iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62467

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Detection of Phishing Websites using Machine Learning

Pathare Ujjwala[1], Lokhande Gayatri[2], Machale Manjushri[3], Gunjal Kajal[4], Prof Gunaware N.G.[5]
Department of Computer Engineering, HSBPVT's GOI FOE, Kashti, Maharashtra, India

Abstract: ***Malicious links or attachments through emails that can perform various functions, including capturing the victim's login credentials or account information is send by phishing. these victims, cause money loss, and identity theft is harmful. In these the phishing problem by developing an extension for the Google Chrome web browser we can contribute to solve this. we used JavaScript PL in development of these feature. a combination of Blacklisting and semantic analysis methods was used to preventing the attack. and the proposed solution was tested and it can be compared to existing approache.***
Keywords: phishing, Machine Learning, Logistic Regression, URL features , Detetction process, Legitimate, truth worthy, false positive, false negative.

I. INTRODUCTION

1) Social engineering and cyber attack is commonly used by phishing
2) Through such attacks, the phisher targets naive online, users by tricking then into revealing confidential information, with the purpose of using it fraudulently.
3) In order to avoid getting phished,
4) Have a backlist which requires the knowledge of website being detected as phishing.
5) The machine learning is proven to be most effective method than other methods.

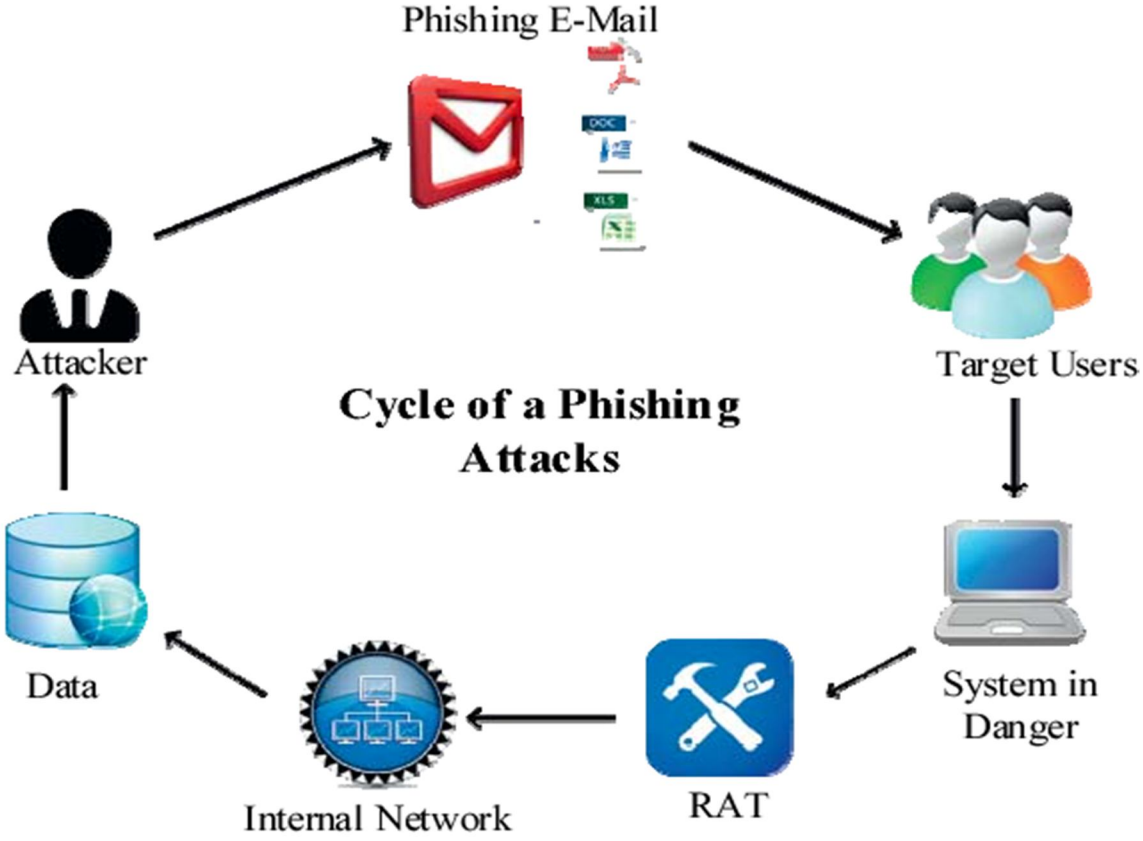


Fig.1: Data Flow

II. RELATED WORK

Here's an outline of related work in this field:

1) Data Collection: Gather a dataset of URLs along with their corresponding labels indicating whether they are phishing or legitimate websites. These labels can be obtained from various sources, including security organizations, blacklists, or manual labeling.
2) Feature Extraction: Extract relevant features from the URLs and associated metadata. These features can include:
3) URL-based features: Length of the URL, presence of special characters, use of IP address instead of domain name, etc.
4) Content-based features: Presence of forms, suspicious scripts, iframes, etc.
5) Domain-based features: Domain age, WHOIS information, SSL certificate validity, etc.
6) Website content analysis: Analysis of webpage content for phishing indicators, such as presence of login forms, misspelled words, or suspicious URLs.
7) Data Preprocessing: Clean the data by handling missing values, encoding categorical variables, and normalizing numerical features.
8) Model Training: Select a suitable machine learning algorithm (such as decision trees, random forests, support vector machines, or neural networks) and train it on the preprocessed data. During training, the model learns to classify URLs as phishing or legitimate based on the extracted features.
9) Model Evaluation: Evaluate the trained model's performance using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, or area under the ROC curve (ROC-AUC). This step helps assess how well the model generalizes to unseen data and identifies any potential issues such as overfitting or underfitting.
10) Model Deployment: Deploy the trained model into a production environment where it can classify new URLs in real-time. This may involve integrating the model into existing security systems or web browsers to automatically flag or block suspicious websites.
11) Monitoring and Updating: Continuously monitor the model's performance in the production environment and update it as necessary to adapt to evolving phishing techniques and new threats. This may involve retraining the model with new data or fine-tuning its parameters to improve its effectiveness.

A. *Benefits and Challenges*

Detecting phishing websites using machine learning offers several benefits and also presents certain challenges:

1) *Benefits:*

- Automation: Machine learning models can automate the detection process, enabling the identification of phishing websites at scale without the need for manual intervention.
- Real-time Detection: ML models can classify URLs in real-time, allowing for immediate response to new phishing attempts and minimizing the window of vulnerability.
- Adaptability: Machine learning algorithms can adapt to evolving phishing techniques and patterns by continuously learning from new data, making them more effective at detecting previously unseen phishing attempts.
- Feature Extraction: ML models can automatically extract relevant features from URLs, such as domain age, presence of suspicious keywords, and similarity to known phishing websites, improving the accuracy of detection.
- Scalability: ML-based phishing detection systems can scale to handle large volumes of incoming URLs, making them suitable for use in enterprise-level cybersecurity solutions.

2) *Challenges:*

- Data Quality: The quality and quantity of labeled data are crucial for training effective machine learning models. Obtaining large, diverse, and accurately labeled datasets for phishing detection can be challenging.
- Imbalanced Data: Phishing datasets often suffer from class imbalance, with a significantly higher number of legitimate URLs compared to phishing URLs.
- Feature Engineering: Identifying informative features for phishing detection and engineering them into a suitable format for machine learning models can be complex and time-consuming.
- Generalization: Ensuring that machine learning models generalize well to unseen phishing techniques and variations is essential. Models that overfit to specific patterns may fail to detect new phishing attempts.

Adversarial Attacks: Phishers may deliberately manipulate their websites to evade detection by machine learning models, introducing adversarial examples that exploit vulnerabilities in the model's decision-making process.
Interpretability: Deep learning models, in particular, are often considered "black boxes," making it challenging to interpret their decisions and understand the rationale behind phishing classifications.
Privacy Concerns: Phishing detection systems that rely on analyzing user behavior or content may raise privacy concerns, as they involve processing potentially sensitive information.

III. METHODOLOGY

The methodology for detecting phishing websites using machine learning involves several structured steps, including data collection, feature extraction, model training, evaluation, and deployment. below are the information about methodology:

1) Data Collection

- Phishing URLs: Obtain lists of known phishing URLs from sources like PhishTank, OpenPhish, and other security databases.
- Legitimate URLs: Gather legitimate URLs from reliable sources such as Alexa's top sites, DMOZ, and other trustworthy databases.

2) Feature Extraction

- Length of URL: Phishing URLs often have unusual lengths.
- Use of Special Characters: Presence of '@', '-', '_', '=', etc.
- Domain Age: Newly registered domains are often used for phishing.
- Presence of IP Address: Phishing URLs might use IP addresses instead of domain names.
- Subdomain Count: Phishing URLs often have multiple subdomains.

3) Data Preprocessing

- Clean and Normalize Data
- Remove Redundant Data: Filter out irrelevant data points and duplicates.
- Normalize Features: Scale features to a consistent range, typically [0, 1].
- Label Encoding
- Encode Labels: Convert categorical labels (phishing or legitimate) into numerical values.

4) Model Training

- Split Data
- Training and Testing Sets: Split the dataset into training (70-80%) and testing (20-30%) sets to evaluate model performance.
- Choose and Train Model
- Select Algorithm: Choose appropriate algorithms such as Random Forest, SVM, Logistic Regression, etc.
- Train the Model: Train the chosen model on the training dataset.

5) Model Evaluation

- Evaluate Model Performance
- Metrics: Use accuracy, precision, recall, F1-score, and ROC-AUC to evaluate the model.
- Cross-Validation
- K-Fold Cross-Validation: Perform cross-validation to ensure the model's robustness and to avoid overfitting.

6) Deployment

- Real-time Detection
- Integrate Model: Deploy the trained model into a web service or application.
- Monitor Performance: Continuously monitor the model's performance in real-time and update as needed.

7) Continuous Improvement

- Model Updates
- Update Dataset: Regularly update the dataset with new phishing and legitimate URLs.
- Retrain Model: Periodically retrain the model with updated data to maintain high detection accuracy.
- Feedback Loop: Incorporate user feedback to improve the model and correct false positives/negatives.
- Collect phishing and legitimate URLs.
- Feature Extraction

IV. SOFTWARE REQUIREMENT

Software Requirements

- Languages: Python,
- Libraries: scikit-learn, TensorFlow/Keras, PyTorch, Pandas, NumPy
- Visualization: Matplotlib, Seaborn, Plotly
- Development: VS Code

V. ALGORITHMS

Here are some common algorithms used in detection of phishing websites using machine learning, along with brief descriptions of their roles:

1) Decision Trees

Decision trees are simple yet powerful models that split the data into subsets based on feature values. They work well for classification problems and can be easily visualized.

2) Random Forest

this algorithm is an ensemble method that builds multiple decision trees and merges them to get a accurate .

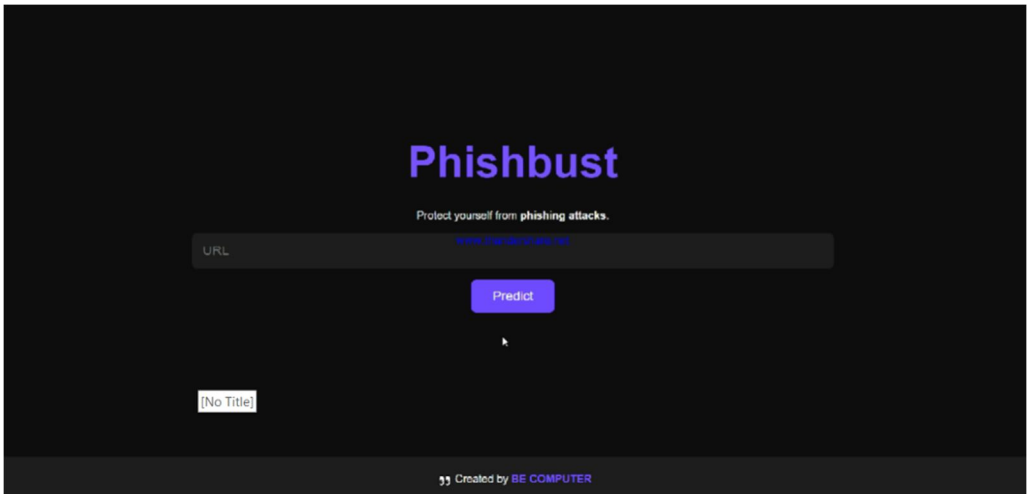
3) Support Vector Machines (SVM)

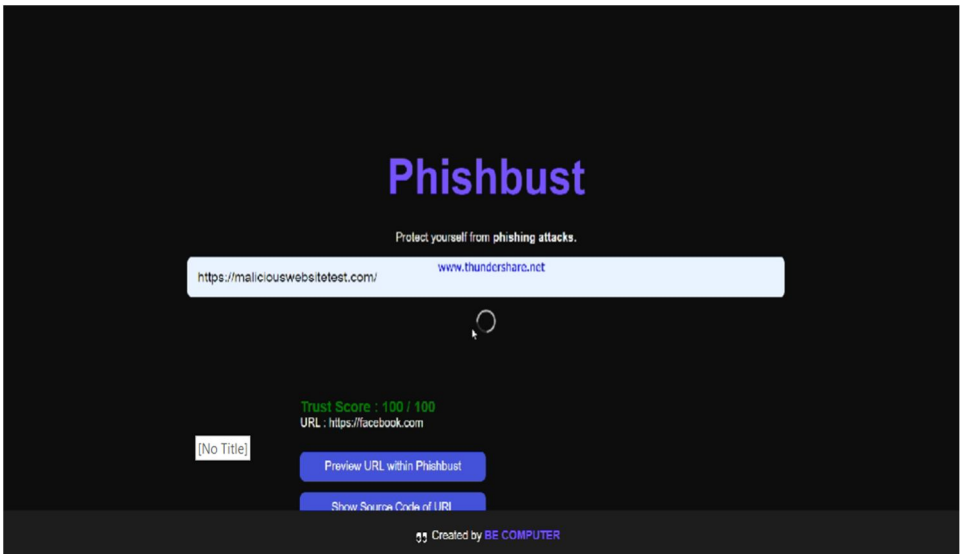
SVM is a classification technique that finds the optimal hyperplane separating different classes in the feature space.
Pros: Effective in high-dimensional spaces and robust to overfitting.
Cons: Requires careful tuning of parameters and can be computationally expensive.

4) Logistic Regression

Logistic regression is a regression analysis in which the outcome variable are binary.

VI. RESULT





VII. CONCLUSION

At present generation attackers are more in networks, and phishing has become major security problem, causing many losses by hacking the legal data that are used by the user

Phishers set up their own fake websites which are exactly like the original websites including applying a DNS server name, setting up a web server, and creating web pages similar to destination websites.

Working on this project is very knowledgeable and worth the effort.

These should classify the inputted URL as legitimate or phishing with the use of the saved model.

VIII. FUTURE SCOPE

1) Enhanced Detection Accuracy

- Improved Algorithms**:** Continuous development of sophisticated algorithms, such as deep learning and ensemble methods, can enhance the accuracy of phishing detection systems.
- Real-Time Analysis**:** Advancements in real-time data processing will enable faster and more accurate detection of phishing websites as they appear.

2) Adaptive and Resilient Models

- Adversarial Learning**:** Incorporating adversarial learning techniques to make models more resilient against evolving phishing tactics.
- Continuous Learning**:** Implementing models that can learn and adapt over time with new data, ensuring they remain effective against new phishing strategies.

3) Integration with Browsers and Email Services

- Browser Integration: Embedding machine learning-based phishing detection directly into web browsers to provide real-time alerts to users.
- Email Filtering: Enhancing email services with advanced phishing detection capabilities to filter out malicious emails before they reach users' inboxes.

4) Improved Feature Extraction

- Advanced Feature Engineering: Utilizing more sophisticated feature extraction techniques to identify subtle patterns that indicate phishing.
- Behavioral Analysis: Analyzing user behavior on websites to detect unusual activities indicative of phishing.

5) Collaboration and Data Sharing

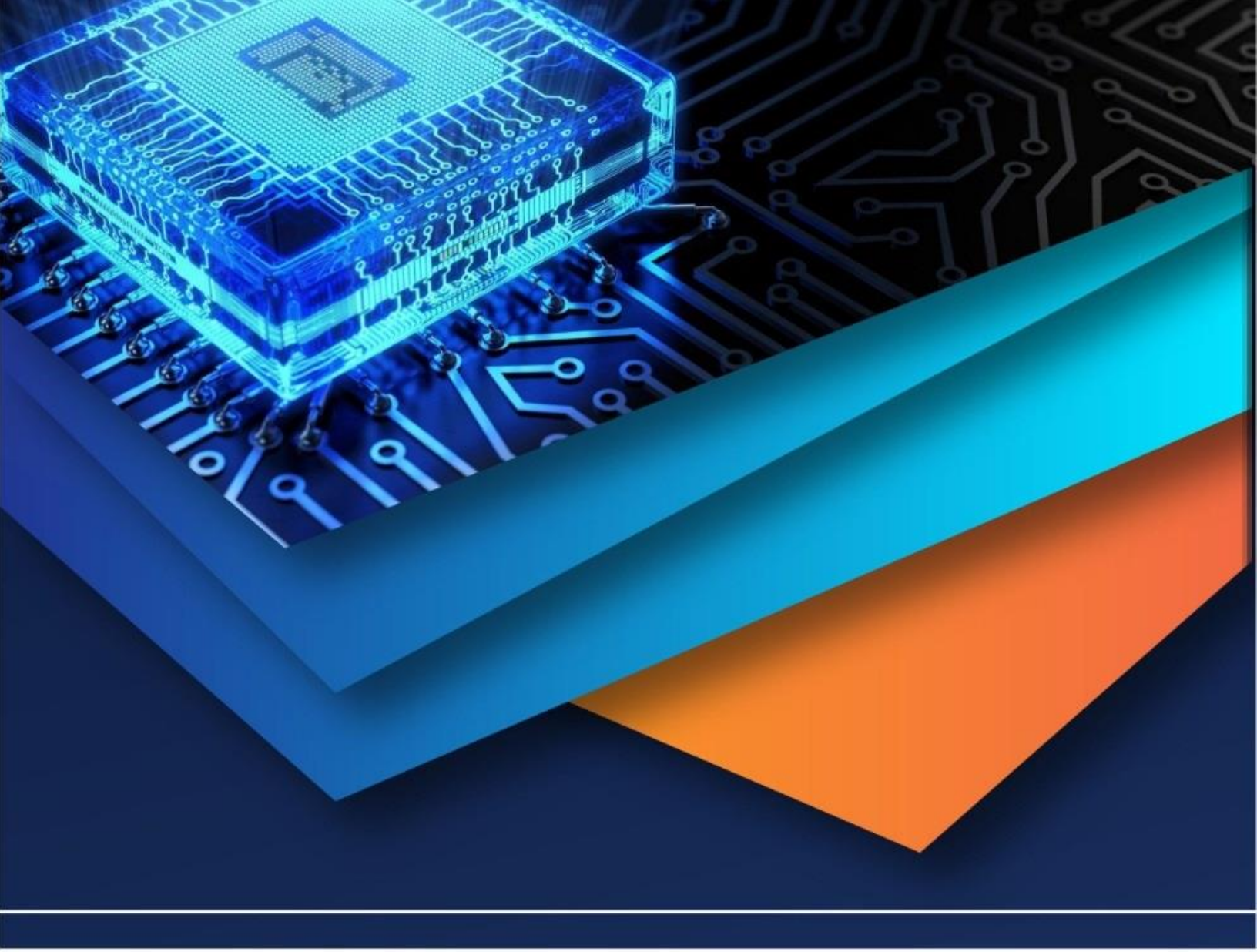
- Shared Databases**:** Creating shared databases of known phishing sites to improve the collective intelligence of detection systems.
- Cross-Platform Collaboration: Collaborating across different platforms and services to create a unified defense against phishing attacks.

6) Regulatory and Policy Support

- Government and Industry Standards: Development of standards and regulations to mandate the use of advanced phishing detection techniques.
- Awareness Programs: Implementing educational programs to increase awareness about phishing and the role of machine learning in combating it.

REFERENCES

[1] https://www.slideshare.net/ummeayesha/phishing-detection
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8504731/
[3] https://search.yahoo.com/search?fr=mcafee&type=E210US885G0&p=mathmatical+model+of+detection+of+phishing+website+using+machine+learning
[4] https://www.researchgate.net/publication/327340841_Detection_and_Prevention_of_Phishing_Attack_Using_Linkguard_Algorithm
[5] https://www.researchgate.net/publication/355263255_Detecting_phishing_websites_using_machine_learning_technique
[6] https://chat.openai.com/c/a2c96f1c-1bb9-4d48-8f44-461610d4af56



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY