International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

URL-Based Phishing Detection Using Machine Learning and Deep Learning

Prof. Alfred Thomas , Sreelekshmi K U
Assistant Professor: Rajiv Gandhi Institute of Technology,
Kottayam*Student: Rajiv Gandhi Institute of Technology, Kottayam

Abstract **—Website phishing is one of the main threats to the present cyber security world. It is a cyber-fraud in which an imposter will be faking a legitimate website in its content such as the website of a bank or any other organization. The fake one will have the complete features of the original website including color theme, logo, texts, and appearance so distinguishing the fake one and legitimate one will be challenging. Phishing can be detected in many ways and using many techniques. URL-based Phishing website detection using Machine Learning (ML) and Deep Learning (DL) is one of the most accurate techniques among them. This project is using ML algorithms such as Random Forest to detect phishing and legitimate websites and comparing the performance with Deep Learning models such as DNN (Deep Neural Networks) and LSTM (Long Short-Term Memory) and Bi-directional LSTM. Data of both legitimate and phishing URLs will be collected using web scraping from the internet and websites like www.phishtank.com instead of using already available datasets. A number of features such as HTML-based features, Domain-related features, and Address bar-related features will be extracted from the raw URLs collected from the internet. Machine learning algorithms are found to be performing very accurately, especially in cases like cyber security where high accuracy performance is demanded. So, machine learning algorithms such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), etc and Deep Learning models such as ANN (Artificial Neural Networks) and DNN are used as the models. For training the LSTM model, the URL data will be processed using Natural Language Processing techniques. The performance of these models is analyzed using performance evaluation measures and metrics such as accuracy, precision and other scores, and the outputs and results will be tabu- lated. The whole system will be converted into a desktop app using Python Tkinter GUI framework.**

Keywords**—Uniform Resource Locator, Artificial Neural Network, Natural Language Processing, Long Short-Term Memory, Deep Neural Network, Deep Learning, Machine Learning, Support Vector Machine, K Nearest Neighbors**

I. Introduction

Identity theft can be a crime where the perpetrator sends a false e-mail, or URL of a website that appears to come from a legitimate source or credible organization, requesting a personal certificate such as bank ownership, username, number, address, capital card details, and more. Fraudulent emails and websites often look strangely legitimate, and even a website whenever a net user is asked to enter personal data, and it sounds fair. Phishing scams are circulating via e-mail, SMS, instant messengers, social networking sites, VoIP, etc., however e-mail that spreads these attacks and phishing scams is achieved by visiting the e-mail link. In addition, the criminal attacks of identity theft are changing dramatically these days. The crime of stealing sensitive information still poses a severe threat to security, and a large number of internet users fall victim to this scam. Moreover, such attacks do not only cause problems for internet users but also for companies that provide online financial services. That is because when users fall victim to such a crime of identity theft, an online service provider often loses its reputation and economic damage.

Phishing costs internet users billions of dollars per year. It refers to luring techniques used by identity thieves to fish for personal information in a pond of unsuspecting Internet users. Phishers use spoofed email, phishing software to steal personal information and financial account details such as usernames and passwords. Social engineering schemes use spoofed emails, pretending to be from legitimate businesses and agencies, designed to lead consumers to phishing websites that trick recipients into revealing confidential data such as usernames and passwords.Technical schemes install malicious software onto computers, to steal credentials directly, often using systems to intercept consumer's online account usernames and passwords. A cybercrime website hits online businesses, banks, Web users, and governments, so it has become a national security issue. It is necessary that this attack be detected early. However, it is difficult to see these attacks because of the new methods used by criminals to steal sensitive information to commit crimes. For successful criminal detection of identity theft to be

achieved, it must be obtained with the highest accuracy and in the shortest possible time. The most common method of detecting identity theft involves black listing and white listing. Criminals use URLs for stealing sensitive information can be obtained with the concept of machine learning, which can be used continuously to prevent such attacks. First, machines used to follow instructions given by man, but now people can train a machine to learn from previous data, build a prediction model and perform very fast, and this is known as machine learning. It is basically the use of tools and technologies that can be used effectively and efficiently. Machine learning is used to make one's work easier, faster, and more accessible by learning from past data and working efficiently now.

The criminals, who want to obtain sensitive data, first create unauthorized replicas of a real website and e-mail,usually from a financial institution or another company that deals with financial information. The email will be created using logos and slogans of a legitimate company. The nature and format of Hypertext Mark-up Language makes it very easy to copy images or even an entire website. While this ease of website creation is one of the reasons. The Internet has grown so rapidly as a communication medium, it also permits the abuse of trademarks, trade names, and other corporate identifiers upon which consumers have come to rely as mechanisms for authentication. Phisher then sends the "spoofed" emails to as many people as possible in an attempt to lure them into the scheme. When these emails are opened and when a link in the mail is clicked, the consumers are redirected to a spoofed website, appearing to be from the legitimate entity.

II. Related Works

Jain, A.K. and Gupta, B.B., observed that attackers steal sensitive information such as personal identification number (PIN), credit card details, log in, password, etc., from Internet users. In this paper, the author has proposed a machine-based reading program based on the Uniform Resource Locator (URL) features. To test the performance of the proposed system, the author has taken 14 features in the URL to find the website as sensitive identity theft or nonsensitive identity theft crime. The proposed approach is being trained using sensitive identity theft and official URLs with SVM and Na¨ıve Bayes divisions. Test results show 90 Purbay M., and Kumar D. examined multiple machine learning methods for obtaining URLs by analyzing various URL parts using machine learning and in- depth learning methods. The authors discussed different ways of reading surveillance to identify criminal URLs that steal sensitive information based on dictionary, WHOIS architecture, PageRank, traffic level information, and key page layouts. Learn how the volume of different training data influences the accuracy of class dividers. The research includes Vector Support Machine (SVM), K-NN, random forest classification (RFC), and Artificial Neural Network (ANN) classification methods. Gandotra E., and Gupta D conducted a comparative study on machine learning based on the outputs and operational selections. They studied 6157 incorrect pages and found several Machine learning methods have been used for best results. The latter job selection method is used to maximize model performance. The random forest algorithm gained accuracy before and after selecting features and significantly increased construction time. Experimental results have shown that using a selective method of machine learning algorithms can improve the performance of classification models to detect the crime of stealing sensitive information without reducing its effectiveness. Hung et al. developed URLNet, a Convolutional Neural-Network (CNN) based on an in-depth reading framework that uses alphabetical characters and URLs to capture semantic information to distinguish malicious and dangerous URLs. Their work has demonstrated a promising approach to URL acquisition through in-depth reading. They discussed the limitations of features obtained using the word bag and mathematical features such as the length of the different segments in the URL. Use CNN to get useful structural information for URLs with two separate databases generated by letters and URL names. Word Level CNN is similar to CNN characters level except that convolution operators are used in words. Database URLs are collected from VirusTotal. They have created a feature set using a training corporation with all the unique words as a dictionary. This method provides another way to separate malicious URLs by capturing a few semantic information via URLNet, which are existing methods based on word tag elements that could not. It provides an essential escape from the AUC beyond the foundation. Kumar J. et al., The author investigated how the URLs of identity theft can be categorized in a set of URLs containing incorrect URLs. They discuss random signal engineering, feature extraction using host-based analysis, and mathematical analysis. In a comparative study, several class dividers were used and found that the results for all the different class dividers were almost identical. Authors argue by suggesting an easy way to remove functionality from URLs with simple common words. Other factors can be tested that lead to better results. The database used in the study includes old URLs. Thus, there is a possibility of inefficiency. Hassan Y. A. Abutair et al. introduce the

CBR-PDS. It relies heavily on the CBR method as the core component. The system is flexible and flexible as it can quickly adapt to detecting cybercrime attacks with a small amount of data set compared to other detectors requiring extensive training in advance. Authors test their system using different scenarios for 572 phishing and official URLs. Studies show that the accuracy of the CBR-PDS system exceeds 95.62 Rao R S and Paris Ali have proposed a clever way to detect sensitive identity theft in 2015 called PhishShield, a desktop application that focuses on detecting identity theft using URLs and website content of sensitive identity theft websites [9]. The features released by PhishShield are minimal link links, zero links in the HTML body, copyrighted content, title content, and website logos. PhishShield is faster, more accurate, and has a broader range of access to criminal websites to steal sensitive information compared to the blacklisting and whitelisting system. However, detection effectiveness decreases when the attacker understands the heuristic process and can successfully pass the heuristic filter. Aljofey et al. demonstrated a solution to detect identity theft using Convolutional Neural Network (CNN) character-based analysis of website URLs using a model based on fast-paced learning solutions. Their model does not include using services from third parties or retrieving content from the targeted website. They capture sequence patterns and URL unit information without the need for an idea about the crime of stealing sensitive information in advance. Consecutive patterns quickly classify the original URL. They also compare different traditional and in-depth machine learning models. Feature sets include handicrafts, embedded characters, Term's Frequency-Inverse Document Frequency (TF-IDF), and calculation vector features at the character level. The experimental results of Aljofey et al. have brought 95.02 AlEroud A and Karabatis G used a productive argument network to classify URLs into categories and bypass criminals to steal sensitive information based on restricted lists. In addition, the researchers argued that the system could surpass both simple ML acquisition strategies and novice ones. Jitendra Kumar, A. Santhanavijayan, B. Janet, B.S. Bindhumadhava, and Balaji Rajendran published "Phishing Website Classification and Detection Using Machine Learning". By making use of lexical structure URL to classify url into different parts and identify the Url whether the given url is phishing url or not. In, this paper, they have compared different machine learning techniques for the phishing URL classification task and achieved the highest accuracy of 96Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani published "Detecting Phishing Websites Using Machine Learning". The system acts as an extra functionality to a web browser as an extension that mechanically notifies the user once it detects a phishing website. The system is predicated on a machine learning method, notably supervised learning. They've selected the Random Forest technique because of its sensible performance in classification. The focus will be on the features combination that we get from Random Forest (RF) technique, as it has good accuracy, is relatively robust, and has a good performance. Recently, there have been several studies that are trying to solve the phishing problem. They can be classified into four types: blacklist, heuristic, content analysis, and machine learning techniques. The blacklisting technique compares the URL with an existing database that contains a list of phishing website URLs. Because of the rapid increase of such phishing attacks, the blacklist approach has become more inefficient in checking whether each URL is a phishing website or not, and this kind of delay can also lead to zero-day attacks from these new phishing sites. Fuma Dobashi, Akihito Nakamura, published "Proactive Phishing Sites Detection,". In this paper, he compared phishing mitigation techniques, such as blacklist, heuristics, visual similarity, and machine learning and concluded that these techniques have limitations in dealing with zero-hour attacks and proactive detection of phishing websites. The authors proposed suspicious URL's generation and to predict likely phishing sites from the given legitimate brand domain name and scores and judge suspects by calculating various indexes to detect phishing websites.

III. PROPOSED WORK

Website phishing is the biggest security threat in the cyber security world and there have been many techniques and technologies implemented to defend it. There are many approaches through which we can do phishing detection such as domain-based analysis, lexical text feature based analysis, etc. Here we are applying a combined approach where we consider all the possible approaches such as domain-related features, address bar related features, and Java script and HTML related features. And finally instead of training with ML models only, we train it with ML and DL models to detect phishing. The problem here is the lack of a efficient system for detecting phishing websites. Website phishing is very common and phishing attackers are adopting advanced technologies like AI/ML to escape from defense mechanisms. So it's a must to develop a phishing detection system that uses these technologies for detection and that is lightweight. The system or the model should also be accurate enough. The developed system should input the URL as a plane text and should

output whether the URL is phishing or legitimate in no time.

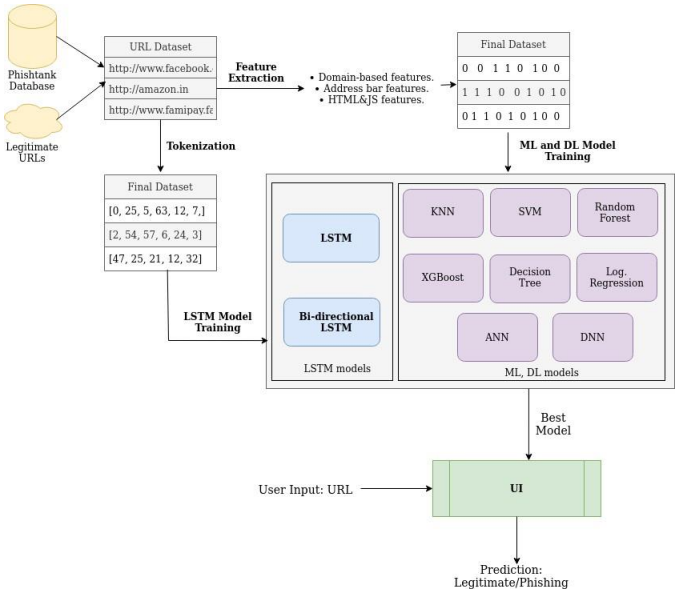

Figure 1: Architecture of the proposed phishing detection system

Figure 1 depicts the detailed design of proposed system. In general the development of the system goes through various steps, First of all, collection of URL data for phishing and legitimate websites. Feature extraction from the URL data. Preparation of final dataset. Load the input data and detection is performed using machine learning models like KNN, Decision Tree, Random Forest, SVM, Logistic Regression and XGBoost. Then Training of deep learning models ANN, DNN, LSTM and Bi-directional LSTM. Testing of the models using the evaluation metrics and tabulation of performance. Deploy the model in a tkinter app and test.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Data Collection

URLs of legitimate and phishing websites are collected for ML model training. The phish URLs are collected from the open-source reservoir of www.phishtank.com. They have listed in their website, urls of verified phish websites. The legitimate URLs are collected from Cyber Security research of University of New Brunswick.

B. Feature Extraction

Three types of features are extracted: Address bar-based features, Domain-based features, HTML and Javascript-based features. After feature extraction the final dataset is split into train and test data as 80:20. The train data will be used to train the models well and the test data will be used to test the model performance on unknown data.

C. Result on Machine Learning Classifiers

ML Model	Accuracy	F1-Score	Recall	Precision
RF	96.6%	97%	99.5%	98.8%
XG Boost	96.5%	96.9%	99.2%	98.6%
Decision Tree	95.6%	96.1%	99.2%	99.3%
SVM	95.6%	96.1%	97.9%	96.6%
KNN	95.5%	96%	99%	99%
Logistic Regression	92.4%	93.2%	94.4%	92.7%

Table I: Result of ML classifiers

Implemented six different classifiers: K-Nearest Neighbors(KNN), Random Forest, Decision tree, XG Boost, Support Vector Machine, Logistic Regression and then compare their performance using Accuracy, Precision, Recall and F1-Score. Performance comparisons are shown in the Fig. 7. Analyzed that the Random Forest classifier performs better with high accuracy of 96.6%.

D. Result on Deep Learning Models

DL Model	Accuracy	F1-Score	Recall	Precision
ANN	96.6%	96.8%	99.2%	98.9%
DNN	96.1%	96.4%	99.4%	99%
LSTM	99.9%	-	-	-
Bi-LSTM	95.6%	-	-	-

Table II: Results of Deep Learning models

Four different deep learning methods are implemented: ANN(Artificial Neural Network), DNN(Deep Neural Network), LSTM(Long Short-Term Memory) and Bi-LSTM(Bidirectional LSTM). On performance evaluation LSTM model outperforms all other DL models with an accuracy of 99.9%.

E. Developing a Desktop App

A final python app developed with minimal GUI features to deploy the models and to detect as shown in the figure. The desktop application is developed using python Tkinter. A desktop application is a user interface for the program, giving an easy way to interact with the code. Python Tkinter is used as Graphical User Interface (GUI). Desktop application contains a window, labels and buttons to execute the actions of the software. GUI will contain a page to give the input url and the output will be predicted on the same page.

Figure 2: Example for detection of legitimate url

Figure 3: Example for detection of phishing url

V. Conclusion

Website phishing is the biggest threat in the cyber security space! Yet, there is not much defense system present that has the characteristics such as high accuracy and light weight. We collected data for both the problems and carried out an extensive preprocessing on it to make it a machine learning trainable dataset. Machine Learning models such as KNN, SVM, Decision Tree, Random forest, Logistics Regression, and XGBoost are trained along with ANN, DNN, LSTM and Bi-drectional LSTM. After testing and evaluating with performance evaluation metrics, Random Forest and LSTM models found to be performing well. Developed a python app in the tkinter GUI framework and deployed the model and tested the app for phishing detection and ransomware detection in real-time.

References

[1] E. Zhu, Z. Chen, J. Cui and H. Zhong, "MOE/RF: A Novel Phishing Detection Model based on Revised Multi-Objective Evolution Optimization Algorithm and Random Forest," in IEEE Transactions on Network and Service Management, doi: 10.1109/TNSM.2022.3162885.

[2] M. Abutaha, M. Ababneh, K. Mahmoud and S. A. -H. Baddar, "URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis," 2021 12th International Conference on Information and Communication Systems (ICICS), 2021, pp. 147-152, doi: 10.1109/ICICS52457.2021.9464539.

[3] S. Alrefaai, G. Özdemir and A. Mohamed, "Detecting Phishing Websites Using Machine Learning," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2022, pp. 1-6, doi: 10.1109/HORA55278.2022.9799917.

[4] B. Geyik, K. Erensoy and E. Kocyigit, "Detection of Phishing Websites from URLs by using Classification Techniques on WEKA," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 120-125, doi: 10.1109/ICICT50816.2021.9358642.

[5] S. Singh, M. P. Singh and R. Pandey, "Phishing Detection from URLs Using Deep Learning Approach," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9277459.

[6] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104161.

[7] S. Parekh, D. Parikh, S. Kotak and S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 949-952, doi: 10.1109/ICICCT.2018.8473085.

[8] . H. Yuan, X. Chen, Y. Li, Z. Yang and W. Liu, "Detecting Phishing Websites and Targets Based on URLs and Webpage Links," 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3669-3674, doi: 10.1109/ICPR.2018.8546262.

[9] M. N. Feroz and S. Mengel, "Phishing URL Detection Using URL Ranking," 2015 IEEE International Congress on Big Data, 2015, pp. 635-638, doi: 10.1109/BigDataCongress.2015.97.

[10] C. L. Tan, K. L. Chiew and S. N. Sze, "Phishing website detection using URL-assisted brand name weighting system," 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2014, pp. 054-059, doi: 10.1109/ISPACS.2014.7024424.

[11] Y. Su, "Research on Website Phishing Detection Based on LSTM RNN," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2020, pp. 284-288, doi: 10.1109/ITNEC48623.2020.9084799.

[12] Masum, Mohammad Hossain Faruk, Md Jobair Shahriar, Hossain Qian, Kai Lo, Dan Adnan, Muhaiminul. (2022). Ransomware Classification and Detection With Machine Learning Algorithms. 10.1109/CCWC54503.2022.9720869.

[13] Choudhary, A. S., Desai, R., Gupta, L., Gedam, M. (2021). Detection and prevention of Phishing Attacks. Asian Journal For Convergence In Technology (AJCT) ISSN -2350-1146, 7(1), 193-196.

[14] Narendra. M. Shekokar, Chaitali Shah, Mrunal Mahajan,Shruti Rachh “An ideal approach for detection and prevention of phishing attacks”, Procedia Computer Science, ISSN: 1877-0509, Vol: 49, Issue: 1, Page: 82-91, 2015.

[15] Mahajan, Rishikesh Siddavatam, Irfan. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications. 181. 45-47. 10.5120/ijca2018918026.

[16] Narendra. M. Shekokar, Chaitali Shah, Mrunal Mahajan,Shruti Rachh “An Ideal Approach For Detection And Prevention of Phishing Attacks”, Procedia Computer Science, ISSN: 1877-0509, Vol: 49, Issue: 1, Page: 82-91,2015.

[17] Khonji, Mahmoud Iraqi, Youssef Jones, Andy. (2013). Phishing Detection: A Literature Survey. IEEE Communications Surveys amp Tutorials. PP. 1-31. 10.1109/SURV.2013.032213.00009.

[18] J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," 2013 International Conference on Control Communication and Computing (ICCC), Thiruvananthapuram, India, 2013, pp. 304-309, doi: 10.1109/ICCC.2013.6731669.