International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Literature Review on Phishing Website Detection Using Deep Learning

Akshatha A P[1], Chaithra R[1], Madhura S[1] , Chandana C Sagar[1] , Hiriyanna G S[2]

** 1UG students, Dept. of CS&E , JNNCE, Shivamogga, India.*
** 2Asst. Prof ., Dept. of CS&E , JNNCE, Shivamogga, India.*

---***---

Abstract - With the escalating threat of phishing attacks on the internet, the need for effective and efficient methods to identify phishing websites has become paramount. This study explores the application of Deep Learning (DL) techniques for the automated detection of phishing websites. Leveraging the power of neural networks, we analyze various features such as URL structure, content, and visual elements to develop a robust model. The proposed deep learning model exhibits high accuracy in distinguishing between legitimate and phishing websites, demonstrating its capability to adapt to evolving phishing techniques. The training process involves a diverse dataset of labeled examples. Experimental results on benchmark datasets showcase the model's superiority over traditional methods, marking a significant stride in the ongoing battle against cyber threats. The study contributes to the development of proactive measures to safeguard users against the pervasive and evolving nature of phishing attacks in the digital landscape.

Key Words: Deep Learning, Phishing detection, CNN, Neural Network.

1.INTRODUCTION

Phishing is a type of cybercrime involving technological and social approaches to collect financial and personal data from clients. Such as personally identifiable information, banking and credit card details, and passwords. Hacker obtains user information through various means, including email, forum postings, URLs, instant chats, text messages, and phone calls. Other than email and website phishing, there is also 'vishing' (voice phishing), 'smishing' (SMS Phishing) and several other phishing techniques cybercriminals are constantly arriving. With the rapid development of machine learning, there are more and more applications in the field of cybersecurity and we have proposed a deep learning-based framework to detect phishing links in a real-time web browsing environment. When the URL of the current tab of the browser is predicted to be a phishing link, the current page will receive an obvious warning prompt. The prediction result is obtained by the core prediction service calling a trained model.

2. LITERATURE SURVEY

Phishing websites continue to pose significant threats to online security, exploiting users' trust to steal sensitive information such as passwords and financial data. In response, researchers have conducted investigations into the detection methods and strategies to mitigate the proliferation of these malicious websites. This literature survey aims to synthesize and analyze the advancements in phishing website detection,

In [1] Link Calculator anti-phishing scheme is based on an algorithm designed to extract link characteristics from loading URLs to determine their legitimacy. Unlike the other link-based extraction approaches, the proposed approach introduced the concept of the weighting of the incoming request for its prediction without using the machine learning approach. The weighting concept allows the system to prevent superfluous computations on non-essential links within the parsed page. The advantage of this is to reduce the problems of false-positive and negatives occasioned by other methods where this idea is missing. This is because certain link information within parsed webpages or requests is sufficient to classify them as phishing without loss of generality.

In [2] the aim is to detect malicious URLs using minimum features by applying deep machine learning techniques. As an input web-page URLs are fed into the feature extractor. The feature extractor extracts the requisite features from the sources such as from URL, hyperlink and third party based and transfers them to Information Gain (IG) feature ranking algorithm. The IG algorithm supports in choosing the best performance features. The finest performance features are again trained over Deep Neural Network (DNN) to find out the output status and to differentiate between legitimate and phishing URLs. Here, a robust system based on deep learning neural network (DNN) is proposed which is highly efficient in detecting phishing websites. To train the deep learning model, URL heuristics and third party-based features have been used. Here we have minimized the number of features as compared to Rao and Pais, thereby reducing the dependence on third party-based amenities which is able to attain an accuracy of 99.90%.

In [3] paper they propose the combination of a convolution operation to model the character-level URL features and a deep convolutional autoencoder (CAE) to consider the nature of zero-day attacks. Extensive experiments on three real-world datasets consisting of 222,541 URLs showed the highest performance among the latest deep-learning methods. They demonstrated the superiority of the proposed method by receiver-operating characteristic (ROC) curve analysis in addition to 10-fold cross-validation and confirmed that the sensitivity improved by 3.98% compared to the latest deep model. The main innovation of this study is the introduction of deep anomaly detection to the field of phishing URL detection and achieving the best performance compared to classification-based deep-learning methods by implementing a neural network structure and an operation optimized for URL modelling. The combination of the encoding/decoding structure to facilitate disentanglement between classes and convolution operation optimized for character-level URL characteristics was utilized to define an anomaly score based on the reconstruction error.

In [4] It depicts the architecture of the components of our proposed framework. There are four modules in terms of data collection tasks, machine learning (ML), cloud application, and web browser extension. The data collection module is an independent scheduled task application. The ML module is used for training modules. The web browser extension is a client-side product. The cloud application is built to deal with false alarms and phishing URLs reported by users from the web browser extension. The core process of this framework is mainly divided into the following six steps: The first is to collect and integrate data from various data sources which is divided into two parts, obtaining data from different data sources, then analysing and storing data. The second step is to combine different data sets for machine learning model training and store the trained model in a file system. This research developed six machine learning models, namely Logistic Regression, Support vector machines (SVM), Random Forest, RNN, RNN-GRU, and RNN-Long short-term memory (LSTM). Parameter Configuration. Data loading.Feature extraction: natural language processing, doc t matrix 1 token =1 word. In deep learning URL to list of character (ASCII) modelling: RNN -several hidden layer - LSTM(hidden layer) GRU both enhance RNN optimizer and loss function: dump model to file system. In third step is that the interface for predicting phishing risk calls the trained model to make predictions.

In [5] it uses character embedding, CNN (Convolutional Neural Network) and RF (Random Forest).Firstly, URL data transformed into character vector using character embedding that convert URLs to normalized matrices. The model is trained using transformed data using CNN. The features extracted gets classified in random forests. The first to seventh layers are the input, convolutional, pooling, linear 1, linear 2, linear 6, and output layers. Ensemble classification is the classification of phishing websites can be achieved using multi-level features to improve the accuracy and generalization ability of the classification algorithm. URL features are extracted using the pooling layer, L1 layer, and L3 layer. Each RF classification contains 100 decision trees with a maximum depth of 5 in the child nodes. The proposed approach has the advantage of strong generalization ability, the low-level features in the hidden layer are common and similar for different but related distributed datasets or tasks; these are combined with the low-latitude features in the hidden layer. Third-party service independence, the proposed method relies only on website URL features for detection, without extracting third-party features. Independence of cybersecurity experts reduced required expert function engineering. Language-independent, the approach proposed in this paper is effective for the detection of websites with content in various languages using character-level features.

In [6] a framework for websites classification (phishing or benign) based on Graph Neural Networks. This framework can be considered as an additional layer to GNN architectures. This architecture is divided into two steps namely Pre-Classification and Message Passing. In Pre-Classification, Initially, the graph comprises n nodes(URL), where each node $x_i(1 \leq i \leq n)$ is a vector of d features(URL features) extracted from the corresponding ith URL. x_1 is the root URL node (website) and every node $x_i(1 < i \leq n)$ represent a link coming from x_1. At this first step, a binary classifier is used to predict in a semi-supervised mode whether a node is phishing or benign, for each feature node $x_i(1 \leq i \leq n)$. The classifier is a function g: $R^d \rightarrow B$, where B is the Boolean domain. After this step, the feature matrix X is transformed to a vector X containing respectively zeroes and ones for legitimate and phishing predictions. In Message Passing, the predictions are then, passed through a traditional message passing GNN with h hidden layers, to propagate the information in the graph and learn node embeddings. A pooling method is used to reduce the dimension of node embeddings to a single node. Finally, the resulting vector contains the probability of belonging into each class: phishing or benign.

In [7] The proposed approach consists of two steps: Preparation of Dataset and Network Architecture and Training Parameters. For effective detection of malicious URLs, the dataset should contain recent URLs which are malicious, for recognizing fresh features to train the model. Attackers will change the production of phishing links by anti-phishing regulations and procedures that have been released. Anti-phishing models and algorithms must also be improved based on new phishing data. Furthermore, the training dataset's quantity and validity significantly impact the performance of machine learning based solutions. The performance of deep learning models increases with the variety of content in the training dataset. Hence, it is advised that phishing URLs and legitimate URLs should be extracted from data repositories. The dataset used in this paper consists of numerous legitimate and malicious URLs which are taken from Phish Tank, OpenPhish, and Common Crawl. It consists of 46839 instances, and we merely looked into the text in the URL and extracted features to train the model. The dataset was split into 75% and 25% for training and testing, respectively.

In [8] they propose a hybrid network architecture, called TCURL, which considers both local and global correlations among the characters of URLs. TCURL has two parallel branches, a convolution branch and a transformer branch, and a fusion block used to deal with messages from the two branches. The convolution branch provides sufficient positional information meaning that no extra positional encoding is needed. Through the embedding process, a given URL is first transformed into a matrix with a shape of (L, C), which is then duplicated and fed into the two branches. Next a transformer decoder layer is used to fuse the outputs from the two branches. Finally, we flatten the output and employ a fully connected layer followed by a SoftMax activation to obtain the final result. TCURL represents a hybrid model designed to address the one-dimensional data binary classification problem. Experiments were designed and conducted to analyse the effect of the various elements in a hybrid transformer and CNN model. A dictionary of 67 unique characters (a valid placeholder, 26 lowercase letters, 10 digits, and 30 special characters) is used to convert the URL into a one-hot encoding matrix. The placeholder channel, which indicates whether the current character is valid, is initialized into ones. If the current position contains a valid character, we set the value of the placeholder channel to 0 and the value of this character channel to 1. Sixty-six-character channels are initialized into zeros. We select 200 as the maximum length and discard any URLs with a length exceeding than 200.

In [9] proposed three distinct approaches in order to train the data so that the output can be achieved efficiently. Numerous methods that assist in detecting phishing attacks have been applied by using different, new, and known features such as URL length, frequency of keywords, lexical features, and by incorporating new features.). The first step is data collection and preprocessing. In this method they have used SelectKBest method is used in order to get optimised set. In training and testing step they have done it in three different ways they are LSTM (long short-term memory) is a form of recurrent neural network (RNN) that gains superior results when dealing with time-series data, removing vanishing gradients and long-term dependencies. The architecture of LSTM is made up of a cell and three gates (input, output, and forget). A Convolutional Neural Network (CNN) is a kind of neural network that requires large, labelled data for training. CNNs play a significant role in many problems such as image classification, object recognition, phishing detection, and diagnosis of medical diseases. Input, convolution, pooling, and fully connected layers are the main layers needed to construct a CNN. Accelerating the learning process has led CNN to accomplish great and high results for many problems. LSTM–CNN architecture involves both CNN and LSTM methods in order to make use of the benefits of both methods and accomplish excellent performance. Since CNN and LSTM show high performance in overcoming classification, detection, and recognition tasks to using these three methods for the phishing detection task is promising.

In [10] it develops and compares four models for investigating the efficiency of using machine learning to detect phishing domains. It also compares the most accurate model of the four with existing solutions in the literature. These models were developed using artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and random forest (RF) techniques. Moreover, the uniform resource locator's (URL's) UCI phishing domains dataset is used as a benchmark to evaluate the models. The proposed models were able to detect different types of attacks from the UCI dataset. This dataset was created to build machine learning-based phishing website detection algorithms. It is comprised of extensive properties that span four distinct categories. They designed and extracted characteristics from the following categories: Address Bar, HTML and JavaScript, Abnormal, and Domain. This dataset has 11,055 records, and each record includes 31 characteristics. The characteristics of the collection are identified by names, such as URL Length, Submitting to Email, Shortening Service, Abnormal URL, Having an At Symbol, and Redirect. To increase accuracy, this paper utilized the Minmax normalization feature as a preprocessing step in each proposed model. Normalization is a useful strategy for improving the accuracy of machine learning models, and it is required for some models to work properly.

In [11] evaluate the performance of our model on the PhishTank dataset, which is a widely used dataset for detecting phishing websites based solely on Uniform Resource Locators (URL) features. Binary-categorical loss and the Adam optimizer are used, the accuracy of the k-nearest neighbours (KNN), Natural Language Processing (NLP), Recurrent Neural Network (RNN), and Random Forest (RF) models is 87%, 97.98%, 97.4% and 94.26%, respectively.1D CNN Architecture is used as Model Architecture. A 1D CNN model is a CNN model that only has one dimension, such as text or time series data. An input layer, one or more convolutional layers, pooling layers, and an output layer make up the fundamental building blocks of a 1D CNN model. Input layer receives input data, typically pre-processed and transformed into numerical representations such as tokenized text or time series data. The pooling layers of a 1D CNN model are responsible for lowering the feature dimensionality. The output layer of a 1D CNN model produces the final prediction.1D CNNs can be trained using the SGD back-propagation algorithm or other optimization algorithms such as Adar ad, Adam, and others.

In [12] proposed an enhanced deep learning-based phishing detection mechanism to effectively identify malicious URLs using variational autoencoders. This approach employs OHE (One-hot-encoding) based preprocessing mechanism that converts every URL string into a numerical vector with N x M dimension. M denotes the length of the maximum number of probable characters that might appear in a URL.N denotes the length of the URL, that is, the number of characters in the URL. After preprocessing the URL inputs, instead of directly feeding them to a neural network-based model for classification, our approach adopts a feature reduction/extraction technique to select certain inherent features of a URL to optimise the performance of the classifier. In order to automatically extract salient features from the input URL vector, they developed an Autoencoder (AE) -based feature extraction approach. The Auto encoders is a special form of feed forward neural network which is mainly designed to encode the input into a compressed and meaningful representation and then decode it back such that the reconstructed input is similar as possible to the original one.

The above surveyed methodologies can be selectively employed based on specific needs and circumstances. The below table 1 highlights the methodologies surveyed with their gaps and the remarks obtained.

Table -1: Taxonomy of surveyed methodologies

Authors	Title	Research focus	Remarks
Orunsolu Abioduna ,SodiyaA. Sb, Kareem S.O[1],2020	"Link Calculator –an efficient link-based phishing detection tool"	Link Calculator anti-phishing scheme is based on an algorithm designed to extract link characteristics from loading URLs to determine their legitimacy.	The problem of link evasion by phisher needs further investigation to prevent null return by the proposed scheme.
Md. Faisal Khana, B L Rahab[2], 2021	"Detection of Phishing Websites Using Deep Learning Technique".	It develops and compares four models for investigating the efficiency of using machine learning to detect phishing domains.	Less heuristic features which prevents which the detection of phishing websites faster and more accurately even if the website includes embedded objects.

Seok-Jun Bu, Sung-bae Cho[3],2021	“Deep Character-Level Anomaly Detection Based on a Convolutional Autoencoder for Zero-Day Phishing URL Detection”	Binary-categorical loss and the Adam optimizer are used, the accuracy of the k-nearest neighbours (KNN), NaturalLanguage Processing(NLP),RecurrentNeuralNetwork(RNN),and Random Forest (RF) models	It was optimized for character-level features among the various features constituting URLs.
Lizhen Tang, Qusay H. Mahmoud [4],2021	“A Deep Learning-Based Framework for Phishing Website Detection”	There are four modules in terms of data collection tasks, machine learning (ML), cloud application, and web browser extension.	Training takes time and to process large number of datasets become tedious tasks.
Rundong Yang, Kangfeng Zheng, Bin Wu, Chunhua Wu and Xiujuan Wang[5],2021	“Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning, Sensors”	URL data transformed into character vector using character embedding that convert URLs to normalized matrices. The model is trained using transformed data using CNN.	The model cannot determine whether the URL is active or not, so it is necessary to test whether the URL is active or not before detection to ensure theeffectiveness of detection.
Tristan Bilot, Gregoire Geis and Badis Hammi [6],2022	“A Phishing Website Detection Framework using Graph Neural Networks”	GNNs (Graph Neural Networks) can handle non-Euclidean data with complex relations between objects.	This method only relies on the HTML content, which could be easily stolen from benign websites in order to build perfect website copies.
Aman Rangpur, Tarun Kanakam and Dhanvanthini P [7],2022	“Phishing Detection Using Deep Recurrent Neural Networks”	The features captured from the URL are fed to the LSTM layer with an orthogonal recurrent initializer.	The main limitations are the absence of comparisons between certain studies.
Chenguang Wang, Yuanyuan Chen[8],2022	“Exploring hybrid transformer and convolutional neural network on phishing URL detection”	Hybrid network architecture, called TCURL.TCURL has two parallel branches, a convolution branch and a transformer branch, and a fusion block used to deal with messages from the two branches.	Layer normalization may not be suitable for architectures that have a large number of parameters or are very deep, as it may not be able to capture the full complexity of the activations.
Zainab Alshingiti ,RabeahAlaqel ,Jalal Al-Muhtadi ,Qazi Emad Ul Haq, Kashif Saleem and Muhammad Hamza Faheem[9] ,2023	“A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN”	Three distinct approaches are used in order to train the data so that the output can be achieved efficiently. Numerous methods that assist in detecting phishing attacks have been applied by using different, new, and known features such as URL length, frequency of keywords, lexical features, and by incorporating new features.	The approach has that the model does not check the status of the URL of the website, i.e., whether the website is active or not, which impacts the results.
ShouqAlnemari, Majid Alshammari[10],2023	“Detecting Phishing Domains Using Machine Learning”	The proposed models were able to detect different types of attacks from the UCI dataset.	It requires features of the URL to be manually extracted which depends on third-party services to obtain certain important features.

Eman Abdullah Aldakheel, Mohammed Zakariah, Ghada Abdalaziz Gashgari, Fahdah A. Almarshad and Abdullah I. A. Alzahrani[11] 2023	"A Deep Learning-Based Innovative Technique for Phishing Detection in Modern Security with Uniform Resource Locators, Sensors"	Evaluate the performance of our model on the PhishTank dataset, which is a widely used dataset for detecting phishing websites based solely on Uniform Resource Locators (URL) features. Binary-categorical loss and the Adam optimizer are used, the accuracy	It requires crawling &analysing URL's which may not be suitable for real-time detecting.
Manoj Kumar Prabakaran , Parvathy Meenakshi Sundaram ,Abinaya Devi Chandrasekar[12],2023	"An enhanced deep learning-based phishing detectionmechanism to effectively identify malicious URLs usingvariational autoencoder"	This approach employs OHE (One-hot-encoding) based preprocessing mechanism that converts every URL string into a numerical vector with N x M dimension. This model suffers from the problem of generalisation.	VAE can suffer from posterior collapse, where the encoder ignores the input data and outputs a trivial latent space, leading to poor representation and reconstruction.

3. CONCLUSIONS

This paper outlined a survey of phishing websites detection using Deep learning. It also outlined the different approaches and techniques in various survey papers as reference points by various authors considering its advantages, and also some key challenges are discussed here. After studying various Machine Learning algorithms, it was found that LSTM and CNN algorithms will produce accurate result. This survey effort will provide a better understanding of algorithms which will be used to develop the model for detection of phishing websites.

REFERENCES

[1] Orunsolu Abioduna,Sodiya A.Sb, Kareem S.O, Link Calculator – an efficient link-based phishing detection tool,Acta Informatica Malaysia, Volume 4, Issue 2, September 2020, p. 37-44.

[2] Md.Faisal Khana, B L Rahab, Detection of Phishing Websites Using Deep Learning Technique, Turkish Journal, Volume 12, Issue 10, April 2021, p.3880-3892 .

[3] Seok-Jun Bu, Sung-bae Cho, Deep Character-Level Anomaly Detection Based on a Convolutional Autoencoder for Zero-Day Phishing URL Detection, Electronics (Korea), Volume10, Issue 12, June 2021.

[4] Lizhen Tang, Qusay H. Mahmoud, A Deep Learning-Based Framework for Phishing Website Detection, IEEE, Volume 10, December 2021, p. 1509 – 1521.

[5] Rundong Yang, KangfengZheng, Bin Wu, Chunhua Wu and Xiujuan Wang, Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning,Sensors (Basel),doi: 10.3390/s21248281, December 2021.

[6] Tristan Bilot, Gregoire Geis and Badis Hammi, PhishGNN: A Phishing Website Detection Framework using Graph Neural Networks, SECRYPT At Lisbo, DOI:10.5220/0011328600003283, July 2022.

[7] Aman Rangpur, Tarun Kanakam and Dhanvanthini P, Phish-Defence, Phishing Detection Using Deep Recurrent Neural Networks, Cornell University,Volume 4, September 2022 .

[8] Chenguang Wang, Yuanyuan Chen, TCURL, Exploring hybrid transformer and convolutional neural network on phishing URL detection, Knowledge-Based Systems, Volume 258, Issue C, December 2022.

[9] Zainab Alshingiti ,Rabeah Alaqel ,Jalal Al-Muhtadi ,Qazi Emad Ul Haq,Kashif Saleem and Muhammad Hamza Faheem, A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN, Vol 23,Jan 2023 .

[10]ShouqAlnemari, Majid Alshammari, Detecting Phishing Domains Using Machine Learning, Applied Sciences (2076-3417), Vol. 13, Issue 8, April 2023, p4649. 16p.

[11] Eman Abdullah Aldakheel, Mohammed Zakariah, Ghada Abdalaziz Gashgari, Fahdah A. Almarshad and Abdullah I. A. Alzahrani,A Deep Learning-Based Innovative Technique for Phishing Detection in Modern Security with Uniform Resource Locators, Sensors, Vol. 23, Issue 9, April 2023.

[12] Manoj Kumar Prabakaran, Parvathy Meenakshi Sundaram ,Abinaya Devi Chandrasekar,An enhanced deep learning-based phishing detectionmechanism to effectively identify malicious URLs usingvariationalautoencoder,IET Information Security, Volume 17, Issue 3, May 2023, p. 423 – 440.

[13] Wei Wei,qiaoke Jakub Nowak,Marcinkorytkowski, Rafat Scherer, Marcin wonxniak, Accurate and fast URL phishing detector, Elsevier, Volume 198, September 2020.

[14] Mohamed A. El-Rashidy, A Smart Model for Web Phishing Detection Based on New Proposed Feature Selection Technique, Menoufia J. of Electronic Engineering Research (MJEER), Vol. 30, No. 1, Jan.2023.

[15] Ahmet Selman Bozkir, Firat Coskun Dalgic, Murat Aydo, GramBeddings: A New Neural Network for URL Based Identification of Phishing Web Pages Through N-gram Embeddings, Elsvier, Volume 124, January 2023.