

PHISHING SITES DETECTION USING RANDOM FOREST ALGORITHM

Amol C. Wani, Associated Professor SSBTCOET Jalgaon MS

Narayan S. Raut, SSBTCOET Jalgaon MS India

Nilesh V. Patil, SSBTCOET Jalgaon MS India

Chandrakant M. Patil, SSBTCOET Jalgaon MS India

Renuka S. Patil, SSBTCOET Jalgaon MS India

Sagar A. Marathe, SSBTCOET Jalgaon MS India

DEPARTMENT OF COMPUTER ENGINEERING

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY, BAMBHORI, JALGAON - 425 001
(MS)

Abstract- With raising in-depth amalgamation of the Internet and social life, the Internet is looking differently at how people are learning and working, meanwhile opening us to growing serious security attacks. The ways to recognize various network threats, specifically attacks not seen before, is a primary issue that needs to be looked into immediately. The aim of phishing site URLs is to collect the private information like user's identity, passwords and online money related exchanges. Phishers use the sites which are visibly and semantically like those of authentic websites. Since the majority of the clients go online to get to the administrations given by the government and money related organizations, there has been a vital increment in phishing threats and attacks since some years. As technology is growing, phishing methods have started to progress briskly and this should be avoided by making use of anti-phishing techniques to detect phishing. Machine learning is a authoritative tool that can be used to aim against phishing assaults. There are several methods or approaches to identify phishing websites. The machine learning approaches to detect phishing websites have been proposed earlier and have been implemented. The central aim of this project is to implement the system with high efficiency, accuracy and cost effectively. It was established that the Random forest classifier provides best accuracy for the selected dataset and gives an accuracy score of 96.82

Keywords-

Internet Protocol	(IP)
Optimal feature selection-neural network	(OFS-NN)
Feature validity value	(FVV)
Graphics user interface	(GUI).
Architecture description languages	(ADLs)
Fuzzy Rough Set	(FRS)

I. INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of 2 billion us dollar per year because their clients become victim to phishing [1]. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as 5 billion dollar [2]. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as "blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation

and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive retain detection is very high [3]. To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing website

II. Literature Survey

To study and analyze more about Phishing Sites Detection using Machine Learning, the following literature survey has been done.

A literature survey is an insightful article that presents the existing information including considerable discoveries just as theoretical and methodological commitments to a specific topic. A very effective detection of phishing website model which is focused on optimal feature selection technique and also based on neural network (OFS-NN) is proposed [23]. In this proposed model, an index called feature validity value (FVV) has been generated to check the effects of all those features on the detection of such websites. Now, based on this newly generated index, an algorithm is developed to find from the phishing websites, the optimal features. This selected algorithm will be able to overcome the problem of over-fitting of the neural network to a great extend. These optimal features are then used to build an optimal classifier that detects phishing URLs by training the neural network. A theory called Fuzzy Rough Set(FRS) [24] was devised to a tool that finds the most appropriate features from a few standardized dataset. These features are then sent to a few classifiers for detection of phishing. To investigate the feature selection for FRS in building a generalized detection of phishing, the models by a different dataset of 14,000 website samples are trained.

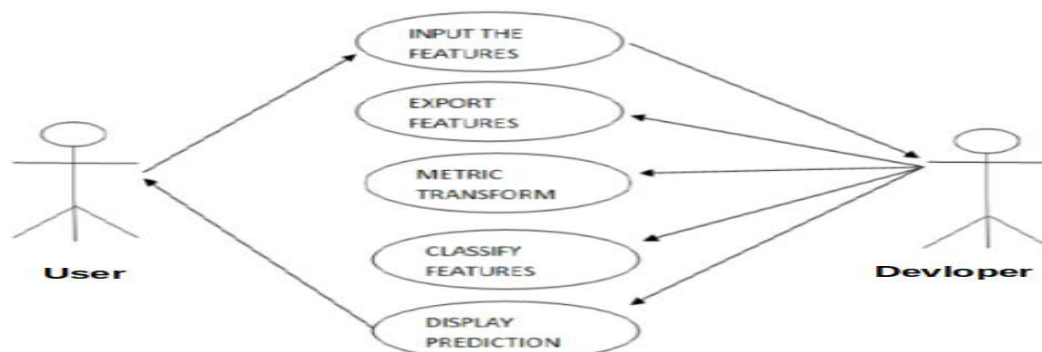
III. Proposed Work

ML methods plays a vital role in many applications of Cybersecurity and shall remain an encouraging path that captivates more such investigations. When coming to the reality, there are several barriers that are limitations during implementations As discussed, there are many approaches earlier proposed for detecting phishing website attack and they also have their own limitations. Therefore, the aim of the project is detection of phishing website attack using a novel Machine learning technique.

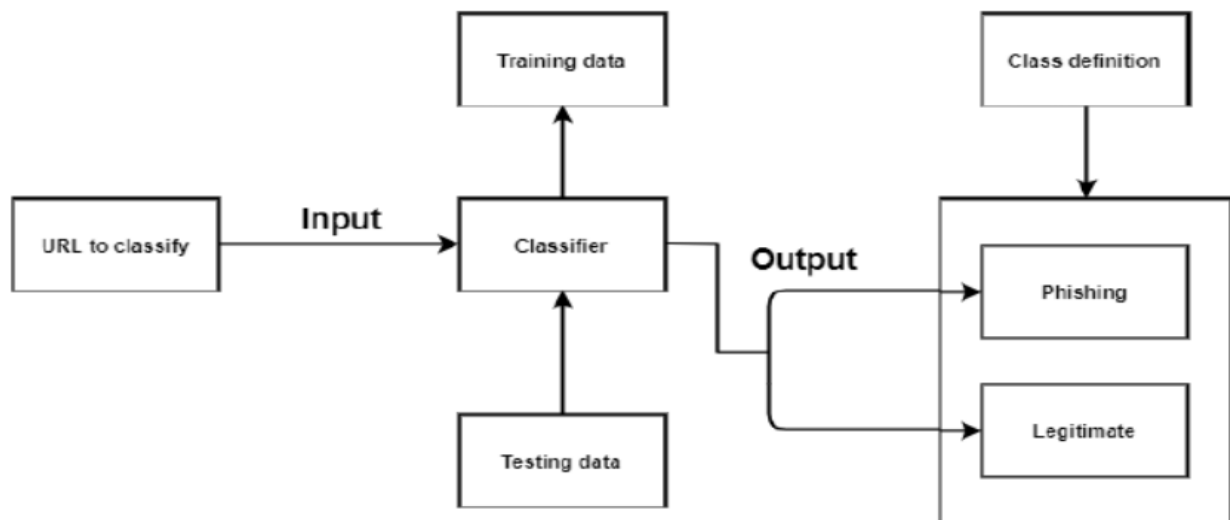
Features of Phishing sites detection System:

- Attractive and user friendly graphics user interface (GUI).
- Fast Phishing website Detection by “Random Forest Classifier” with highest accuracy.
- System independent, maintainable and Portable web application.
- Detection is in real time.
- Detection is based on input features of URL and model can retrain itself.

IV. Use Case Diagrams



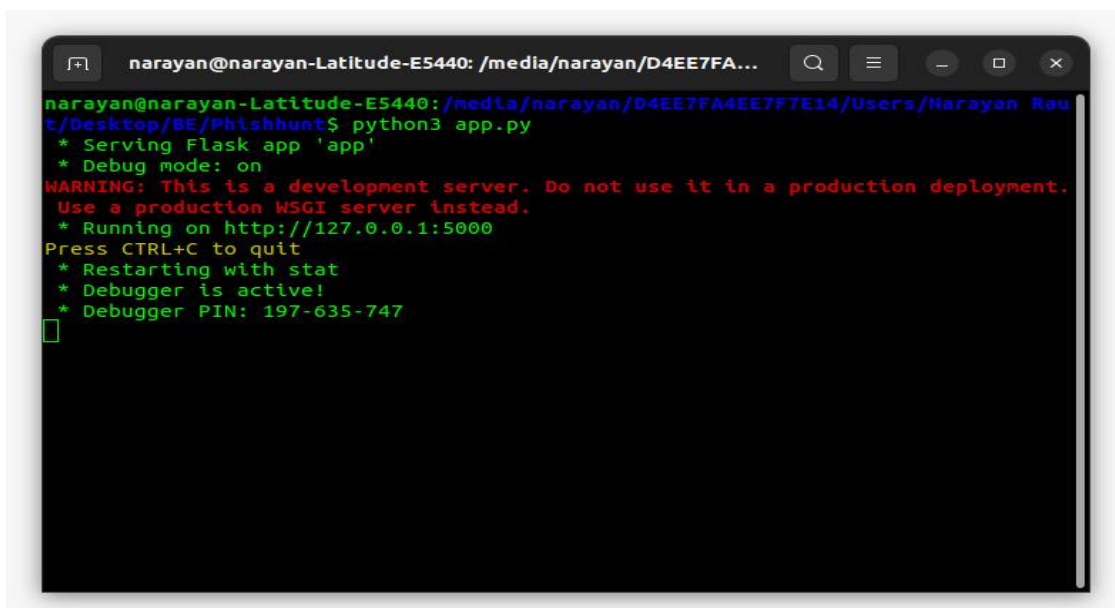
V. System Architecture



A system architecture or systems architecture is the conceptual model defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way supports reasoning about the structures and behaviour structure of the system. It can consist of system components and the sub-systems developed, will work together to implement the overall system. It have been efforts to formalize languages to describe system architecture, collectively the are called architecture description languages(ADLs) Various organizations can define systems architecture in different ways, including the fundamental organization of a system,embodied in its components, the relationships to each other and to the environment, and the principles governing its design and evolution. The figure shows system architecture of project.

VI. Results and Discussion

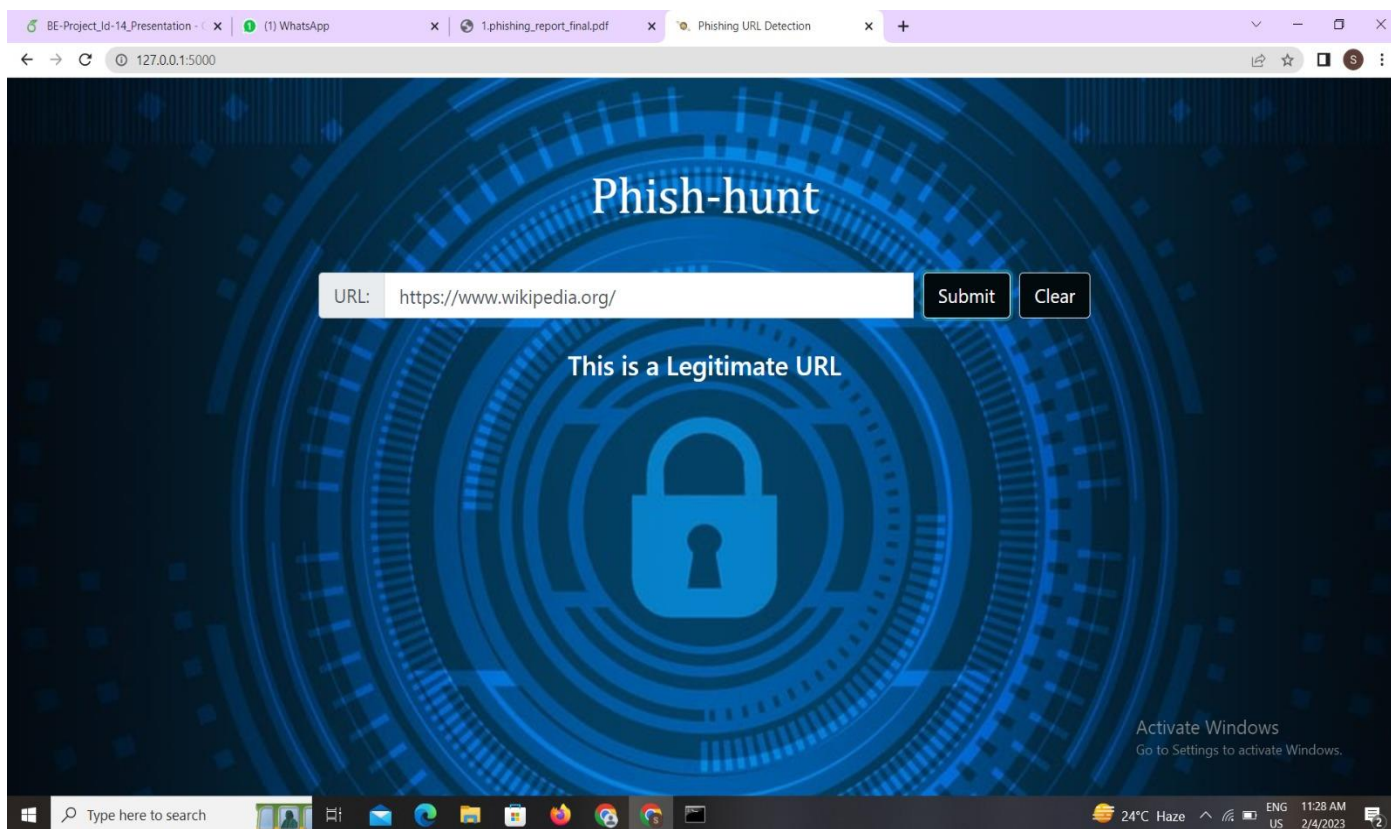
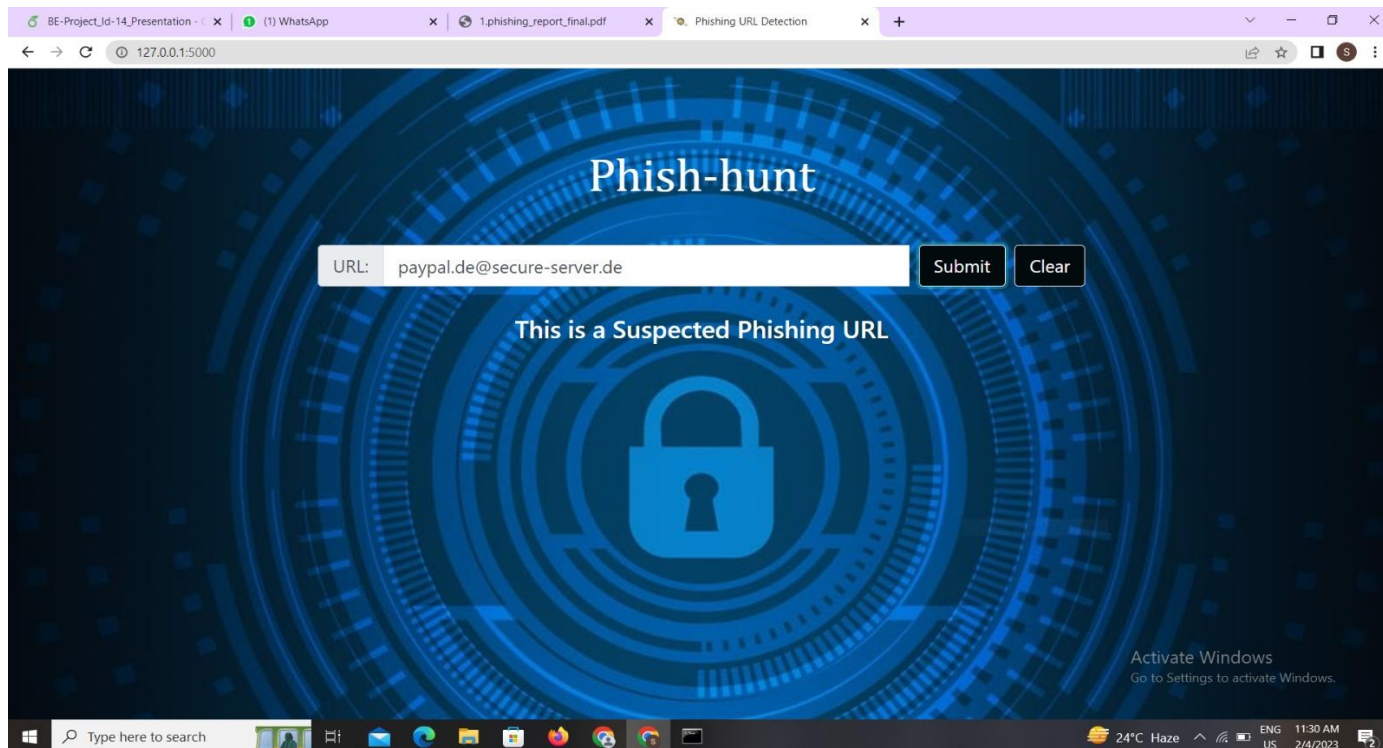
The results section is a section containing a description about the main findings of a research,whereas interprets the results for readers and provides the significance of the finding. Writing the results as separate sections allows to focus first on what results can obtained and set out clearly what happened in the experiments or investigations without worrying about the implication.



```

narayan@narayan-Latitude-E5440: /media/narayan/D4EE7FA...
narayan@narayan-Latitude-E5440:/media/narayan/D4EE7FA4EE7F7E14/Users/Narayan Rau:
t/Desktop/BE/Phishhunt$ python3 app.py
* Serving Flask app 'app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 197-635-747

```



VII. Conclusion and Future work

In the project, It has been a great pleasure for work on the exciting and challenging project. Phishing is defined as mimicking a creditable company's website aiming to take private information of a user. In order to eliminate phishing, different solutions proposed. However, only one single magic bullet cannot eliminate this threat completely. Data mining is a promising technique used to detect phishing attacks. In this paper, an intelligent system to detect phishing attacks is presented. We used different data mining techniques to decide categories of websites: legitimate or phishing. Different classifiers were used in order to construct accurate intelligent system for phishing website detection. Classification accuracy, area under receiver operating characteristic (ROC) curves (AUC) and F-measure is used to evaluate the performance of the data mining techniques. Results showed that Random Forest has outperformed best among the classification methods by achieving the highest accuracy 97.36percent. Random forest runtime are quite fast, and it can deal with different websites for phishing detection.

VIII. Bibliography

- [1] M. H. Alkawaz, S. J. Steven and A. I. Hajamydeen, "Detecting Phishing Website Using Machine Learning," 2020 16th IEEE International Colloquium on Signal Processing and Its Applications (CSPA), 2020, pp. 111-114, doi:10.1109/CSPA48992.2020.9068728.
- [2] V. K. Nadar, B. Patel, V. Devmane and U. Bhawe, "Detection of Phishing Websites Using Machine Learning Approach," 2021 2nd Global Conference for Advancement in Technology (GCAT), 2021, pp. 1-8, doi: 10.1109/GCAT52182.2021.9587682.
- [3] Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018, doi: 10.1007/978-981-10-8536-944.
- [4] Tony Yiu. "understanding random forest, how the algorithm works and why it is so effective". <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>, 06 2019.
- [5] J. Li and S. Wang. Phishbox: An approach for phishing validation and detection. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), pages 557–564, 2017.
- [6] S. Parekh, D. Parikh, S. Kotak, and S. Sankhe. A new method for detection of phishing websites: Url detection. In 2018 Second International Conference on Intelligent Communication and Computational Technologies (ICICCT), pages 949–952, 2018.
- [7] M. M. Yadollahi, F. Shoeleh, E. Serkani, A. Madani, and H. Gharaee. An adaptive machine learning based approach for phishing detection using hybrid features. In 2019 5th International Conference on Web Research (ICWR), pages 281–286, 2019.
- [8] E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu. Ofs-nn: An effective phishing websites detection model based on optimal feature selection and neural network. IEEE Access, 7:73271–73284, 2019.