Artificial Neural Network Deep Learning Approach for Phishing Websites Prediction

Gaurav Tiwari
Research Scholar
Department of CSE
OIST, Bhopal, India

Atul Barve
Associate Professor
Department of CSE
OIST, Bhopal, India

ABSTRACT
Phishing is known as one of the oldest forms of Cyber attacks. A phishing website is a domain similar in name and appearance to an official website. They're made in order to fool someone into believing it is legitimate. Today, phishing schemes have gotten more varied, and are potentially more dangerous than before. Artificial intelligence based machine and deep learning techniques is capable to predict the phishing websites. The detection of phishing websites using machine learning can be conducted using machine and deep learning classification technique. This paper presents an artificial neural network based deep learning technique for detection of phishing websites.

Keywords
Phishing Websites, ANN, AI Model, Deep Learning, Accuracy.

1. INTRODUCTION
Now day's digital operations became more important, and people started to depend on new initiatives such as the cloud and mobile infrastructure. Consequently, the number of cyberattacks such as phishing has increased. Machine learning can identify phishing sites by separating them from trustworthy sources [1].

Phishing is an email scam when the sender poses as a legitimate business or individual in an effort to get sensitive information such as login credentials or credit card numbers. Scammers often utilise emails purporting to come from official-looking organisations like social networks, banks, auction sites, and IT administrators to trick the unwary. It's a malicious sort of social manipulation.

Bots that "scrape" the web are just computer programmes that do this task mechanically. The prices of items are copied and pasted unlawfully from other online retailers onto their website. Approximately half of all website traffic, according to various online traffic assessments, is generated by automated programmes. Even though there has been a lot of progress made in the realm of E-commerce, there are still numerous security problems that arise while operating an online store. Competitor pricing scraping is one of the most destructive assaults on e-commerce businesses [4].

Web users' browsing habits may be collected and correlated thanks to third-party monitoring. As ad-blocking software and other anti-tracking measures gained popularity, tracking service providers developed a new method dubbed "CNAME cloaking" to circumvent these safeguards. As a result, the browser thinks the request came from the visited site, even if the subdomain in question employs a CNAME to resolve to a tracking-related third-party domain. Thus, the privacy safeguards against third-party targeting are bypassed by using this method. The research presented in this article aims to better understand CNAME cloaking-based tracking and to develop methods for detecting and safeguarding against it. First, utilising a CNAME blocklist, we define CNAME cloaking-based tracking by scanning the most popular pages from the Alexa Top 300,000 sites and examining the utilisation of CNAME cloaking.

These days, cybercrime comes in many forms. One kind of cyberattack, known as "phishing," involves hackers posing as representatives of a trustworthy company or institution in an effort to trick their targets into divulging private data (such as account or social security numbers) by email, text message, ads, or other methods. The frequency with which phishing attacks are occurring has been rising at an exponential rate. Most of those whose data was compromised in this assault were unwitting victims. Phishing assaults are a common method of hacking success, in which victims are duped into providing sensitive information by clicking on seemingly authentic links.

Huge institutional and personal security concerns have arisen as a result of the widespread use of technologies like ubiquitous network access, big data, the Internet of Things (IoT), global digitization, and the usage of social networking sites and applications. When it comes to protecting people and businesses online, the traditional security system frequently falls short. AI has the intelligence and adaptability to deal with the ever-changing cyber security landscape. The identification of spam, spyware, and botnets [3, 4] are only some of the areas where AI has proven useful.

There has been a recent uptick in interest in the use of artificial intelligence (AI) to the field of web development. Artificial intelligence (AI) is expanding and maturing, and it's becoming more crucial in the realm

of web app development. The technologies involved continue to play an increasingly important role in the creation of new and advanced online applications. With the involvement of the internet into our daily lives, particularly businesses are enjoying the aspects of AI. Precisely, companies use AI in proper marketing of their products and enhancing their brand visibility by building their websites and web applications. AI or Machine Learning (ML) models are able to help web app developers to solve problems related to security, user experience, content analysis, quality assurance and much more. This presents the need for a framework or tool that can allow third party developers to seamlessly build an AI based app [7].

This paper is organised into the IV section. I section provides the overview & introduction of the anomaly detection. The II section provides the proposed methodology, III section provides the simulation and results and IV section provides the conclusion of this research paper.

2. PROPOSED METHODOLOGY

The proposed methodology is explained using following flow chart-

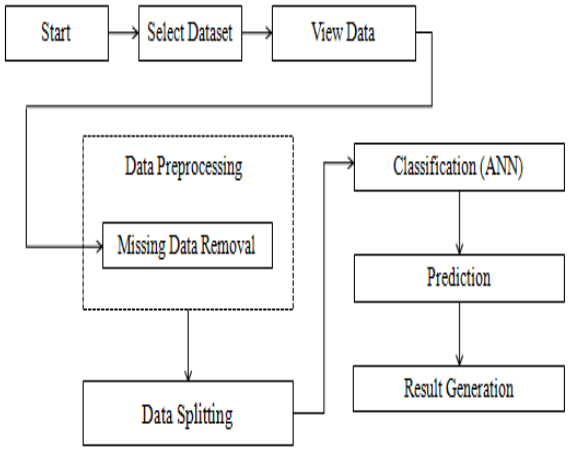

Figure 1: Flow Chart

Steps-

- Firstly, finalize the dataset [13] based on the phishing website, taken from publicly available large dataset repository.
- Now preprocessing of the data, here handing the missing dataset. Remove the null value or replace from common 1 or 0 value.
- Now apply the classification method based on the artificial neural network approach.
- Now check and calculate the performance parameters in terms of the precision, recall, F_measure, accuracy and error rate.

The methodology of the proposed research is based on the following sub modules-

- Data Selection and Loading
- Data Preprocessing
- Splitting Dataset into Train and Test Data
- Feature Extraction
- Classification
- Prediction
- Result Generation

Data Selection and Loading

- The data selections are the process of selecting the dataset and load this dataset into the python environment.

Data Pre-processing

- Data pre-processing is the process of removing the unwanted data from the dataset.
- Missing data removal
- Encoding Categorical data
- Missing data removal: In this process, the null values such as missing values are removed using imputer library.

Splitting Dataset into Train and Test Data

- Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
- One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.

Feature Extraction

Feature extraction is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Classification

ANN- artificial neural network is best represented as a directed graph with weights assigned to the artificial neurons that make up the network. We may think of the connection between a neuron's output and input as directed edges with weights. The Artificial Neural Network takes in data from the outside world as a vectorized pattern or picture. For each n number of inputs, a mathematical notation x(n) is used to ascribe a value.

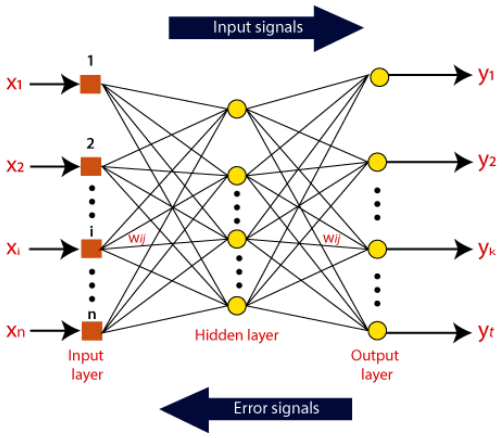

Figure 2: ANN layer

After that, multiply each input by its associated weight (these weights are the details utilised by the artificial neural networks to solve a specific problem). These weights often stand in for the robustness of the neural network's internal connections. Inside the computer, all of the weighted inputs are tallied.

If the total weighted value is 0, then bias is used to make the output greater than zero. The input for bias and the value of weight are both 1. In this case, the sum of the weighted inputs might be any positive number. Here, a maximum value is used as a reference to ensure that the response stays within the acceptable range, and the activation function is applied to the sum of the weighted inputs.

The activation function is the collection of transfer functions that produces the desired result. Each activation function is unique, but most fall into one of two categories: linear or non-linear. Binary, linear, and Tan hyperbolic sigmoidal activation functions are three examples of popular families of activation functions.

Prediction

- It's a process of predicting android malware from the dataset.
- This research is effectively predicted the data from dataset by enhancing the performance of the overall prediction results.

Evaluation

The confusion metrics used to evaluate a classification model are accuracy, precision, and recall.

- Precision = True Positive/(True Positive + False Positive)
- Recall = True Positive/(True Positive + False Negative)
- F1-Score = 2x (Precision x Recall)/(Precision + Recall)
- Accuracy = [TP +TN] / [TP+TN+FP+FN]
- Classification Error = 100- Accuracy

Result Generation

The final result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like accuracy, error rate etc.

3. SIMULATION AND RESULTS

The simulation is performed using the Python Spyder IDE 3.7 software.

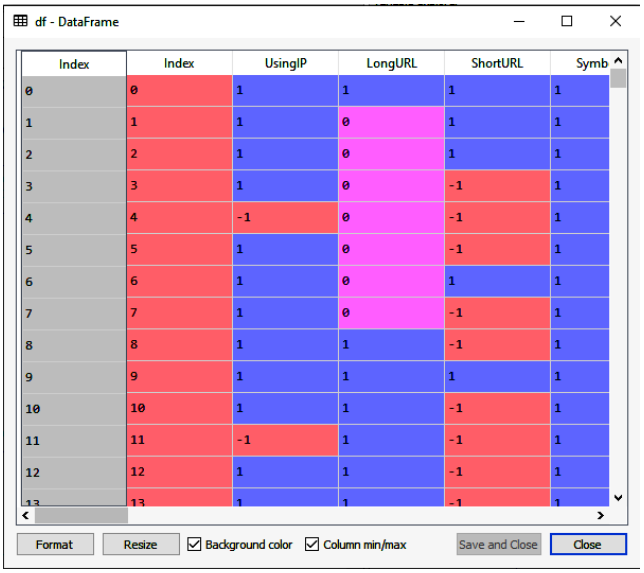

Figure 3: Dataset

Figure 3 is showing the dataset in the python environment. The dataset have various numbers of rows and column. The features name is mention in each column.

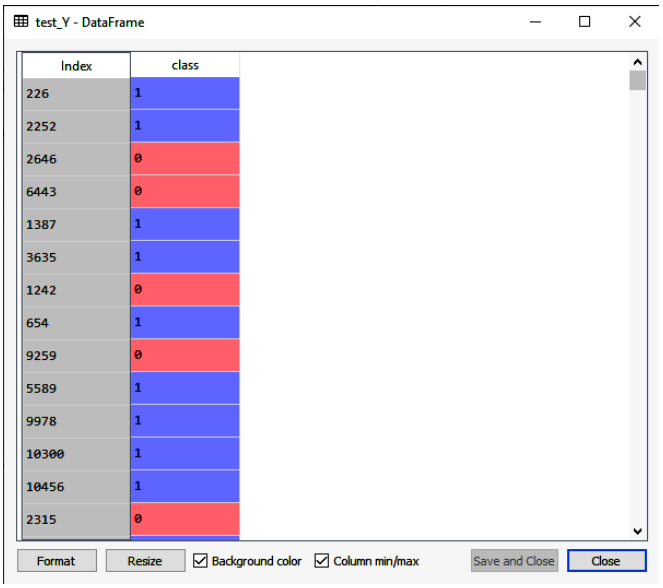

Figure 4: Y test

Figure 4 is showing the y test of the given dataset. The given dataset is divided into the 20-30% part into the train dataset.

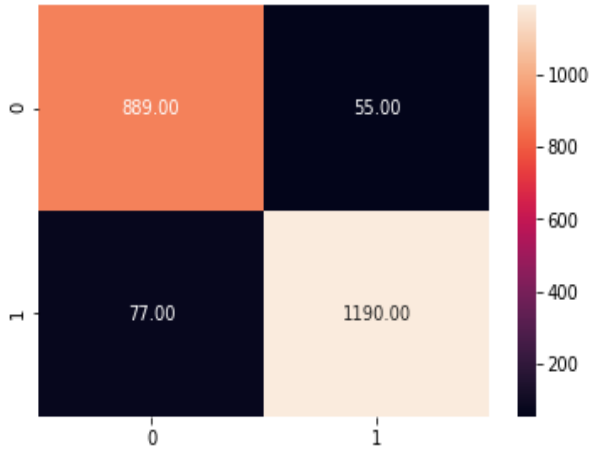

Figure 5: Confusion matrix heat map

Figure 5 is showing the heat map confusion matrix of the ANN classification technique. It is an N x N matrix used for evaluating the performance of a classification model.

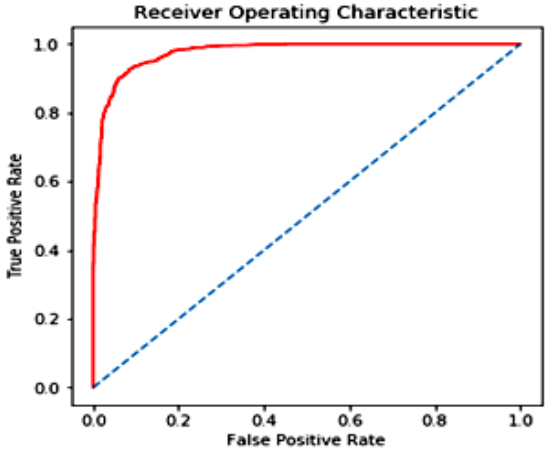

Figure 6: ROC

Figure 6 is presenting the receiver operating characteristic curve. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance.

Table 1: Result Comparison

Sr. No.	Techniques	Accuracy (%)
1	Decision Tree	91.51
2	K-Nearest Neighbor	97.69
3	Random Forest	94.44
4	ANN (Proposed)	98.3

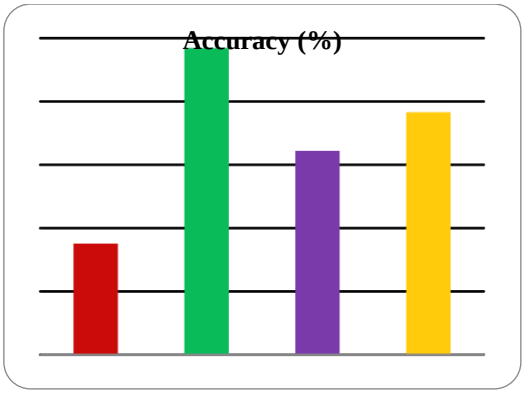

Figure 7: Accuracy Result graph

Figure 7 is presenting the graphical representation of the accuracy. The proposed work achieved better accuracy then existing work.

4. CONCLUSION

The websites exactly seems to be semantically as well as visually to the original websites. The main idea of the phisher or hackers is to gain and purloin the critical information such as credential account, username, password and other private information related to any organization and company. According to phishing or web spoofing techniques is one examples of social engineering attack. This paper presents the support vector machine learning technique for detection of phishing websites. The simulated results shows that the proposed support vector machine learning classification technique achieve better accuracy rather than existing techniques. The ANN achieved 98.3% accuracy while existing KNN achieve 97.69% accuracy.

5. REFERENCES

[1] F. Yahya et al., "Detection of Phishing Websites using Machine Learning Approaches," 2021 International Conference on Data Science and Its Applications (ICoDSA), 2021, pp. 40-47, doi: 10.1109/ICoDSA53588.2021.9617482.

[2] K. S. Swarnalatha, K. C. Ramchandra, K. Ansari, L. Ojha and S. S. Sharma, "Real-Time Threat Intelligence-Block Phising Attacks," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1-6, doi: 10.1109/CSITSS54238.2021.9683237.

[3] S. M. Istiaque, M. T. Tahmid, A. I. Khan, Z. A. Hassan and S. Waheed, "Artificial Intelligence Based Cybersecurity: Two-Step Suitability Test," 2021 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), 2021, pp. 1-6, doi: 10.1109/SOLI54607.2021.9672437.

[4] R. Yaqoob, Sanaa, M. Haris, Samadyar and M. A. Shah, "The Price Scraping Bot Threat on E-commerce Store Using Custom XPATH Technique," 2021 26th International Conference on Automation and Computing (ICAC), 2021, pp. 1-6, doi: 10.23919/ICAC50006.2021.9594223.

[5] M. Min, J. J. Lee, H. Park and K. Lee, "Honeypot System for Automatic Reporting of Illegal Online Gambling Sites Utilizing SMS Spam," 2021 World Automation Congress

(WAC), 2021, pp. 180-185, doi: 10.23919/WAC50355.2021.9559478.

[6] H. Dao, J. Mazel and K. Fukuda, "CNAME Cloaking-Based Tracking on the Web: Characterization, Detection, and Protection," in IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3873-3888, Sept. 2021, doi: 10.1109/TNSM.2021.3072874.

[7] R. Nanjundappa et al., "AWAF: AI Enabled Web Contents Authoring Framework," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-5, doi: 10.1109/INDICON49873.2020.9342385.

[8] K. E. Aydın and S. Baday, "Machine Learning for Web Content Classification," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, pp. 1-7, doi: 10.1109/ASYU50717.2020.9259833.

[9] U. Iqbal, P. Snyder, S. Zhu, B. Livshits, Z. Qian and Z. Shafiq, "AdGraph: A Graph-Based Approach to Ad and Tracker Blocking," 2020 IEEE Symposium on Security and Privacy (SP), 2020, pp. 763-776, doi: 10.1109/SP40000.2020.00005.

[10] N. Megha, K. R. Remesh Babu and E. Sherly, "An Intelligent System for Phishing Attack Detection and Prevention," 2019 International Conference on Communication and Electronics Systems (ICCES), 2019, pp. 1577-1582, doi: 10.1109/ICCES45898.2019.9002204.

[11] S. S. Hashmi, M. Ikram and M. A. Kaafar, "A Longitudinal Analysis of Online Ad-Blocking Blacklists," 2019 IEEE 44th LCN Symposium on Emerging Topics in Networking (LCN Symposium), 2019, pp. 158-165, doi: 10.1109/LCNSymposium47956.2019.9000671.

[12] T. Vo and C. Jaiswal, "ADREMOVER: THE IMPROVED MACHINE LEARNING APPROACH FOR BLOCKING ADS," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2019, pp. 1-4.

[13] https://www.kaggle.com/datasets/isatish/phishing-dataset-uci-ml-csv?select=uci-ml-phishing-dataset.csv.