JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Phishing Detection System based on Deep Learning

Raj Anant[1],S. S. Bhosale[2],Ayesha Qureshi[3], Chandan Yadav[4] S. N. Kamble[5], C. G. Patil[6]

Department of E&TC,SKNCOE,SPPU,Pune

[1]rajanant424@gmail.com

[2]sonali.bhosale_skncoe@sinhgad.edu

[3]ayeshafaryal222@gmail.com

[4]ychandan590@gmail.com

[5]sarika.kamble_skncoe@sinhgad.edu

[6]chandrashekhar.patil_skncoe@sinhgad.edu

Abstract— **In 21st Century, Our generation is increasingly moving toward digital future and integrating ourselves in the world of internet but this has led to one of the major problem in cyberspace is how to maintain cyber security. Cyber security is often threatened by phishing attacks. In recent years there has been increase of over 40% in phishing attacks.Phishing is a type of cyber attack which is used to steal the sensitive personal information of unaware victim. These attacks have increased because the value of ones personal information has increased rapidly. To prevent this a system is proposed for phishing websites detection by utilizing Deep learning methods. Convolutional Neural Networks (CNN) is one of the method of deep learning.Dataset containing legitimate and phishing website URLs is going to be used in for proposed system. With the CNN based model excellent accuracy can be achieved for distinguishing phishing websites apart from legitimate websites. The deep learning methods used to prevent phishing attacks can help to reduce cybercrime. The system that is proposed in this paper has shown accuracy of 94.96% which is quite good. Along with it a web application has been proposed which utilizes trained CNN model.**

Keywords— **Deep Learning, CNN, Phishing, Cyber attack, URL, Personal Data, Classification.**

I. INTRODUCTION

URL (Uniform Resource Locator) is generally referred as web address. Tim Berners- Lee defined Uniform Resource code in 1994. URL has many parts some of them are mandatory and some are optional. Parts of URL are domain, subdomain, path and parameters etc. URL is the address of a unique resource on web. An URL can be used to get resource behind it by just simply typing it in address bar of browser.As 21st century is progressing our generation is entering the age of internet. Many areas our life are being transformed due to digitization and use of internet. With the use of internet there are many advantages but there are also disadvantages. One of the major problem facing today is phishing attacks. Phishing is a kind of social engineering based attack which can grant cybercriminals the ability to steal credentials, distribute ransomware, carry out financial fraud and theft.Phishing is a cyber crime in which the perpetrator attempt to steal sensitive information like user name, passwords, debit/credit card information etc.by sending a mail or SMS to the users that look similar to the URL .

Phishing is a threat leading to loss of a lot of money.There are two types of phishing one is Spear phishing and second is Clone phishing. In spear phishing, attackers try to target personal information which can result into the higher chances to succeed in phishing attack. In the case of clone phishing, attackers use emails which look like they are from legitimate sources but attackers alter some of its content to add malicious content like links for phishing site.

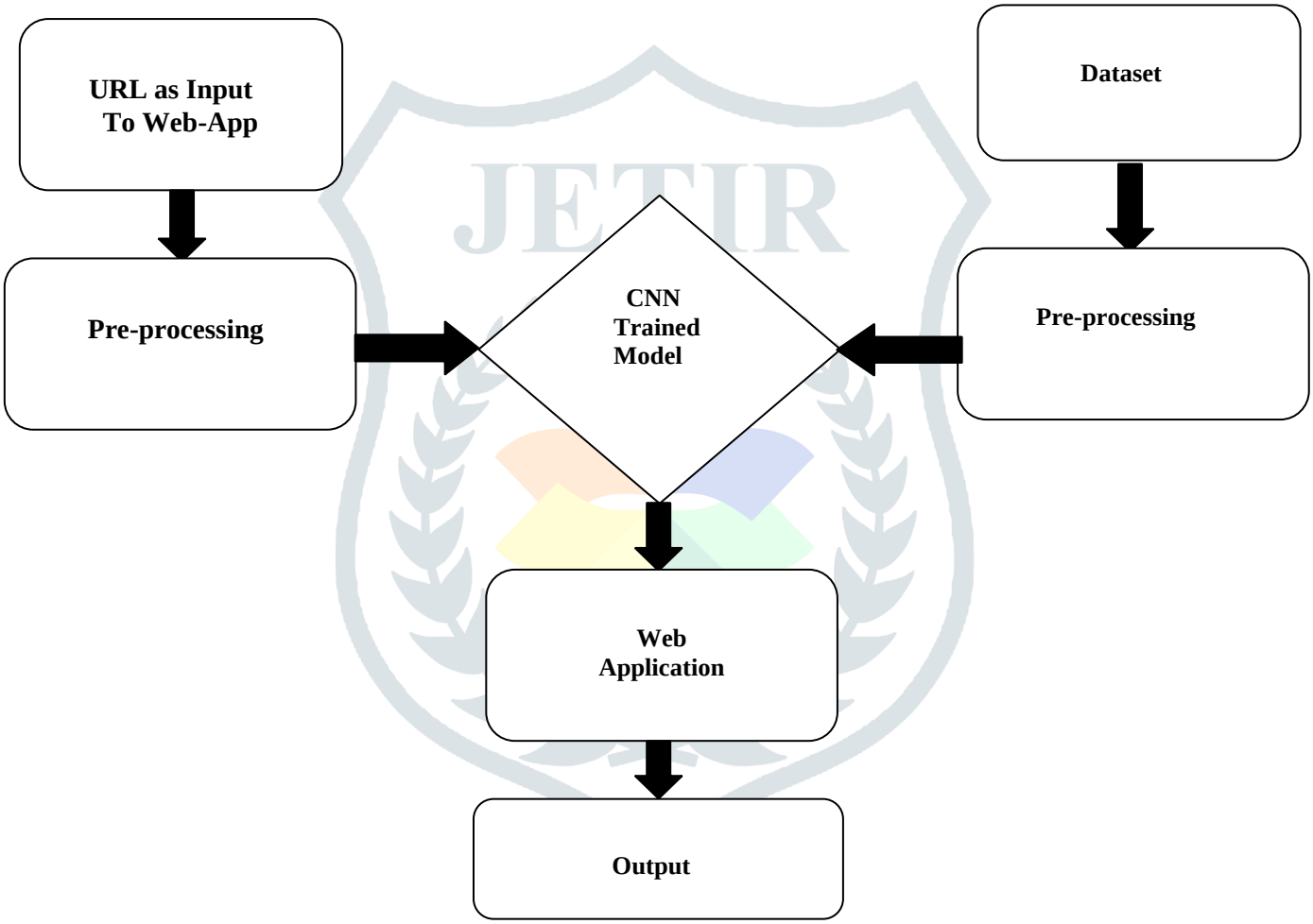
Clone phishing is mostly used for purpose of fraud.[7] Some of the older phishing attacks involve sending SMS to the mobile phone of the user. In this type of attack, perpetrator pretends to be from bank or from other government authority. This attacks are easily recognizable because of the awareness about them. As world is entering in digital age, there has been many approaches that are used to detect the phishing websites like approach based on blacklists and whitelists , visual similarity of webpage and URL and content features of website.First approach which is based on whitelists and blacklists which is dependable on existing database to detect existing phishing websites. Because of this deficiency,this approach is unable to detect new phishing websites unless database is updated. Now considering second approach which is based on visual similarities between web-pages. In this approach visual features are extracted from web-pages and theses extracted features are compared to get classification for that website . This approach is also flawed because if there is any distortion in webpage contents then is difficult to extract visual features which can lead to wrong classification. Third approach is most reliable one, in this approach the URL and web content features are used for classification of phishing websites. In past many machine learning based algorithms like Random Forest, SVM (Support Vector Machine), Linear Regression are used for detection of phishing websites.[1]To prevent phishing , new technology and methods should be utilized . Deep learning presents that opportunity.

II. LITERATURE REVIEW

As for the related work various papers are referred . These papers have data related to various techniques which are used for detecting phishing websites.In this paper, authors have proposed a machine learning based system which has been determined by utilizing best machine learning algorithms. First dataset is collected from UCI Machine Learning Repository. This dataset is used for training and testing of three machine learning algorithms which are SVM classifier, Neural Network Classifier with backpropagation and Random Forest classifier. The backpropagation is used in NN classifier is to reduce error in final result. SVM classifier showed best accuracy among all classifiers used in testing . The accuracy of NN classifier, Random Forest classifier and SVM classifier is 87.34% , 89.63% and 89.84% respectively. A chrome extension is created using trained model.[3].In this paper authors have given information about previous conventional methods to detect phishing websites like blacklist and whitelist approach , Heuristic Approach and visual similarity approach. To avoid the weaknesses of these approaches authors have used machine learning approach. In this paper authors have suggested a system which includes blacklist whitelist approach with the machine learning algorithms. So in this system a input url is given to the system which is firstly compared to whitelist and backlist approach if the given url is not in that database then this input url is further moved for feature extraction. There are two methods are used for feature extraction one is based on heuristics approach and other one is based on visual similarity approach. These features are utilized by machine learning algorithms, for this system authors have used multiple machine learning algorithms like Logistic Regression , Decision Tree and Random Forest and compared these algorithms with each other. If the given input url is detected as phishing url then system will block that url. In this system logistic regression has given 96.23% accuracy , Decision Tree has given 96.23% accuracy and Random Forest has given 96.58% accuracy in detecting phishing URLs.[17]In this paper Authors have used deep learning methods to detect phishing websites. Convolutional Neural Network (CNN) based model is used identify URLs of phishing websites from URLs of legitimate URLs. Various URL and web content features have been used in proposed system. Dataset containing instances from phishing as well as legitimate URLs is used to train and test the CNN model. Sigmoid activation function is used for final classification purpose. ReLU (Rectified Linear Units) Activation function is to mitigate vanishing and exploding gradient issues. CNN model is tested on various measures like accuracy , precision, recall and F1 score. This CNN model uses 1D CNN (one dimensional) layers. This model is compared with other models based on machine learning algorithms like Naive Bayes , SVM , Random Forest and Bayes Net. This proposed system have given upto 97% accuracy in detecting phishing URLs. Precision is upto 0.970. Recall and F1 score of proposed system are 0.982 and 0.976 respectively. This deep learning approach performs better than other machine learning algorithms. As this modern technology can be used to avoid serious problems like phishing .[1]vTo detect the phishing URLs authors have designed a deep learning based system. This system uses CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) for its base model. In order to use deep learning models, a dataset is required. For this system authors have created their own dataset containing 97,000 phishing URLs and 97,400 Legitimate URLs. Phishing URLs are labeled as 0 and legitimate URLs

as 1 in dataset. Then dataset is pre-processing is done .In pre-processing the number of characters in URL is limited to 75 and excessive characters are removed. The deep learning model consists CNN layers and RNN layers. The URL which is preprocessed is passed to an embedding layer,which is first hidden layer of this model. Embedding layer is used to reduce the feature space. The dimension of embedding layer is 75*32. Then three convolutional layers are with 256 filters and kernel size of 5,6 and 7 are used. The ReLU (Rectified Linear Units) activation function is used. LSTM layer is used which has 32 neurons in it and tanh activation function is used . The last layer consists of a fully connected layer with sigmoid activation function. Adam optimizer was used with a learning rate of Le-4 . Model was tested at 200 epochs.After training and testing of model, this model achieved excellent accuracy in detection of phishing URLs.[2] These papers give brief idea about phishing.

III. METHODOLOGY



3.1 Block Diagram of Deep Learning Based Phishing Detection System

A) Dataset Collection :

For this system which is designed to detect phishing URLs , this system is based on deep learning approach. For Deep Learning model, Dataset is required to train this model. For this system large dataset is required to train and test proposed model. The data for this dataset is acquired from various sources like Phishtank website and Kaggle.com. Dataset that is acquired has around 200000 URLs of different websites. From those 200000 URLS , 100000 URLs are legitimate URLs and 100000 are phishing URLs. This dataset has been used to train our deep learning based model

B) Pre-processing :

Dataset which is used needs some modifications in order to make it suitable to use with CNN model. For that pre-processing of data set is done. Data pre-processing is technique that is used to convert the raw data into a clean data.The URL length in this dataset has been limited to 80 characters. This dataset will be used for training an testing of our CNN based deep learning model.[2]

C) CNN Model :

This system is based on deep learning. Deep learning involves various types of methods. Deep learning is a type of Artificial intelligence and machine learning that copies the way humans gain some types of knowledge. For this System, CNN (Convolutional Neural Network) is used, CNN is a type of Neural Networks. For Convolutional Neural Network two types of layers are used those layers are convolutional layers and maxpooling layers. Convolutional layer is used to extract the optimal features and max pooling layer is used to reduce the dimensions of convolutional layer features. Proposed 1D CNN model for dataset which is used to train this model. Following 1D CNN model can detect new phishing websites effectively. This system utilizes convolutional layers and maxpooling layers.In model first layer is embedding layer after that there are 4 one dimensional convolutional layers which have kernel size of 5,6,7and 8 respectively those layers a maxpooling layer for each convolutional layer. Padding is same for all layers. Dropout is given between the layers in order to avoid over fitting of model. Output from last convolutional layer is flattened and given to last layer. Last output layer is a dense layer which uses sigmoid activation. There are two functions used in this model, ReLU (Rectified Linear Units) and Sigmoid Activation Function. Adam optimizer is used to optimize this model. This model is trained using 80% of dataset while remaining 20% is used for testing. This model is trained at 220 epochs with batch size 40. This model gives accuracy of 94.96%.[2]

Table 1. Parts of URL

Parts Of URL	Name of Part
https://	Scheme/ protocol
www.	Subdomain
example.	Domain
Co.uk	Top level domain
:444	Port no.
/blog/article/search	path
?	Query string
docid=730&hl=en	Parameter
#daytwo	Fragment

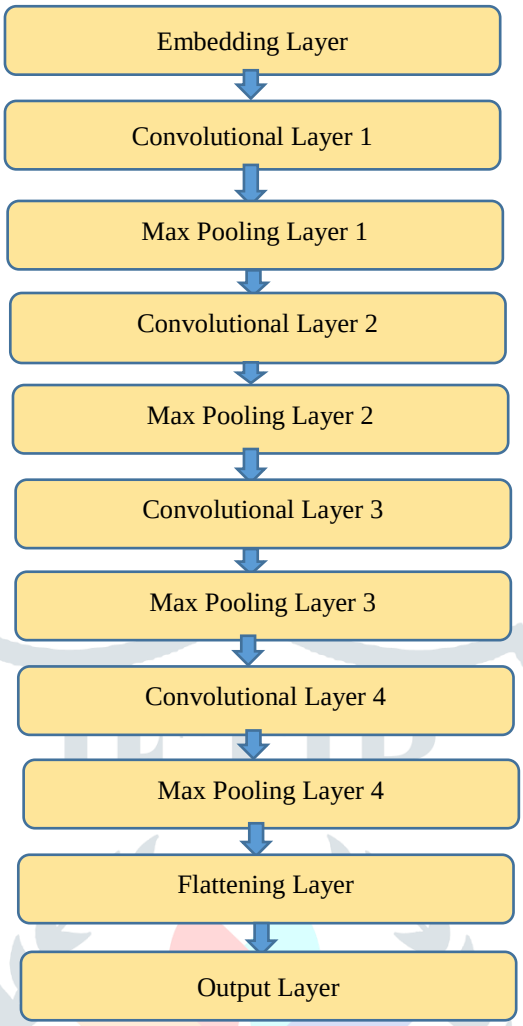


Fig.1. Architecture of 1D CNN Model

IV. IMPLEMENTATION

For the development of proposed system various Libraries are used which involve.

Keras : Keras is a python library which is open source and mostly used for to have interface with artificial neural network. Keras is high level API of Tenserflow. The development of Keras was done with focus on enabling rapid experimentation. Keras is simple and it is not simplistic. Keras has ability to adopt the principle of progressive disclosure of complexity which makes simple workflows to happen quick and easy. One of the major key points of Keras is that it gives industry level strength performance and scalability. It is widely used by various big companies and organizations. Keras can run on TPU or on GPU clusters . The Keras models can be exported to run on mobile devices or on browsers.

Tensorflow : Tensorflow is a open source library which is used primarily in works related to building neural networks. Tensorflow was developed by Google's Google Brain Team to be used in internal research and production. It is primary software tool of deep learning. Tensorflow uses flow graphs in building of models. It gives developers a ability to develop large models consisting many layers. It helps to scale models to a large size models. Tensorflow is mostly used with python as programming language but it can also be used with Javascript and C++. The flexibility of Tensorflow allows us the easy deployment of computation across various platforms. Dataflow graphs are used to express the computations of Tensorflow. The set of optimizers which includes ADAM, ADAGRAD and SGD are offered by Tensorflow. Tensorflow helps into loading of data which is used for training of model. Tensorflow is used in wide variety of fields like social media ,medical , Retail sector and search engines.

Pandas : Pandas stands for Python Data Analysis Library. Pandas is library which is designed for python programmers, this library is mainly used for data analysis and its manipulation. It is a open source library. It offers various operations and functions for data analysis and modifications. It offers features like data alignment , reshaping and pivoting of datasets, column insertion and deletion and also dataset joining and merging.

V. EXPERIMENTATION

Various URLs are tested for their classification. First the testing for the URL of a legitimate website is done , for which this system predicted its a legitimate URL and after that the testing is done for a phishing URL which is taken from Phishtank website for this URL, system predicted it as phishing URL which is correct.This way experimentation is done for this system.

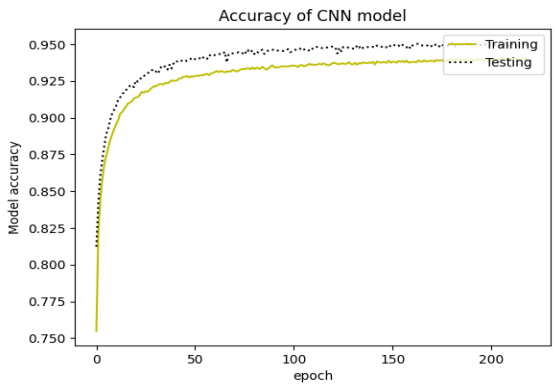


Fig. 2. Graph of Accuracy of Trained Model

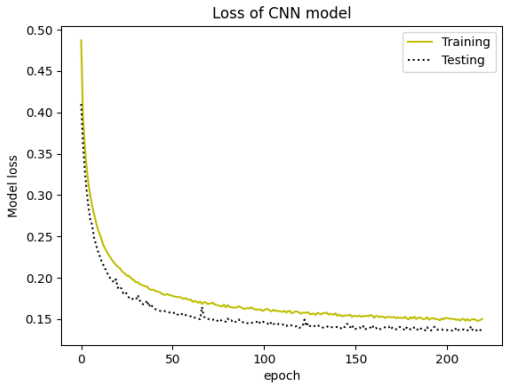


Fig.3.Graph of Loss of Trained Model

Table 2. Various Results about Model

Name of Metric	Value
Accuracy	**0.95**
Precision	**0.95**
Recall	**0.95**
F1 score	**0.95**

In above figures and table the results of trained model are included which involve the graph of accuracy and graph of loss. These graphs show that first accuracy is low while training and then it increases as training of model goes forward. As for the loss , it is huge during initial phase of training and then it decreases. Table 2 indicates various results of metrics used for checking of trained model.

VI. CONCLUSION

The value of personal sensitive data has increased in recent times. This increase is because of the integration of internet in people's lives. Services like online payment methods, bank account and social media accounts are essential in today's world. The increase in use of internet has given rise to cyber attacks which includes phishing attacks.

Phishing attacks are generally targeted for stealing the personal information from victims like usernames , passwords and debit /credit card credentials. These attacks are carried out by phishing websites. To prevent these phishing attacks, a deep learning based system which can detect these websites from their URL is developed. This system utilizes deep learning methods like Convolutional Neural Networks (CNN). This deep learning based approach is better at detecting new phishing websites than the previous machine learning approaches. CNN based model which is used in this project has given excellent accuracy of 94.96% . The development of web application is also done, this web application is light and fast. This project has a potential to be used in many areas like in web browsers and other Internet using Applications. This ensure security of personal and commercial data quite efficiently.

REFERENCES

[1] Suleiman Y. Yerima & Mohammed K. Alzaylaee "High Accuracy Phishing Detection Based on Convolutional Neural Networks" (ICCAIS 2020).

[2] Yazhmozhi V. M. , B. Janet & Srinivasulu Reddy " Anti-phishing System using LSTM and CNN" (2020 IEEE).

[3] Smita Sindhu , Sunil Parameshwar Patil, Arya Sreevalsan and Faiz Rahman "Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation" IEEE(2020).

[4] Farid Tajaddodianfar, Jack w. Stokes and Arun Gururajan " Texception: A Character/Word-Level Deep Learning Model For Phishing URL Detection" IEEE(2020).

[5] Athulya A. A. and Praveen K. " Towards The Detection of Phishing Attacks" IEEE (2020).

[6] M. A. Adebowale, K.T. Lwin & M. A. Hossain "Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection" (2019 IEEE).

[7] Happy Chapla , Riddhi Kotak & Mittal Joiser "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier" (IEEE 2019).

[8] Mahdieh Zabihimayvan and Derek Doran " Fuzzy Rough Set Feature Selection to Enhance Phishing Detection" IEEE (2019).

[9] Rabab Alayham Abbas Helmi, Chua Shang Ren, Arshad Jamal and Muhammad Irsyad Abdullah "Email Anti-Phishing Detection Application" IEEE(2019) and ICSET(2019).

[10] S. Roopak, Athira P. Vijayaraghavan and Tony Thomas "On Effectiveness of Source and SSL Based Features for Phishing Website Detection" IEEE(2019).

[11] Ivan Ortiz Garces, Maria Fernada Cazares and Roberto Omar Andrade "Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture" IEEE(2019).

[12] Yongjie Huang, Jinghui Qin and Wushao Wen " Phishing URL Detection Via Capsule-Based Neural Network" IEEE(2019).

[13] Merlin. V. Kunju, Esther Dainel, Heron Celestie Anthony and Sonali Bhelwa " Evaluation of Phishing Techniques Based on Machine Learning" IEEE(2019) and ICCS(2019).

[14] G.Jaspher Willsie Kathrine, Paradise Mercy Praise, A. Amrutha Rose and Eligious C. Kalaivani " Variants of Phishing Attacks and Their Detection Techniques" IEEE(2019) and ICOEI(2019).

[15] Akihito Nakamura and Fuma Dobashit " Proactive Phishing Sites Detection" IEEE(2019).

[16] Peng Yang, Guangzhen Zhao and Peng Zeng " Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning" IEEE (2019).

[17] Vaibhav Patil , Pritesh Thakkar & Chirag Shah " Detection and Prevention of Phishing Websites using Machine Learning Approach" (2018 IEEE).

[18] Shraddha Parekh, Dhawani Parikh & Srushti Kotak " A New Method for Detection of Phishing Websites: URL detection" IEEE (2018) & ICICCT (2018).

[19] Muhammet Baykara and Zahit Ziya Gurel " Detection of Phishing Attacks" IEEE (2018).

[20] Tianrui Peng, Ian G. Harris and Yuki Sawa " Detecting Phishing Attacks Using Natural Language Processing and Machine Learning" IEEE(2018).

[21] Mehek Thaker, Mihir Parikh , Preetika Shetty, Vinit Neogi and Shree Jaswal " Detecting Phishing Websites using Data Mining" IEEE(2018) , ICECA (2018).

[22] Xiaoqing Zhang, Dingee Shi, Hongpo Zhang, Wei Liu and Runzhi Li " Efficient Detection of Phishing Attacks with Hybrid Neural Networks" IEEE(2018).