International Journal of Research in Engineering, Science and Management
Volume 5, Issue 6, June 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Prediction of Phishing Websites Using Machine Learning

S. Bargunan[1], R. Revathi[2], A. Priyadharshini[3*]
[1]Assistant Professor, Department of Computer Science and Engineering, Agni College of Technology, Chennai, India
[2,3]Under Graduate, Department of Computer Science and Engineering, Agni College of Technology, Chennai, India

***Abstract*: A large number of people buy things online and pay for them using different websites. Several websites ask users for personal information such as usernames, passwords, and credit card numbers, among other things, for harmful purposes. Phishing websites are exactly what they sound like. We suggested an intelligent, versatile, and effective solution based on machine learning techniques for detecting and predicting phishing websites. To extract the phishing data sets criteria and identify their authenticity, we used a classification algorithm and approaches. In the final phishing detection rate, URL and Domain Identity, as well as security and encryption criteria, can be used to detect the phishing website. Our system will utilize a machine learning algorithm to determine whether or not the website is phishing. Many E-commerce businesses can utilize this programme to make the entire transaction process secure. In comparison to other classic classification algorithms, the machine learning algorithm utilized in this system performs better. This technique also allows users to purchase things online without fear of being scammed.**

***Keywords*: Anti-phishing, machine learning, random forest, prediction.**

1. Introduction

Phishing is the criminally fraudulent process of attempting to get sensitive information such as usernames, passwords, and credit card numbers by impersonating a trustworthy institution in an electronic communication, according to the field of computer security. A phishing website is a widely distributed social engineering attack that seeks to defraud people of their personal information, such as credit card numbers, bank account numbers, social security numbers, and passwords, in order to fraudulently exploit these details against them. Phishing has a significant detrimental influence on a company's income, client relationships, marketing activities, and overall reputation. Communication that appears to be typical in order to entice the naïve population. Phishing is usually done by e-mail or instant messaging, and it often urges people to submit information on a phoney website that looks and feels remarkably identical to the real one. Phishing is a type of social engineering approach that takes Dominance of the poor usability of current web security systems to mislead users.

2. Literature Survey

A literature review is a piece of writing that seeks to summarise the most important aspects of current knowledge and/or methodological approaches to a specific issue. It is a secondary source that discusses published material in a specific subject area, as well as information in a specific subject area within a specific time period. Its ultimate goal is to keep the reader up to date on current literature on a topic, and it serves as the foundation for other goals, such as future research that may be required in the field. It comes before a research proposal and may simply be a list of sources. It usually follows a pattern and includes both synthesis and summarization. A summary is a re-organization and reshuffling of information, but a synthesis is a re-organization and reshuffling of information. It could offer a fresh perspective on old material, or blend new and old perspectives, or it could chart the field's intellectual evolution, including significant disagreements. The literature review may evaluate the sources and advise the reader on the most topical or relevant ones, depending on the situation.

A Bio-Inspired Self-learning Coevolutionary Dynamic Multiobjective Optimization Algorithm for Internet of Things Services, Zhen Yang, Yaochu Jin, and Kuangrong Hao, 2018.

The Internet of Things' (IoT) ultimate purpose is to provide pervasive services. Many obstacles must be overcome in order to reach this aim. This work offers a bio-inspired self-learning coevolutionary algorithm (BSCA) for dynamic multiobjective optimization of IoT services to reduce energy consumption and service time, based on cooperative mechanisms across different systems in humans. The BSCA is made up of three tiers. Multiple subpopulations evolve cooperatively to produce distinct Pareto fronts in the first layer. The second layer tries to expand the diversity of solutions by building on the first layer's solutions. The third layer improves the accuracy of the answers identified in the second layer by employing an adaptive gradient refinement search technique and a dynamic optimization method to cope with changing concurrent multiple service requests. Experiments on agricultural IoT services in the presence of dynamic requests are carried out using two service-providing methodologies, namely single service and collaborative service. The simulation findings show that BSCA outperforms four existing algorithms on IoT services,

*Corresponding author: priyadharshiniamsaraj@gmail.com

especially when dealing with high-dimensional challenges. In this research, a three-layer progressive bio-inspired self-learning coevolutionary algorithm (BSCA) is presented for dynamic multi-objective optimization of IoT services to minimize service costs and time. BSCA is based on human neurological, endocrine, and immunological systems' mechanisms for tracking moving Pareto optimal solutions in the face of changing requirements. The results of the simulations show that the suggested method is competitive in the dynamic optimization of agricultural IoT services. In practice, an IoT service system may choose one of the extreme solutions or other Pareto optimal options on the front end based on the decision-service maker's strategy.

A Prediction Model of DoS Attack's Distribution Discrete Probability, Wentao Zhao, Jianping Yin, Jun Long, 2008.

Prediction analysis is a strategy or technology for discovering or stimulating new, unknown, or challenging intermediate processes based on prior and current circumstances, and then guessing on the results [5]. The goal of a network offence and defense task in an early warning system is to accurately predict DoS attacks. The detection of DoS assaults using abnormity is effective. Several research [2][6][10] focused on DoS assaults in various ways. These solutions, on the other hand, required a priori information and made distinguishing between regular burst flow and DoS attack flux unfeasible. Furthermore, they required a huge quantity of historical information and were unable to predict such attacks well. We provide a prediction model of DOS attack distribution discrete probability based on data from flow inspection and intrusion detection using a genetic algorithm clustering methodology and a Bayesian method. The frequency of DoS attacks is considered a random variable due to numerous interfering effects.

New Attack Scenario Prediction Methodology, Seraj Fayyad, Cristoph Meinel, 2013.

The intrusion detection system generates a lot of information about harmful behaviours that happen on the network. The data generated by IDS is saved in the IDS database. This information represents the history of network attacks. The primary objective of an IDS system is to improve network protection technology. Other strategies, such as Attack graph, are also utilised to improve network defence. The network attack graph is used for a variety of purposes, including predicting the attacker's next attack step. We present a real-time prediction methodology in this research for anticipating the majority of possible assault stages and attack scenarios. The proposed methodology takes advantage of network attack history as well as attack graph source data. It does not require a lot of calculation, such as examining the library of assault strategies. It can forecast parallel assault situations in real time.

3. Existing System

Lookup systems, fraud cue-based approaches, and deep representation-based methods are the three types of CTI for phishing website detection currently available. The lookup system detects a phishing website by "looking up" the URL against a blacklist of phishing URLs, and an alarm is triggered if the URL is on the list. To detect phishing websites, blacklists consist of classifiers (e.g., SVM, decision tree) and novel machine learning methods (Approaches based on statistical learning theory, genre tree kernel methods, and the recursive trust labelling algorithm, for example). Similarly, fraud cues based on website traffic must monitor website traffic over time, making it difficult to achieve the real-time detection criteria.

Disadvantages of existing system:

1. If we wish to update any features and train the model, the transfer learning takes longer.
2. They make no mention of the model's accuracy.
3. There are no performance measures such as recall F1 score or machine learning algorithm comparison.
4. The performance is poor, and it is becoming more difficult for other networks

4. Proposed System

The proposed model is for developing a machine learning model for detecting anomalies. Anomaly detection is a critical approach for detecting fraud, suspicious activity, network intrusion, and other unusual events that may be relevant but are difficult to notice. The machine learning model is created using proper data science techniques such as variable identification, the process of identifying the dependent and independent variables. The data is then visualised to reveal the data's insights. The model is built on the prior dataset, and new techniques are employed for better comparisons when the algorithm learns data and is trained. The metrics of performance are calculated and compared.

Advantages of the proposed system:

1. Machine learning can be used to automate the anomaly detection process.
2. In order to get a better model, the accuracy level of Machine Learning Algorithm Model is calculated and performance metrics are compared.

A. Modules

Data Pre-processing: Machine learning validation approaches are used to calculate the error rate of the Machine Learning (ML) model, which is as close to the genuine error rate of the dataset as possible. Validation approaches may not be required if the data volume is large enough to be representative of the population. However, in real-world circumstances, it is necessary to work with data samples that are not always representative of the population of a dataset. Identify the missing value by duplicating the value and the data type description, whether the variable is a float or an integer variable.

Exploration data analysis of visualization: In applied statistics and machine learning, data visualization is a crucial ability. Statistics is concerned with quantitative data descriptions and estimations. Data visualization is a valuable set of tools for acquiring a qualitative understanding of data. This might be useful for spotting patterns, faulty data, outliers, and other things when exploring and getting to know a dataset. Data visualizations can be used to express and demonstrate crucial relationships in plots and charts that are more visceral and

meaningful to stakeholders than measurements of association or significance with a little subject knowledge. Machine learning validation approaches are used to calculate the error rate of the Machine Learning (ML) model, which is as close to the genuine error rate of the dataset as possible. Validation approaches may not be required if the data volume is large enough to be representative of the population. However, in real-world circumstances, it is necessary to work with data samples that are not always representative of the population of a dataset. Identify the missing value by duplicating the value and the data type description, whether the variable is a float or an integer variable.

Algorithm:

Random Forest Classifier:

We are using four classification to find a best accuracy for our project and that are Logistic regression, Random forest, Decision Tree, Naive Bayes. We got more accuracy in random forest so that's why We're using the random forest method in our project to improve accuracy. Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other tasks that uses a large number of decision trees to train and then output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

```
print("Accuracy result of Random Forest Classifier is:",accuracy.mean() * 100)
LR=accuracy.mean() * 100

Classification report of Random Forest Classifier Results:

              precision    recall  f1-score   support

          -1       0.90      0.91      0.90       211
           0       0.84      0.84      0.84        31
           1       0.87      0.85      0.86       164

    accuracy                           0.88       406
   macro avg       0.87      0.87      0.87       406
weighted avg       0.88      0.88      0.88       406


Confusion Matrix result of Random Forest Classifier is:
 [[192   2  17]
 [  1  26   4]
 [ 21   3 140]]

Sensitivity :  0.9896907216494846

Specificity :  0.9629629629629629

Cross validation test results of accuracy:
[0.86346863 0.89667897 0.88929889 0.91111111 0.90740741]
```

Fig. 1. Classification report of random forest

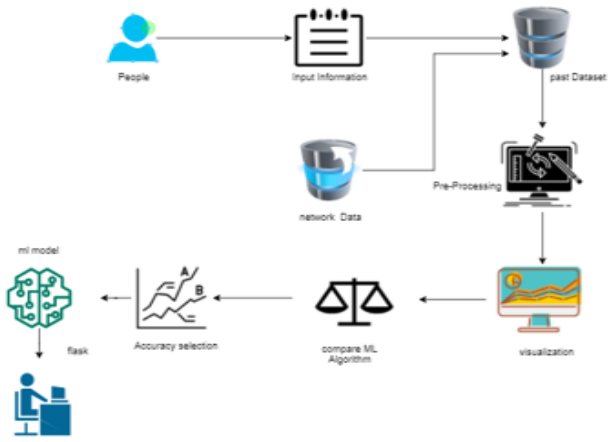

Fig. 2. System architecture

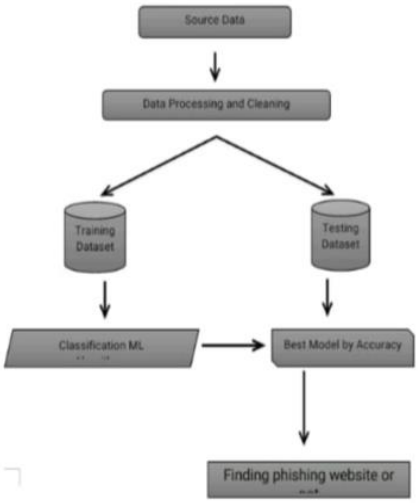

Fig. 3. Flow graph

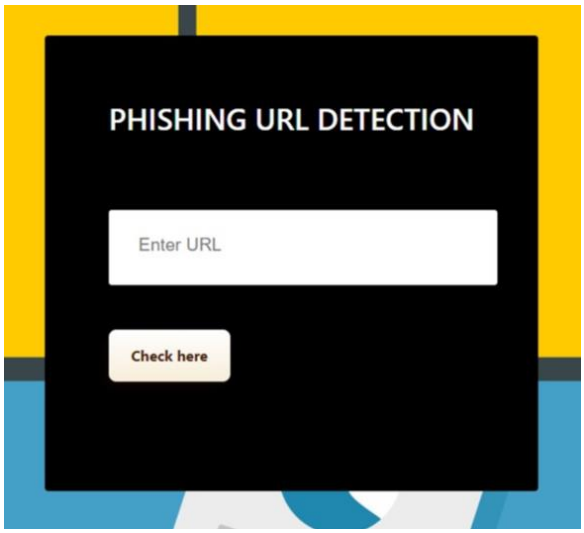

Fig. 4. Output screen

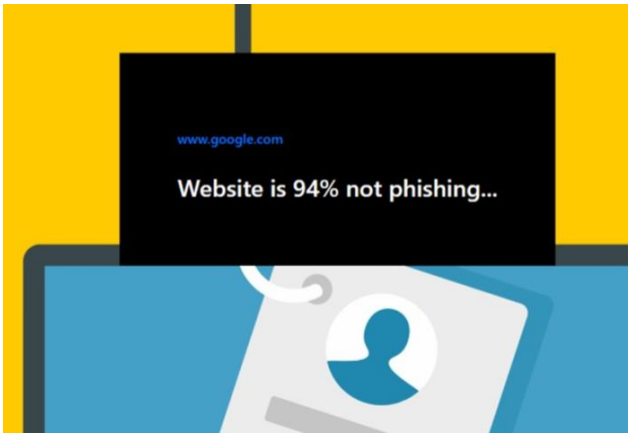

Fig. 5. Output result screen

5. Conclusion

The goal of this initiative was to improve information security. Data cleaning and processing, missing value analysis, exploratory analysis, and model creation and evaluation were all part of the analytical process. The best accuracy on a public test set will be discovered, as will the highest accuracy score.

This application can assist in determining whether a website is a phishing site or not.

References

[1] Mohammed Hazim Alkawaz; Stephanie Joanne Steven; Asif Iqbal Hajamydeen, "Detecting Phishing Website Using Machine Learning" 2020 16th IEEE International Colloquium on Signal Processing & its Applications (CSPA).

[2] Yuji Sakurai; Takuya Watanabe; Tetsuya Okuda; Mitsuaki Akiyama; Tatsuya Mori, "Discovering HTTPS Phishing Websites Using the TLS Certificates Footprints," 2020 IEEE European Symposium on Security and Privacy Workshops (Euro S&P W).

[3] Waleed Ali; Sharaf Malebary, "Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Websites Detection," IEEE Access 2020.

[4] Ayman El Aassal; Shahryar Baki; Avisha Das; Rakesh M. Verma ''An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs'' IEEE Access 2020

[5] Rizal Dwi Prayogo; Siti Amatullah Karimah, "Optimization of Phishing Websites Classification Based on Synthetic Minority Oversampling Technique and Feature Selection," 2020 International Workshop on Big Data and Information Security (IWBIS).

[6] A. S. Patrick, A. C. Long, and S. Flinn, "HCI and Security Systems," presented at CHI, Extended Abstracts (Workshops). Ft. Lauderdale, Florida, USA., 2003.

[7] K. Gilhooly, "Biometrics: Getting Back to Business," in Computer world, May 09, 2005.

[8] A. Jain, L. Hong, and S. Pankanti, "Biometric identification," Communications of the ACM, vol. 33, pp. 168-176, 2000.

[9] D. Weinshall and S. Kirkpatrick, "Passwords You'll Never Forget, but Can't Recall," in Proceedings of Conference on Human Factors

[10] in Computing Systems (CHI). Vienna, Austria: ACM, 2004, pp. 1399-1402.

[11] D. Davis, F. Monrose, and M. K. Reiter, "On user choice in graphical password schemes," in Proceedings of the 13thUsenix Security, Symposium. San Diego, CA, 2004.

[12] A. Jain, L. Hong, and S. Pankanti, "Biometric identification," Communications of the ACM, vol. 33, pp. 168-176, 2000.