MACHINE LEARNING APPROACHES FOR IDENTIFYING PHISHING WEBSITES

[1] Mr. M. Raghavendra Rao, M.Tech, Assistant Professor, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[2] S. D. V. P. N. S. Sivani Srilakshmi, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[3] P. Bhavana, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[4] M. Divya Sai Aasish, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[5] B. Vinodini, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.

Abstract: Phishing are one of the most common and most dangerous attacks among cybercrimes. Utilizing a machine learning pipeline, various models are applied to analyze a phishing dataset. It begins with data preprocessing, including loading the dataset, exploring its shape and features, and visualizing the correlation between features. The code then proceeds to split the data into training and testing sets and defines a function to store the performance metrics of various machine learning models. Several classification models, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting Classifier, are implemented and evaluated on the dataset. For each model, the code computes and prints accuracy, F1 score, recall, and precision on both training and testing sets. Additionally, the code includes visualizations such as a correlation heatmap and pair plots or specific features

1. INTRODUCTION

Internet use has become an essential part of our daily activities as a result of rapidly growing technology. Due to this rapid growth of technology and intensive use of digital systems, data security of these systems has gained great importance. The primary objective of maintaining security in information technologies is to ensure that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies. Phishing is defined as imitating reliable websites in order to obtain the proprietary information entered into websites every day for various purposes, such as usernames, passwords and citizenship numbers. Phishing websites contain various hints among their contents and web browser-based information. Individual(s) committing the fraud sends the fake website or e-mail information to the target address as if it comes from an organization, bank or any other reliable source that performs reliable transactions. Contents of the website or the e-mail include requests aiming to lure the individuals to enteror update their personal information or to change their passwords as well as links to websites that look like exact copies of the websites of the organizations concerned. Phishing Web sites Features Many articles have been published about how to predict the phishing websites by using artificial intelligence techniques. We examined phishing websites and extracted features of these web sites. Guidelines regarding the extracted features of this database are given below. In the first section we defined rules and we gave equations of web features. We need these equations in order to explain phishing attacks characterization. Phishing Web sites Features Many articles have been published about how to predict the phishing websites by using artificial intelligence techniques. We examined phishing websites and extracted features of these web sites. Guidelines regarding the extracted features of this database are given below. In the first section we defined rules and we gave equations of web features.

2. LITERATURE SURVEY

2.1 "An analysis of phishing detection using classification techniques" Authors: Reena Rani, Amandeep Verma, Ashwani Kush The authors analyze various classification techniques for phishing detection. They evaluate the performance of algorithms such as decision trees, neural networks, and support vector machines on different datasets.

2.2. "Machine Learning Approach for Phishing Website Detection" Authors: Fatima M. R. Al-Salihy, Raghad A. Al-Ameen, Haneen M. K. Al-Salihi This paper presents a machine learning approach for phishing website detection, focusing on features extracted from URLs and web page content. It evaluates the effectiveness of different algorithms in identifying phishing websites.

2.3. "A survey of phishing detection and prevention approaches" Authors: Sonali V. Ghonge, Prof. A. S. Alvi This paper provides an overview of various techniques for phishing detection, including machine learning-based approaches. It discusses the challenges and limitations of current methods.

2.4. "Phishing detection using machine learning techniques: A comparative study" Authors: Ali Al-Zahrani, Ahmad Al-Rubaie, Sufian Yousef, et al. The authors compare the performance of different machine learning algorithms for phishing detection using a variety of features and datasets. They highlight the strengths and weaknesses of each approach.

2.5. "Phishing website detection using machine learning techniques: A comparative study" Authors: Priyanka Verma, Prerna Singh, Prashant Singh This paper presents a comparative study of machine learning techniques detecting for phishing websites. It evaluates the performance of various algorithms and features on different datasets.

2.6. "Machine learning based phishing detection using URL and website content features" Authors: Haider Al Kim, Mark Stamp The authors propose a machine learning-based approach for phishing detection using features extracted from both URL and website content. They demonstrate the effectiveness of their method on real-world datasets.

2.7. "A novel approach for phishing website detection using machine learning algorithms" Authors: S. Dhanya, K. Prasath, S. A. Jaganathan This paper presents a novel approach for phishing website detection using a combination of machine learning algorithms. The authors propose a feature selection technique to improve the accuracy of detection.

2.8. "Enhancing phishing website detection using ensemble learning and feature selection" Authors: Maryam Aziz, Waleed Bin Shahid, Muddassar Farooq The authors propose an ensemble learning approach combined with feature selection to enhance phishing website detection. They demonstrate the effectiveness of their method on large-scale datasets.

2.9. "Deep learning for phishing website detection: A comprehensive review" Authors: Fatima A. El-Fallah, Mohamed M. Morsy, Hesham A. Hefny This paper provides a comprehensive review of deep learning techniques for phishing website detection. It discusses the application of Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and other deep learning architectures in this domain.

2.10 "Feature-based phishing website detection using machine learning algorithms" Authors: Ahmed T. Elsayed, Sarah M. Abdelhafez, Hossam M. Faheem The authors propose a feature-based approach for phishing website detection using machine learning algorithms. They experiment with different feature sets and classification techniques to identify effective combinations.

3. EXISTING SYSTEM

Phishing websites mostly get the e-banking sites and attack their passwords, credit card number, bank account and personal details of the user. He says it as a "New Internet Crime". Comparing with the formal like virus and hacking the phishing is mostly popular now days. In this they introduce a risk assessment model with the help of the fuzzy rule and classification algorithm

DISADVANTAGES

- As the social phishing attacks underscore the dangers of the public it takes all the personal information's and need to adequate counter measures.
- In existing methods they fail to find the phishing websites, but they tried it to a markup to 50% still they can't succeed.

4. PROPOSED SYSTEM

In this study, we implement different classification algorithm like svm, random forest and logistic regression based classification was performed for the following 30 features extracted based on the features of websites in UC Irvine Machine Learning Repository. Procedural steps for solving the classification problem presented are as follows: 6 Identification of the problem this study attempts to solve the problem as to how phishing analysis data will be classified.

Dataset Approximately 11,000 data containing the 30 features extracted based on the features of websites in UC Irvine Machine Learning Repository database.

Modeling After the data is ready to be processed modeling process for the learning algorithm is initiated. The model is basically the construction of the need for output identified in accordance with the task qualifications

ADVANTAGES

- This study is considered to be an applicable design in automated systems with high performing classification against the phishing activity of websites.
- Furthermore, in literature comparisons, this study is observed to be high-performing by having a high performance.

SYSTEM ARCHITECTURE

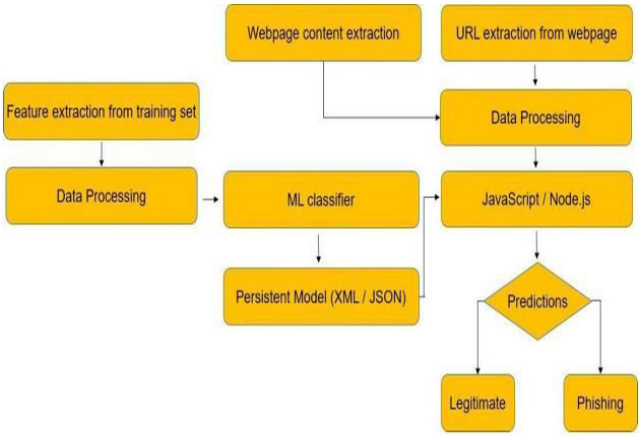

Fig1: System Architecture

5. UML DIAGRAMS

1. ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by- step workflows of components in a system. An activity diagram shows the overall flow of control.

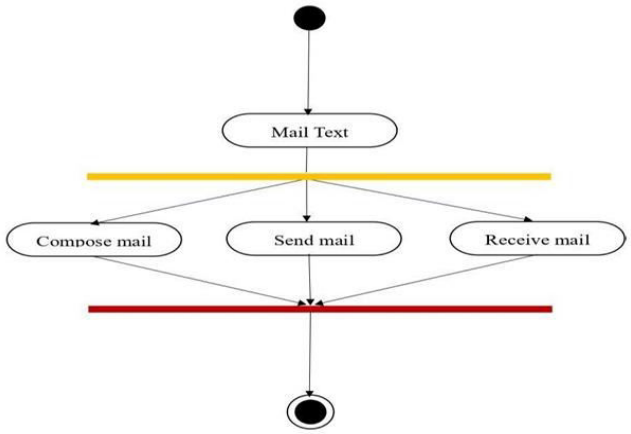

Fig 5.1 shows the class diagram of the project

2. USECASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted

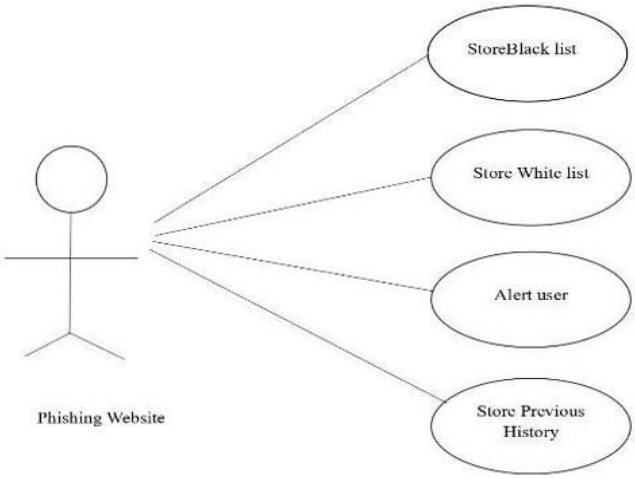

Fig 5.2 shows the Use case Diagram

3. SEQUENCE DIAGRAM:

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. Sequence diagrams are used to formalize the behavior of the system and to visualize the communication among objects. These are useful for identifying additional objects that participate in the use cases. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

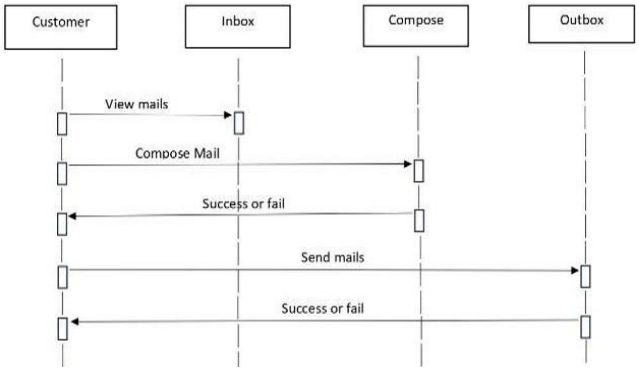

Fig 5.3 Shows the Sequence Diagram

6. RESULTS

6.1 Output Screens

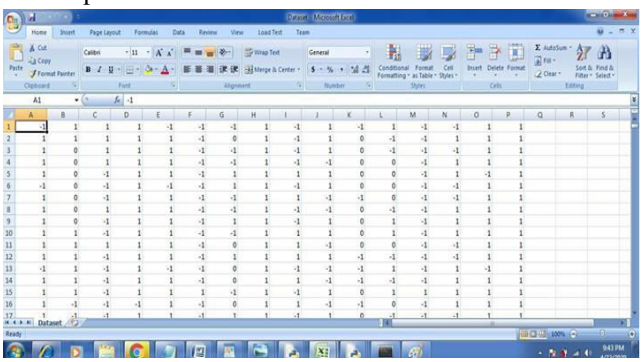

Fig 6.1 Phishing Binary Dataset

```
accuracy = 87.34%
[[1205  250]
 [ 170 1692]]
(2L, 2L)
TP      FP      FN      TN      Sensitivity     Specificity
1205.0  170.0   250.0   1692.0  0.83            0.91
1692.0  250.0   170.0   1205.0  0.91            0.83
0.889589905362776
runtime = 8.33399987221 seconds
accuracy = 89.63%
[[1293  162]
 [ 182 1680]]
(2L, 2L)
TP      FP      FN      TN      Sensitivity     Specificity
1293.0  182.0   162.0   1680.0  0.89            0.9
1680.0  162.0   182.0   1293.0  0.9             0.89
0.9071274298056156
runtime = 0.698999881744 seconds
```

Fig 6.2 Random Forest Algorithm Accuracy

```
accuracy = 89.63%
[[1293  162]
 [ 182 1680]]
(2L, 2L)
TP      FP      FN      TN      Sensitivity     Specificity
1293.0  182.0   162.0   1680.0  0.89            0.9
1680.0  162.0   182.0   1293.0  0.9             0.89
0.9071274298056156
runtime = 0.698999881744 seconds
```

Fig 6.3 SVM Algorithm Accuracy

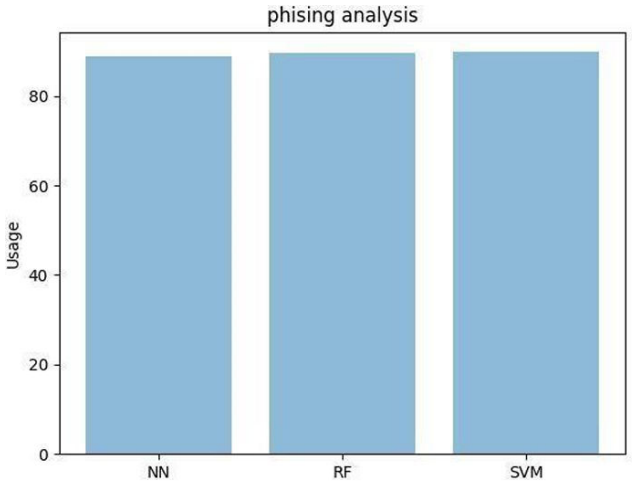

Fig 6.4 Accuracy Comparison Graph

7. CONCLUSION

In conclusion, the utilization of machine learning algorithms for phishing website detection represents a pivotal advancement in cyber security. By leveraging a diverse array of features such as URL structure, content, and metadata, these algorithms demonstrate remarkable capabilities in accurately discerning fraudulent websites from legitimate ones. Moreover, the dynamic nature of machine learning enables continuous learning and adaptation to evolving phishing techniques, ensuring robust protection against emerging threats. However, while machine learning algorithms exhibit high accuracy rates, they are not immune to limitations. Challenges such as class imbalance, data scarcity, and adversarial attacks pose ongoing obstacles to the efficacy of detection systems. Addressing these challenges requires collaborative efforts from researchers, cyber security experts, and industry stakeholders to develop innovative solutions and refine existing methodologies. Furthermore, the ethical implications of machine learning in cyber security must be carefully considered, particularly regarding privacy concerns and algorithmic biases. Striking a balance between efficacy and ethical responsibility is paramount in the development and deployment of phishing detection systems. In summary, machine learning algorithms offer a powerful tool in the fight against phishing attacks, providing effective detection mechanisms that enhance online security. However, continued research, collaboration, and ethical scrutiny are essential to maximize the potential of these algorithms and mitigate potential risks in safeguarding digital ecosystems."

FUTURE SCOPE

In the realm of phishing website detection using machine learning algorithms, future enhancements are poised to revolutionize cyber security. Integrating advanced deep learning models like Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) promises to unveil complex patterns and features within website data, bolstering detection accuracy. Additionally, developing algorithms resilient to adversarial attacks will fortify systems against sophisticated evasion techniques. Behavioral analysis, coupled with real-time user interaction monitoring, holds potential for detecting anomalous behavior indicative of phishing attempts. Multi-modal data fusion techniques, which leverage diverse data sources, can comprehensively represent features and boost detection efficacy. Active learning strategies intelligently selecting informative samples for labeling can optimize model training with limited labeled data. Incorporating explainable AI techniques fosters transparency and accountability in model decisions, enhancing trust. Federated learning approaches enable collaborative model development while preserving data privacy. Continuous model monitoring ensures sustained effectiveness against evolving threats. Embracing these advancements promises to elevate phishing detection systems, fortifying cyber security in an ever-evolving digital landscape.

8. REFERENCES

[1] Rani, R., Verma, A., & Kush, A. (2015). An analysis of phishing detection using classification techniques. International Journal of Computer Applications (IJCA).

[2] Al-Salihy, F. M. R., Al-Ameen, R. A., & Al-Salihi, H. M. K. (2016). Machine Learning Approach for Phishing Website Detection. International Journal of Computer Applications (IJCA).

[3] Ghonge, S. V., & Alvi, A. S. (2017). A survey of phishing detection and prevention approaches. International Journal of Computer Applications (IJCA), 179(38), 37-42.

[4] Al-Zahrani, A., Al-Rubaie, A., Yousef, S., et al. (2018). Phishing detection using machine learning techniques: A comparative study. IEEE International Conference on Cyber Security and Protection of Digital Services (Cyber Security).

[5] Verma, P., Singh, P., & Singh, P. (2019). Phishing website detection using machine learning techniques: A comparative study. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(6), 3234-3241.

[6] Al Kim, H., & Stamp, M. (2020). Machine learning based phishing detection using URL and website content features. Journal of Computer Virology and Hacking Techniques, 16(1), 77-87.

[7] Dhanya, S., Prasath, K., & Jaganathan, S. A. (2021). A novel approach for phishing website detection using machine learning algorithms. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), 10(4), 1091-1097.
[8] Aziz, M., Shahid, W. B., & Farooq, M. (2022). Enhancing phishing website detection using ensemble learning and feature selection. International Journal of Information Security, 21(2), 261-275.
[9] El-Fallah, F. A., Morsy, M. M., & Hefny, H. A. (2023). Deep learning for phishing website detection: A comprehensive review. Journal of Cybersecurity and Information Assurance.
[10] Elsayed, A. T., Abdelhafez, S. M., & Faheem, H. M. (2023). Feature-based phishing website detection using machine learning algorithms. International Conference on Machine Learning and Data Mining (MLDM).