need 3 things. Encoding function, decoding function, and a loss function. The encoder and decoder are parametric and is chosen from a neural network.

E. Performance Evaluation

Seven measures that affect the performance of the model are accuracy, sensitivity, specificity, false-positive rate, false-negative rate, precision, error rate. Accuracy is overall. Accuracy is defined as the ratio of correct predictions to the total prediction. Sensitivity also called a true positive rate. Specificity also called a true negative rate. The recall is the proportion of the actual positive which is predicted positive. Precision is the proportion of the actual positives which is positive. The matrices used in the proposed method are shown in figure 4. [15].

F. Comparison with PCA

Principal Component Analysis (PCA) is a dimensionality reduction method. It is also a supervised learning method. In this section, our proposed method is compared with PCA Plus other machine learning methods to detect which method has greater accuracy.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$False\ Positive\ Rate(FPN) = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$False\ Negative\ Rate(FNR) = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$Error\ rate = \frac{\text{FP} + \text{FN}}{\text{P} + \text{N}}$$

Fig .4: Performance Matrices

IV. Conclusion and Future Works

This paper presented a fully automated method for the identification of a phishing website using a sparse decomposition feature selection method and Autoencoder machine learning technique. The proposed model search time is very fast compared to the existing schemes. We present a new feature selection method sparse decomposition to find the correlated features. The proposed method has a low computational cost. In the future, our method should be tested with a deep Boltzmann machine and deep neural network with a sparse decomposition feature selection method.

References

[1] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey", IEEE communications surveys& tutorials, Volume 15, Issue No: 4, fourth quarter 2013.

[2] C.Whittaker, B.Ryner, and.Nazif "Large scale automatic classification of phishing page "in NDSS 2010.

[3] ZaKwihoon Kim, Yong-Geun Hong," General labeled data generator framework for network machine learning", International Conference on Advanced Communications Technology(ICACT), ISBN 979-11-88428-01-4, ICACT2018 February 11 ~ 14, 2018".

[4] Mohammad Mehdi Yadollahi, Farzaneh Shoeleh, "An Adaptive Machine Learning-Based Approach for Phishing Detection Using Hybrid Features", 2019 5th International Conference on Web Research (ICWR), 2019 IEEE.

[5] Anit Kumar Jain, B.B Gupta "A novel approach to protect against phishing attack at client side using auto-updated whitelist", EURASIP journal on information security,2016

[6] Ankit Kumar Jain and B.B Gupta, "Phishing Detection: Analysis of Visual Similarity-Based Approaches", National Institute of Technology, Kurukshetra, India, Hindawi, vol.2017.july

[7] Luong Anh Tuan Nguyen, Ba lam To, Huu Khuong Nguyen, "A Noval approach for phishing detection using URL based heuristic", International conference on computing, management, and telecommunications, IEEE 2014

[8] CARLOS LA ORDEN, BORJA SANZ, DeustoTech Computing, "Collective classification for spam filtering", Logic Journal of IGPL Advance Access published July 30, 2012

[9] Jian Feng, Lianyang Zou, and Tianzhu Nan, "A phishing webpage detection method based on the stacked Autoencoder and correlation coefficients", Journal of computing and information technology vol.27 no2 June 2014.

[10] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp. 639–648.

[11] Venkatesh Ramanathan and Harry Wechsler, Department of Computer Science George Mason University, "Phishing Website Detection Using Latent Dirichlet Allocation and AdaBoost", Washington, D.C., USA,2012 IEEE.

[12] Mohammad Mehdi Yadollahi, Farzaneh Shoeleh, Elham Serkani, "An Adaptive Machine Learning-Based Approach for Phishing Detection Using Hybrid Features", 2019 5th International Conference on Web Research (ICWR).

[13] shantTyagi, Jatin Shad, Shubham Sharma, Siddharth Gaur, Gagandeep Kaur, "A Novel Machine Learning Approach to Detect Phishing Websites",5th International Conference on Signal Processing and Integrated Networks (SPIN).2018.

[14] Christopher N. Gutierrez, Taegyu KimyNorthrop Grumman Corporation, Learning," from the Ones that Got Away: Detecting New Forms of Phishing Attacks", DOI 10.1109/TDSC.2018.2864993, IEEE

[15] Priyanka Singh, Yogendra P.S. Maravi, "Phishing Websites Detection through Supervised Learning Networks", School of Information Technology2015 IEEE.

[16] Mazharul Islam, Department of Computer Science & Engineering, "Phishing Websites Detection Using Machine Learning-Based Classification Techniques", International Islamic University of Chittagong, Bangladesh

Aswathy.S S is an M.Tech scholar in Dept. of CSE, College of Engineering and Management Punnapra affiliated to APJ Abdul Kalam Technological University. She has completed her bachelor's degree from Sri Vellappally Natesan College of Engineering in 2015. Her area of interest includes Machine learning. and artificial Intelligence.
Email id: aswathyshreyas32@gmail.com

Mrs. Prameela.S is working as Assistant Professor in the Department of Computer Science and Engineering, College of Engineering and Management, Punnapra. Her area of interest includes Bioinformatics and Machine Learning Email id: s_prameela@yahoo.co.in

Suresh Kumar N is working as an Assistant professor in the Department of Computer Science Engineering at the College of Engineering and Management, Punnapra. His area of interest includes the Internet of Things, Digital Image Processing, and Machine Learning.
Email id:cnsuresh2000@gmail.com