IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

# Artificial Neural Network based Prediction of Phishing Website

**[1]Prof. Priyanka Soni, [2]Prof. Poonam Choubey, [3]Prof. Sneha Sule**
[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor
Department of Computer Science and Information Technology
Sagar Institute of Research & Technology, Bhopal, India

***Abstract :*** Phishing is one of the first types of cyber assaults. Phishing websites are domains that have names and layouts that are very similar to those of legitimate websites. They are created with the intention of convincing someone that they are genuine when they are not. These days, there is a greater variety of phishing scams, each of which has the potential to be more harmful than its predecessor. It is possible to identify potential phishing websites using machine learning and deep learning approaches, both of which are supported by artificial intelligence. The identification of phishing websites that use machine learning may be accomplished by utilizing a classification approach that combines machine learning with deep learning. This article proposes a method for the identification of phishing websites that is based on the use of artificial neural networks deep learning technique.

***IndexTerms*** - Phishing Websites, ANN, AI Model, Deep Learning, Accuracy**.**

## I. 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. Phishing websites can be detected using machine learning by classifying the websites into legitimate or illegitimate websites [1].

Phishing is the process of attempting to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a trustworthy entity using bulk email which tries to evade spam filters. Emails claiming to be from popular social web sites, banks, auction sites, or IT administrators are commonly used to lure the unsuspecting public. It's a form of criminally fraudulent social engineering.

Scraping bots or scrapers are computer programs that automatically fetch data from the Internet. The price of products is illegally copied from different e-commerce stores and pasted into their sites for their benefit. Various web traffic surveys show that automated programs account for approximately half of all website traffic. However, an E-commerce website concerns many security issues that remain unresolved despite a significant increase in E-commerce development. One of the most damaging attacks on E-commerce sites is the price scraping when it comes to competitors [4].

Third-party tracking on the Web has been used for collecting and correlating user's browsing behavior. Due to the increasing use of ad-blocking and third-party tracking protections, tracking providers introduced a new technique called CNAME cloaking. It misleads Web browsers into believing that a request for a subdomain of the visited website originates from this particular website, while this subdomain uses a CNAME to resolve to a tracking-related third-party domain. This technique thus circumvents the third-party targeting privacy protections. The goals of this paper are to characterize, detect, and protect the end-user against CNAME cloaking based tracking. Firstly, we characterize CNAME cloaking-based tracking by crawling top pages of the Alexa Top 300,000 sites and analyzing the usage of CNAME cloaking with CNAME blocklist, including websites and tracking providers using this technique to track users' activities [6].

Now-a-days there are different types of cybercrime, Phishing is one of cyber-attacks where attackers impersonate as a member of legitimate institutions or organizations through an email, text message, advertisements or through any means to steal sensitive information which results to loss of personal and sensitive information such as account no, social security no, credit card no etc. Phishing attack has been increasing exponentially. In this attack mostly innocent users are comes to losses their sensitive, unique, personal, valuable and secure data and information's. Many hackers are accomplished through phishing attacks where client are trapped into interacting with web-pages which looks like to be legitimate websites.

Enormous network connectivity, big data, the Internet of Things (IoT), digitalization of the world, and the use of social websites and apps have brought enormous institutional and individual security challenges. The conventional security system often fails to provide cyber security to institutions and individuals. Artificial Intelligence (AI) is highly adaptive and smart to handle the volatile cyber security environment. AI plays a prudent role in access control, user authentication and behavior analysis, spam, malware, and botnet detection [3].

Artificial intelligence (AI) in web development is a new sector that a lot of people are into recently. AI continues to evolve and grow, and plays an increasingly important role in the web app development space. When it comes to developing innovative and more sophisticated web applications, the involved technologies continue to play a bigger role. 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].

## II. METHODOLOGY

The proposed research work can be understand by using the following flow chart-

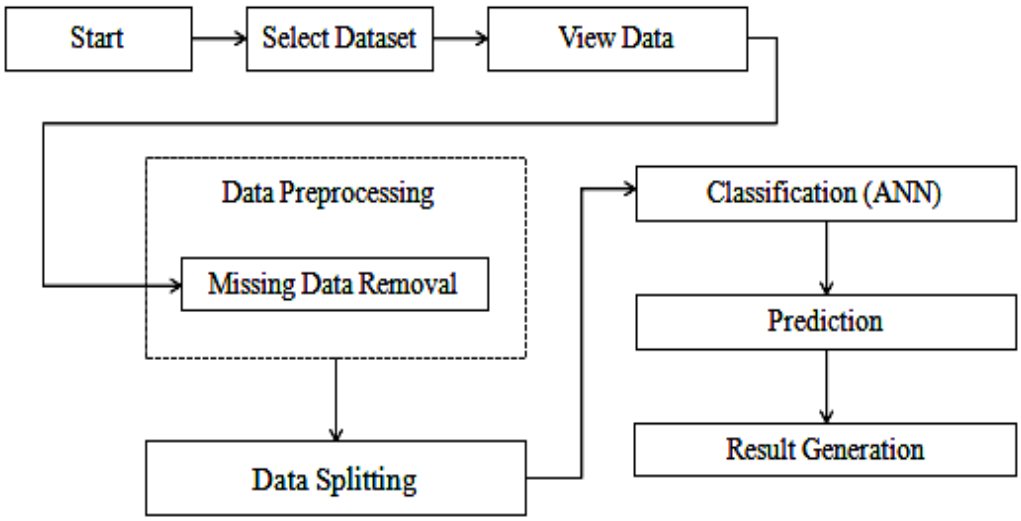


Figure 1: Flow chart

**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.

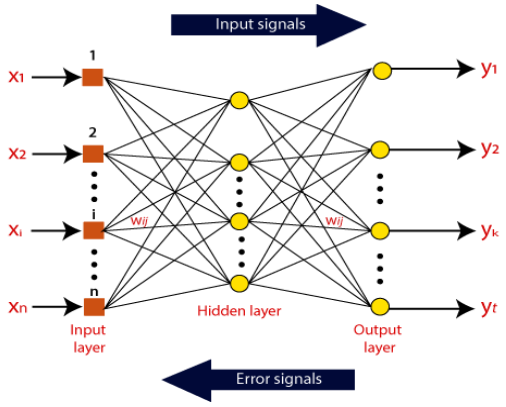

Figure 2: ANN layer

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

## III. SIMULATION 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.

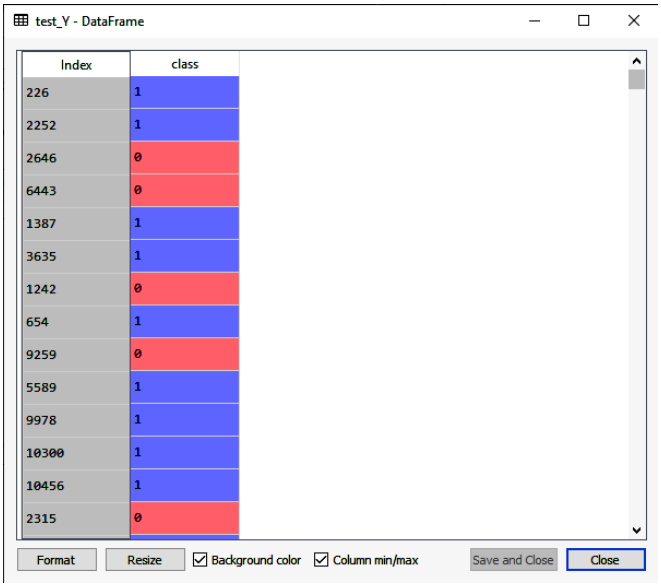


Figure 4: Y test

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.

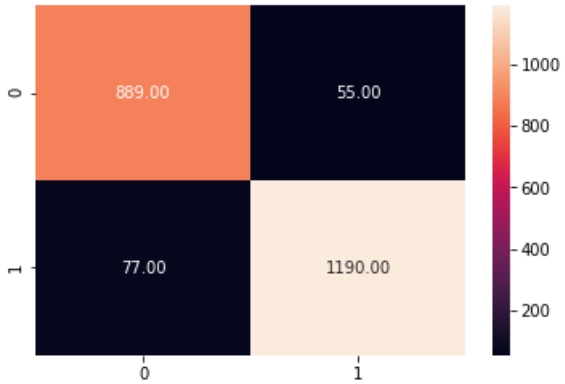


Figure 5: Confusion matrix heat map

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.

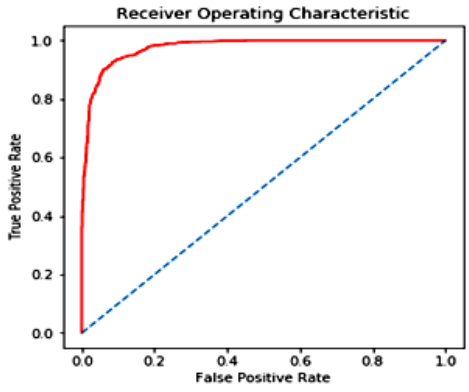


Figure 6: ROC

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) | 98.2 |

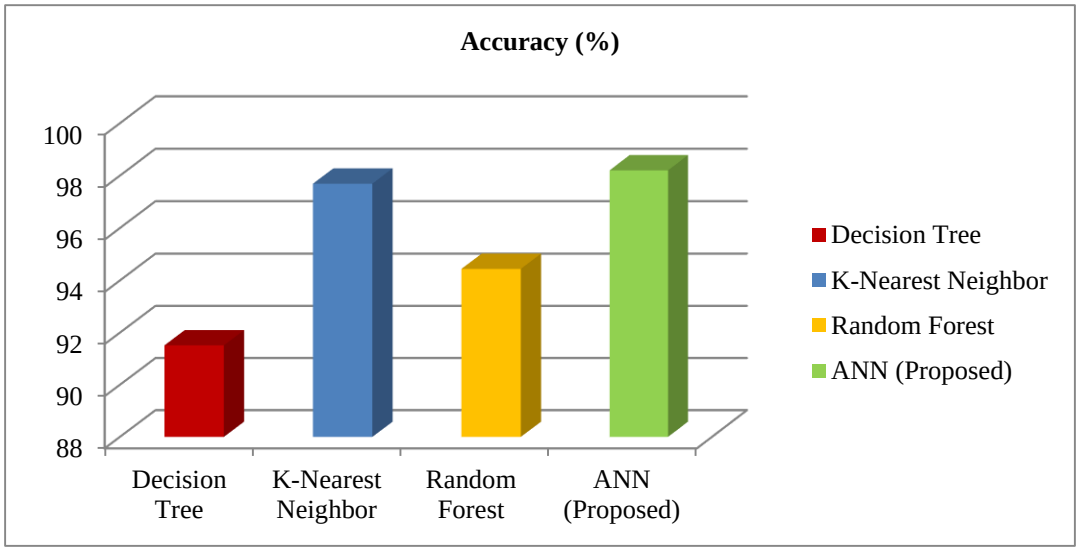


Figure 7: Accuracy Result graph

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

## IV. 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 98.2% accuracy while existing KNN achieve 97.69% accuracy.

## 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.