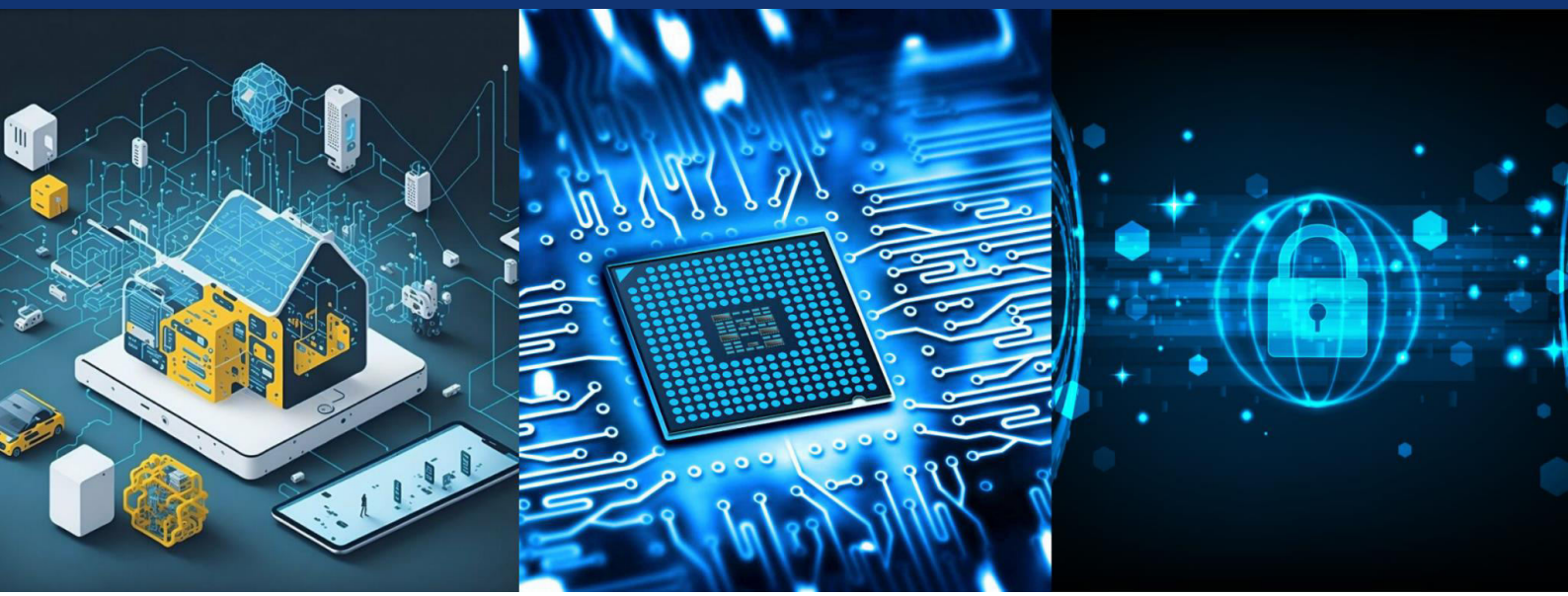
IJIRCCE

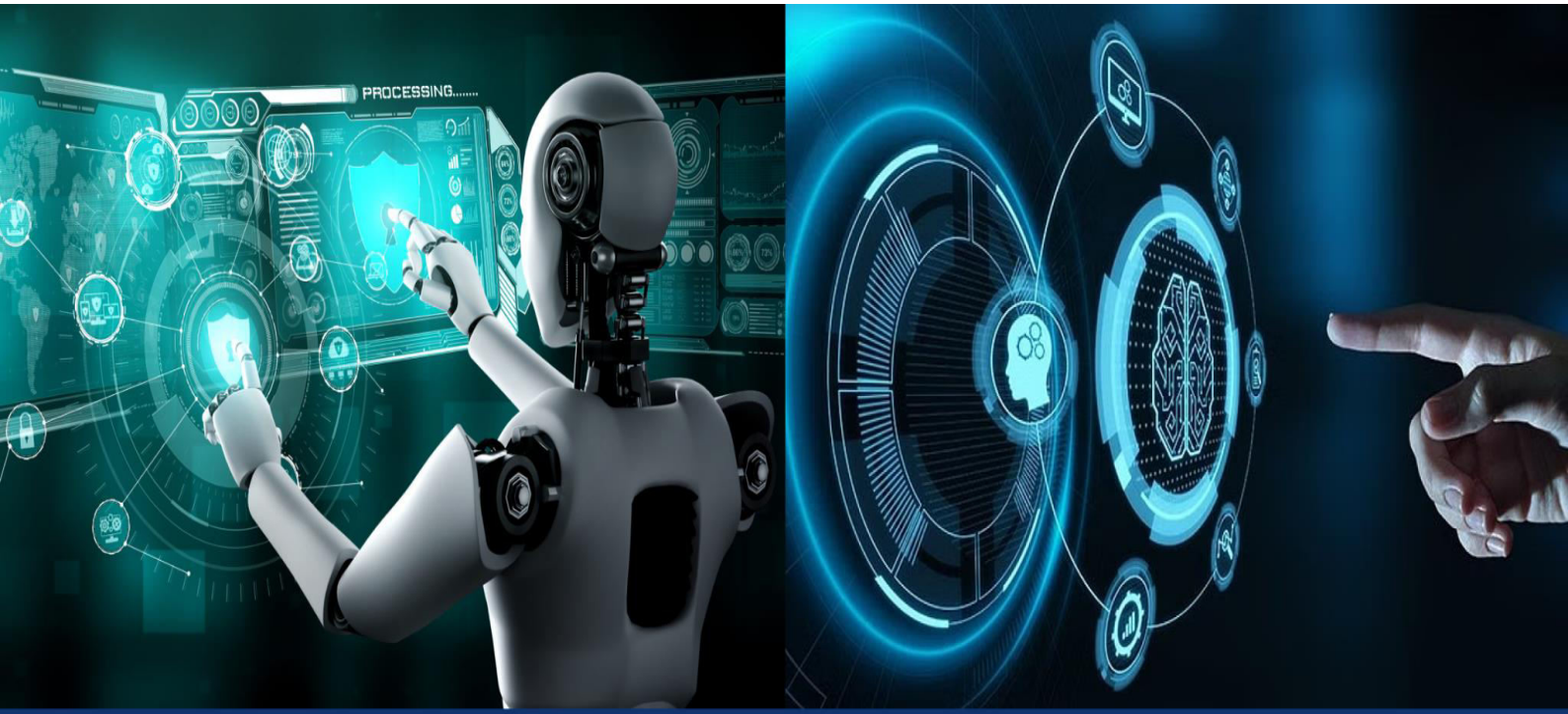
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Phishing URL Detection Behaviour Analysis System

Sindhu Priyanka Chadalavada [1], Sanapala Aparna [2], Yenugu Jaya Saisudha [3], Setti Sajith Kumar [4], Nerusu Srinivasa Rao [5]

Associate Professor, Department of CSE, Eluru College of Engineering & Technology, Eluru, India[1]

B. Tech Student, Department of CSE, Eluru College of Engineering & Technology, Eluru, India[2,3,4,5]

ABSTRACT: Many people utilize different websites to make payments and make purchases of goods online. Phishing websites are cyber-attacks that aim to steal sensitive data from internet users, such as financial information and login credentials. Phishing is still one of the best and most successful ways for hackers to cheat users out of our money and steal our personal and financial information.

These kinds of websites are commonly known as Phishing websites. The "Phishing Website Detection Using Machine Learning Classification Data Mining Algorithms to Identify and Anticipate Phishing Websites" categorizes websites into two groups: phishing and legitimate. We suggested an intelligent, adaptable, and efficient system based on a classification data mining algorithm to identify and anticipate phishing websites. To extract the phishing data set criteria and classify them according to their legality, we used classification algorithms and methodologies.

The URL, domain identification, and other key feature scan be used to identify the phishing website. The objective is to develop a robust and reliable system capable of identifying phishing websites with high accuracy.

KEYWORDS: Detection, Data Mining, Identify and Phishing.

I. INTRODUCTION

Phishing website detection involves analyzing URL features to identify potential threats. Key features include domain length, subdomain count, use of special characters, and presence of HTTPS. Machine learning models can be trained on these features to distinguish between legitimate and phishing URLs, enhancing cyber security efforts. In the realm of phishing website detection, the identification of potential threats relies heavily on various URL features. One critical aspect is the length of the domain, as legitimate domains typically adhere to a standard length, while phishing URLs often exhibit longer or unconventional domain names. Another significant factor is the count of subdomains, with an anomalous number potentially indicating a phishing attempt, as legitimate websites tend to have a limited number of subdomains. Examining the use of special characters in URLs is also crucial, as phishing sites may employ these to mimic legitimate domains, and analyzing their presence and placement aids in detection. The use of HTTPS is a common practice for secure communication on legitimate websites, and the absence or deceptive usage of HTTPS, such as employing HTTP, on a site may suggest a potential threat. Domain reputation, assessed through databases storing information about known phishing domains, is another valuable feature in determining the authenticity of a given domain. Additionally, the presence of multiple redirects or the use of URL shortness can be indicative of phishing attempts. Analyzing the re directionchain helps in identifying suspicious behavior.

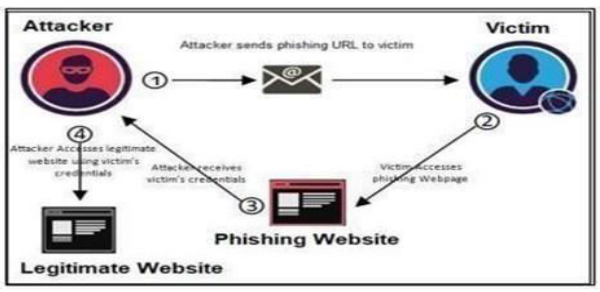


Fig 1:Phishing Website Life cycle

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

1.1 Motivation of the Project

- The project focuses on addressing the significant threat posed by phishing websites, which aim to steal sensitive information from internet users, including financial data and login credentials. Phishing attacks remain one of the most successful tactics used by hackers to deceive users and extract valuable personal and financial information.
- To combat this threat, the project proposes the development of an intelligent, adaptable, and efficient system for detecting and anticipating phishing websites. The system employs machine learning classification algorithms and data mining techniques to categorize websites into two groups : phishing and legitimate.

1.2 Problem Definition

With existing phishing detection systems: their inefficiency in predicting new threats and adapting to ever-changing phishing tactics by cybercriminals. If you're working on a project or researching this topic, I can help you brainstorm solutions, find recent developments, or refine the problem statement further.

1.3 Objective of the project

Objectives for the project "Phishing Website Detection Using Machine Learning Classification Data Mining Algorithms to Identify and Anticipate Phishing Websites": Develop a robust and reliable system capable of accurately identifying phishing websites to mitigate the risk of financial and personal information theft.

- Utilize machine learning classification algorithms to categorize websites into two groups: phishing and legitimate, based on key features such as URL and domain identification.
- Explore various classification algorithms and methodologies to identify the most effective approach for detecting and anticipating phishing websites.
- Train and evaluate machine learning models using collected data sets to ensure high accuracy and reliability in identifying phishing websites.
- Design an intelligent, adaptable, and efficient system capable of continuously learning and adapting to new phishing tactics and techniques.

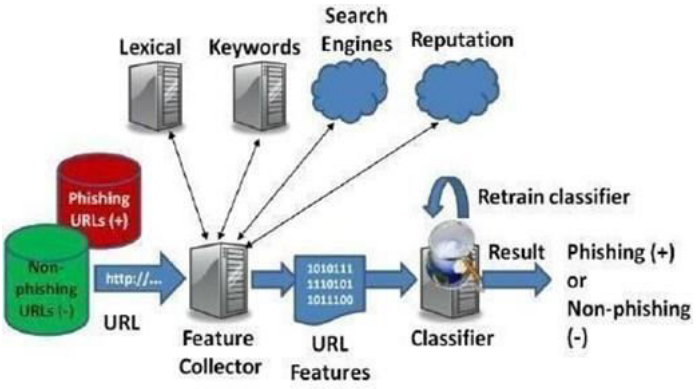


Fig 2: Detection of spam url

1.4 Scope

The scope for phishing website detection using URL features is vast and pivotal in addressing the ever-evolving landscape of cyber threats. Leveraging URL characteristics provides a comprehensive framework to proactively identify and mitigate phishing attacks. This approach extends across multiple domains, including user protection, financial security,and data integrity. The scope encompasses the development of advanced machine learning models trained on diverse datasets, enabling the detection of subtle patterns and variationsion deceptive URLs. As phishing techniques continue to grow in sophistication, the scope for URL feature analysis widens, creating opportunities for continuous innovation in cyber security measures.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE SURVEY

We made this project by referencing to some Journals and Research papers in IEEE, ACM, Springer websites. This made us a better understanding of detecting an spoofed web page and carious phishing attacks to develop a useful interface that is essential to check the spam URLs. In Referencing to IEEE research paper, we got to know about URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis, this is a high- level methodology done using gradient Boosting, Random Forest, Support Vector Machine. These experiments show high results and the effectiveness of the approach adopted in handling the imbalanced dataset presented.

Moreover, no over fitting for the data trained and tested was presented in all experiments conducted. In general, the results showed that SVM outperforms the other classifiers as its accuracy achieved 99.896%. In referencing to an ACM paper, we got know about Segmentation-based Phishing URL Detection, they proposed a novel URL-tokenizer to decrease unknown words in URLs by combining Bert tokenizer and Word Segment tokenizer using natural language process. These Information, Methodology, Process of Execution is very useful in completing our project to implement few steps and process based on above references and developed an efficient prototype that is used to detect a phishing URL.

Detecting malicious url techniques IEEE symposium Series on Computational Intelligence, 2016
All observations on different techniques are known. Therefore, the ranking of the methods can be accepted as a correct and significant ranking in terms of prediction accuracy. Methods adapted are Decision Tree, K- nearest neighbor, Bayesian networks, Random forest, Support vector machine, Multi- layer perceptron. Research gaps are The results of this paper suggest that the classification methods achieve competitive prediction accuracy rates for URL classificationwhen only the numerical features are used for training.

Malicious URL Detection based on Machine Learning International Journal of Advanced Computer Science and Applications, 2020
Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF). Key findings are the combination between easy-to- calculate attributes and big data processing technologies to ensure the balance of the two factors is the processing time and accuracy of the system. Research gaps are In this study, we do not use special attributes, nor do we seek to create huge datasets to improve the accuracy of the system as many other traditional publications.

Malicious URL Detection using Machine Learning: A Survey Journal of Cornell University, 2017
Methodology adapted are Batch Learning, Online Learning, Representative Learning. Key findings are In practice, these methods and features can complement each other to improve the performance of the machine learning models. Research gaps are In this survey, we categorized most, if not all, the existing contributions for malicious URL detection in literature and have many challenges.

Phishing Detection IEEE Communications Surveys & Tutorials, April 2013
Methodology adapted are Blacklists, Rule- based, heuristics Visual and similarity Machine Learning-based classifiers. Key findings are User education or training is an attempt to increase the technical awareness level of users to reduce their susceptibility to phishing attacks. Research gaps are analyzes the phishing detection techniques from the perspective of their computational cost and energy consumption.

Detection and Analysis of Drive-By- Download Attacks and Malicious Javascript Code Proceedings of the 19th International Conference on World Wide Web (WWW 2010) -New York, New York, USA
Methodology adapted is JavaScript code is also used to carry out attacks against the user's browser and its extensions. Key findings are the system is able to identify anomalous JavaScript code by emulating its behavior and comparing it to the established profiles. Research gaps are Unfortunately, the dynamic nature of the JavaScript language and its tight integration with the browser make it difficult to detect and block malicious JavaScript code.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic SocialEngineering Attacks by ACM computer surveys
Methodology adapted is platform or system applications. Key findings are this article presents a taxonomy of semantic attacks, as well as a survey of applicable defences. Research gaps are the threat landscape and the associated mitigation techniques are used.

SYMANTEC INTERNET SECURITY THREAT REPORT (ISTR) 2019
Methodology adapted is insights into global threat activities. Key findings are we share the latest insights into global threat activity, cyber criminal trends, and attacker motivations. Research gaps are It requires more knowledge and cost to perform the trends.

Mobile Phishing Attacks and Mitigation Techniques Journal of Information Security,2015
Methodology adapted is Tradition webattacks and Phishing. Key findings are This paper discusses various phishing attacks using mobile devices followed by some discussion on countermeasures . Research gaps are the web attacks and counter measures are difficult to find out and apply.

Identification of Malicious Webpages with static Heuristics by Victoria University of Wellington,2014
Methodology adapted are Http response with html and honeypot capture hpcv2.1. Key
findings are this paper presents a simple yet effective classification method for detecting malicious web pages that requires assessing attributes of the initial HTTP response. Research gaps are The acquired knowledge needs to be updated periodically to ensure that the majority of malicious pages are detected with the presented classification method.

III. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Adaptability of Phishing Techniques: Phishers continuously evolve their tactics to mimic legitimate URLs effectively. Existing systems struggle to keep pace with the dynamic nature of phishing attacks, making it challenging to reliably detect deceptive URLs.
Mimicry of Legitimate URLs: Some phishing sites employ sophisticated
techniques to closely mimic the appearance of legitimate URLs. Detection systems may struggle to differentiate between authentic and deceptive URLs, leading to potential false negatives.

3.1.1 Disadvantages

Limited Detection Accuracy: Existing systems may struggle with accurately identifying all malicious URLs due to the constantly evolving tactics used by attackers. False positives and false negatives can undermine the effectiveness of the system.
Over-reliance on Static Features: Some systems primarily rely on static features such as URL structure or domain reputation, which may not capture the dynamic behavior of malicious URLs. This can lead to a lack of robustness in detection, especially against sophisticated attacks that obfuscate their URLs.
Scalability Issues: As the volume of online data continues to grow exponentially, existing systems may face scalability challenges in processing and analyzing a large number of URLs in real-time. This can result in delays or inefficiencies in detection.
Limited Coverage of Attack Vectors: Some systems may focus on specific types of malicious URL behavior (e.g., phishing, malware distribution) while neglecting other attack vectors. This limited coverage leaves gaps in protection and allows attackers to exploit overlooked vulnerabilities.

3.2 PROPOSED SYSTEM

Real-Time Monitoring: Establish a robust real-time monitoring component that can quickly assess and categorize URLs, addressing scalability challenges and ensuring timely detection of phishing threats.
Contextual Analysis: Enhance the system with contextual analysis, considering the broader context in which a URL is embedded, to improve the accuracy of threat assessment and reduce false positives.
Continuous Learning and Updates: Ensure the system can continuously learn fromnew data and updates to stay ahead of evolving phishing tactics, providing ongoing protection against emerging threats.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Dynamic Feature Analysis: Implement a system that dynamically analyzes URL features, considering changes over time, to capture evolving tactics used by phishers and adapt to new threats effectively.

3.2.1 Advantages

- Multi-Context Analysis
- Real-Time Detection
- Reduced Dependency on Static Features
- Privacy Preservation
- Integration with Existing Security Infrastructure
- User-Friendly Interface

3.3 MODULES

3.3.1 Training

Data Splitting: Divide the data set into training and validation sets, allocating a significant portion (e.g., 70-80%) to training and the remainder to validation.

Model Training: Train the machine learning model using the training data set. Various algorithms such aslogistic regression, random forest, or neural networks can be employed for this purpose.

Hyper parameter Tuning: Fine-tune the model's hyper parameters to optimize its performance. Techniques like gridsearch or random search can be used to search for the best combination of hyper parameters.

Data Preprocessing: Handling missing values: Imputation techniques such as mean imputation or interpolation. Normalization/Standardization: Scaling features to ensure they have similar ranges. Feature selection: Identifying the most relevant features using techniques like feature importance or dimensionality reduction.

Evaluation Metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE): Common metrics for regression tasks.

R-squared (R^2): Measure of how well the model fits the data relative to a simple average.

Cross-validation: Assessing model performance on unseen data to avoid over fitting.

3.3.2 Validation

Temporal Validation: Temporal validation in URL spam detection involves splitting data based on time, training the model on earlier data, and evaluating its performance on later data, ensuring the model's effectiveness in detecting evolving spam patterns over time.

Spatial Validation: Spatial validation in URL spam detection entails splitting data geographically, training the model on one region's data, and evaluating its performance on another, ensuring the model's adaptability across diverse geographic contexts for effective spam detection.

Cross-Validation: Employ techniques like k-fold cross-validation to assess the model's robustness. This involves splitting the dataset into k subsets, training the model on k-1 subsets, and validating it on the remaining subset. Repeating this process k times ensures that each subset serves as the validation set exactly once.

Feature Importance: Feature importance in URL spam detection identifies which features (e.g., domain reputation, URL length) have the most significant impact on classification, enabling prioritization of key factors for accurate spam detection and model refinement, enhancing overall performance and inter pretability.

Outlier Detection: Outlier detection in URL spam detection identifies URLs that deviate significantly from normal behavior, potentially indicating malicious activity, enhancing the system's ability to detect novel spam tactics and mitigate emerging threats effectively, bolstering overall cybersecurity measures.

IV. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

A software architecture is a set of principles that define the way software is designed and developed.An architecture defines the structure of the software system and how it is organized. It also describes the relationships between components, levels of abstraction, and other aspects of the software system. An architecture can be used to define the goals of a project, or it can be used to guide the design and development of a new system. A software architecture is a set of principles that define the way software is designed and developed. An architecture defines the structure of the software system and how it is organized. It also describes the relationships between components, levels of abstraction, and other aspects of the software system. An architecture can be used to define the goals of a project, or it can be used to guide the design and development of a new system. A software architecture is a set of principles that define the way

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

software is designed and developed. An architecture defines the structure of the software system and how it is organized. It also describes the relationships between components, levels of abstraction, and other aspects of the software system. An architecture can be used to define the goals of a project, or it can be used to guide the design and development of a new system.

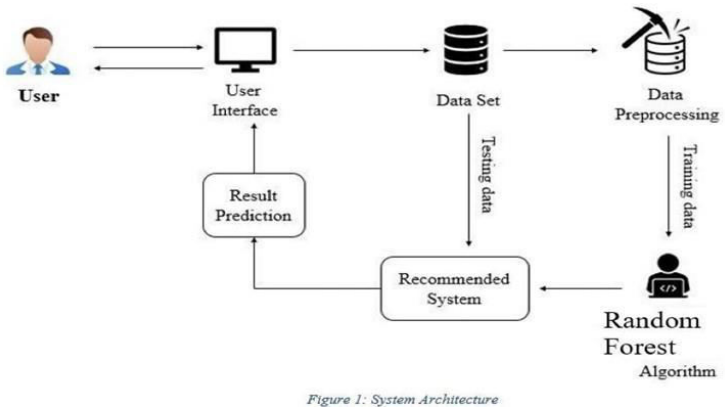


Fig 3: System Architecture

Data Collection:
Gather a diverse set of data samples comprising both phishing and legitimate websites. Sources may include publicly available phishing data sets, online repositories, and data scraping techniques.
Ensure the data collected covers a wide range of phishing tactics and techniques to train the classification models effectively.

Data Preprocessing:
Clean the collected data to remove duplicates, irrelevant information, and noise.
Perform data transformation, normalization, and encoding to prepare the data for analysis.
Split the data into training and testing sets for model evaluation.

Feature Engineering:
Identify relevant features that can distinguish between phishing and legitimate websites. Extract features such as URL structure, domain characteristics, presence of HTTPs, website content analysis, and other metadata.
Utilize domain-specific knowledge and expertise to select informative features for model training.

Model Deployment:
Deploy the finalized system in production environments, either as standalone software or integrated into existing cyber security frameworks.
Provide documentation, training, and support to users for effectively utilizing the system. Implement monitoring and maintenance protocols to ensure the system's continued effectiveness and reliability over time.

4.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general- purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to or associated with, UML. The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

UML was created as a result of the chaos revolving around software development and documentation. In the 1990s, there were several different ways to represent and document software systems.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

4.2a GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of object oriented tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

V. RESULTS

5.1 Algorithms used

We have already applied, studied, and experimented with the decision tree approach. It won't be too hard to understand the random forest method, then. It is an extremely well- liked ensemble learning algorithm. It is a composite of many randomly selected trees from various data set subsets. A random forest network with an abundance of random trees has a high model accuracy. The tree with the most votes is chosen to make the decision.

The random forest method has the best accuracy, at 93.73% percent. Finally, we use this approach to the creation of our model and the launch of our project, which is a website. The random forest model is depicted in "Fig. random-forest" below. Generally speaking, a forest is made up of several types of trees. In this case as well, we can state that random forests are composed of different decision tree subsets.

5.1.1 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to
the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given data set and takes the average to improve the predictive accuracy of that data set." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The below diagram explains the working of the Random Forest algorithm:

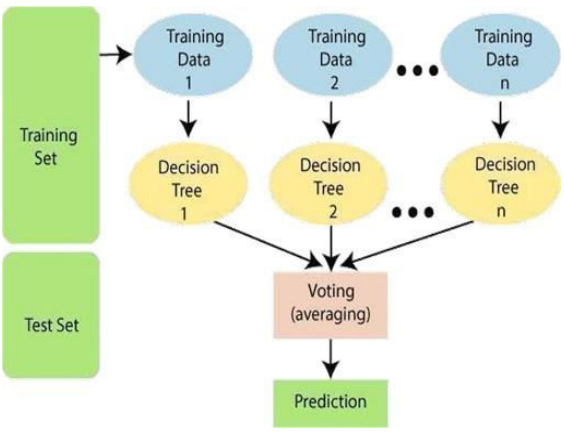


Fig 4: Random forest method

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

5.2.2 Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the data set, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier: There should be some actual values in the feature variable of the data set so that the classifier can predict accurate results rather than a guessed result. The predictions from each tree must have very low correlations.

5.2.3 Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large data set it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

5.2.4 How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Example: Suppose there is a data set that contains multiple fruit images. So, this data set given to the Random forest classifier. The data set is divided into subsets and given to each decision tree.

During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:

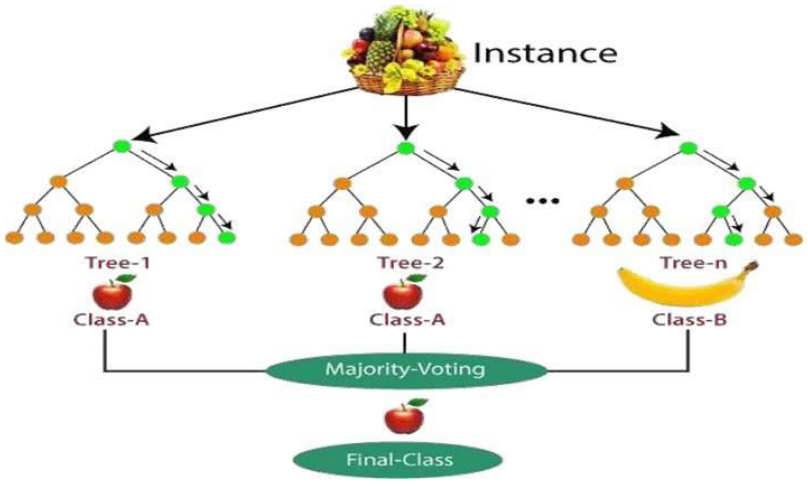


Fig 5: Example of Random Forest

5.2.5 Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

Banking: Banking sector mostly uses this algorithm for the identification of loan risk.

Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.

Land Use: We can identify the areas of similar land use by this algorithm.

Marketing: Marketing trends can be identified using this algorithm.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

5.2.6 Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over fitting issue.

5.2.7 Disadvantages of Random Forest

Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

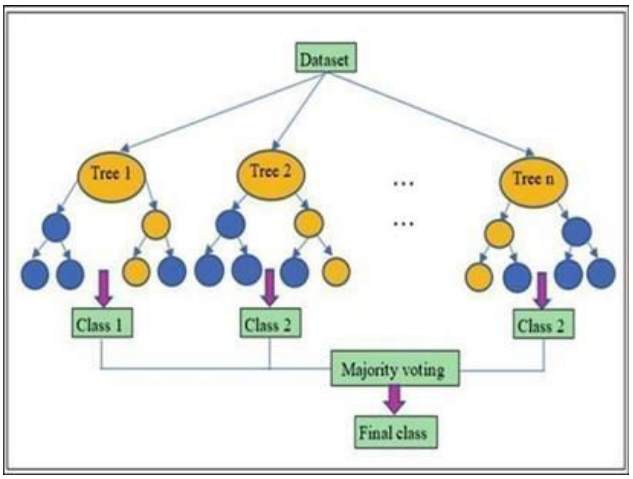


Fig 6: Random-forest

5.2.8 Implementation Steps:

1. Start with home page
2. And then design predict
3. Collect datasets
4. Data Preprocessing
5. Create a ML model
6. Train and Test the Model
7. Predict the Output
8. Display Output on Result page
9. Finally Create Dashboard Page

The following figures present the sequence of screenshots of the results.

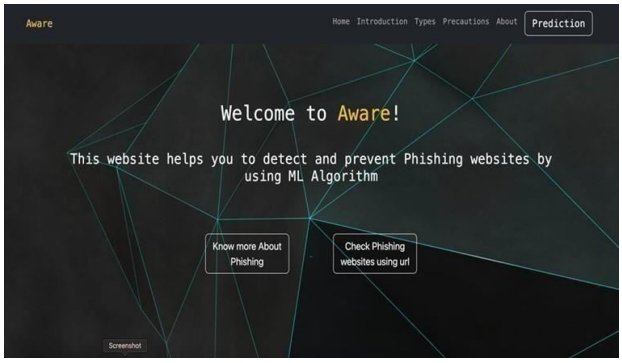


Fig 7a: Introduction to webpage

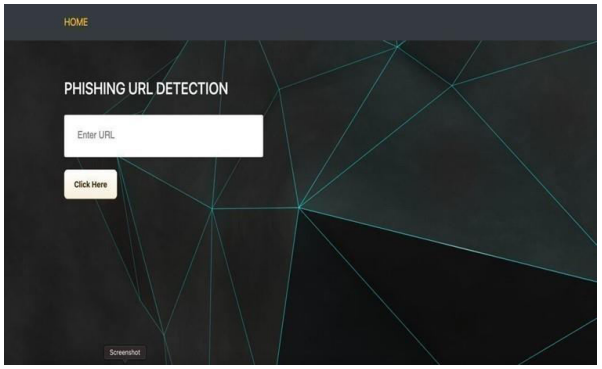


Fig 7b: Phishing URL detection page

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304166

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

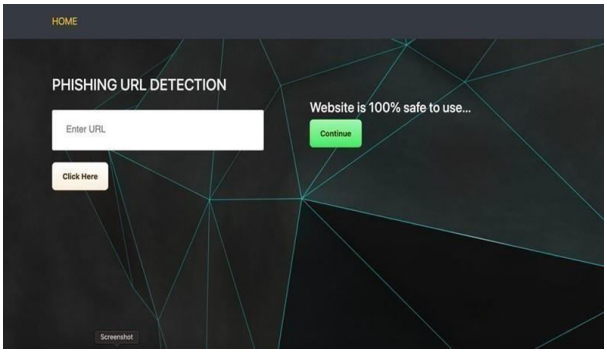


Fig 7c: legitimate page

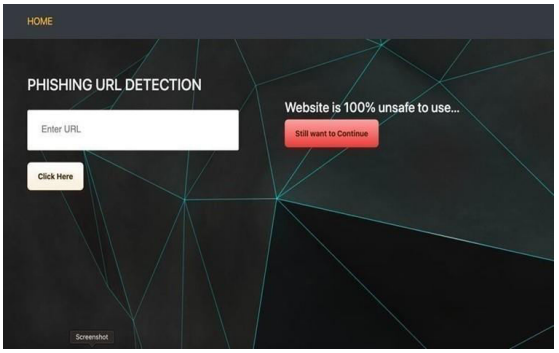


Fig 7d: Phishing page

VI. CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

In conclusion, the detection of phishing websites through URL features is a critical and evolving aspect of cyber security. While current systems face challenges such as adaptability issues, false positives, and static analysis limitations, proposed solutions involving dynamic feature analysis, behavioral assessment, advanced machine learning models, and real-time monitoring hold promise for a more robust defense. The continuous evolution of phishing tactics demands a proactive and adaptive approach, integrating innovative technologies to counteract new threats effectively. As online threats persist and grow in sophistication, enhancing phishing detection mechanisms remains pivotal in safeguarding user trust, financial security, and the integrity of digital ecosystems on a global scale. The landscape of phishing website detection using URL features is dynamic, demanding constant innovation to counter evolving cyber threats.

6.2 FUTURE WORK

Looking ahead, we can make phishing website detection even better by using more advanced technologies. Imagine using super-smart computer programs that understand tricky patterns in fake website addresses, making them better at spotting scams. We could also make these programs clearer, so people can understand why they make certain decisions, making the whole process more trustworthy. Another idea is to use a special kind of technology called block chain, making it harder for bad actors to mess with the information about dangerous websites. By paying attention to how people behave online and teaming up with other security groups worldwide, we can create a strong defense against new types of scams. This might also involve having experts work alongside the smart programs to make sure everything stays secure. Overall, the goal is to keep improving and working together globally to stay ahead of the bad guys trying to trick us online.

REFERENCES

[1] T. Manyumwa, P. F. Chapita, H. Wu, and S. Ji, ‘‘Towards fighting cybercrime: Malicious URL attack type detection using multiclass classification,’’ in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 1813–1822.

[2] M. Alshehri, A. Abugabah, A. Algarni, and S. Almotairi, ‘‘Characterlevel word encoding deep learning model for combating cyber threats in phishing URL detection,’’ *Comput. Electr. Eng.*, vol. 100, May 2022, Art. no. 107868.

[3] (2022). *Making the World’s Information Safely Accessible.* [Online]. Available: https://safebrowsing.google.com/

[4] D. R. Patil and J. B. Patil, ‘‘Feature-based malicious URL and attack type detection usingmulticlass classification,’’*Int. J.Inf. Secur.*, vol.10, no. 2, pp. 141–162, 2018.

[5] F. O. Catak, K. Sahinbas, and V. Dörtkardeş, ‘‘Malicious URL detection using machine learning,’’ in *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*. Hershey, PA, USA: IGI Global, 2021, pp. 160–180.

[6] E. Benavides, W. Fuertes, S. Sanchez, and M. Sanchez, ‘‘Classification of phishing attack solutions by employing deep learning techniques: A systematic literature review,’’ in *Developments and Advances in Defense and Security* (Smart Innovation, Systems and Technologies), vol. 152, Á. Rocha and R. Pereira, Eds. Singapore: Springer, 2020, doi: 10.1007/978981-13-9155-2_5.

[7] K. Krombholz, P. Frühwirt, P. Kieseberg, I. Kapsalis, M. Huber, and E. Weippl, ''QR code security: A survey of attacks and challenges for usable security,'' in *Proc. Int. Conf. Hum. Aspects Inf. Secur., Privacy, Trust*. Cham, Switzerland: Springer, 2014, pp. 79–90.

[8] C. D. Xuan, H. Dinh, and T. Victor, ''Malicious URL detection based on machine learning,'' *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 148–153, 2020.

[9] Y. Liang, Q. Wang, K. Xiong, X. Zheng, Z. Yu, and D. Zeng, ''Robust detection of malicious URLs with self-paced wide & deep learning,'' *IEEE Trans. Depend. Secure Comput.*, vol. 19, no. 2, pp. 717–730, Mar. 2022.

[10] F. Sadique, R. Kaul, S. Badsha, and S. Sengupta, ''An automated framework for real-time phishing URL detection,'' in *Proc. 10th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2020, pp. 335–341.

[11] D. Huang, K. Xu, and J. Pei, ''Malicious URL detection by dynamically mining patterns without pre-defined elements,'' *World Wide Web*, vol. 17, no. 6, pp. 1375–1394, Nov. 2014.

[12] M. Alsaedi, F. Ghaleb, F. Saeed, J. Ahmad, and M. Alasli, ''Cyber threat intelligence-based malicious URL detection model using ensemble learning,'' *Sensors*, vol. 22, no. 9, p. 3373, Apr. 2022.

[13] M. Aljabri, H. S. Altamimi, S. A. Albelali, M. Al-Harbi, H. T. Alhuraib, N. K. Alotaibi, A. A. Alahmadi, F. Alhaidari, R. M. A. Mohammad, and K. Salah, ''Detecting malicious URLs using machine learning techniques: Review and research directions,'' *IEEE Access*, vol. 10, pp. 121395– 121417, 2022.

[14] A. S. Rafsanjani, N. B. Kamaruddin, H. M. Rusli, and M. Dabbagh, ''QsecR: Secure QR code scanner according to a novel malicious URL detection framework,'' *IEEE Access*, vol. 11, pp. 92523–92539, 2023.

[15] A. S. Rafsanjani, N. Kamaruddin, N. N. A. Sjariff, N. Firdaus, N. Maarop, and H. M. Rusli, ''A evaluating security and privacy features of quick response code scanners: A comparative study,'' *Open Int. J. Informat.*, vol. 10, no. 2, pp. 197–207, 2022.

[16] J. Yuan, Y. Liu, and L. Yu, ''A novel approach for malicious URL detection based on the joint model,'' *Secur. Commun. Netw.*, vol. 2021, pp. 1–12, Dec. 2021.

[17] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, ''URLNet: Learning a URL representation with deep learning for malicious URL detection,'' 2018, *arXiv:1802.03162*.

[18] M. Akiyama, T. Yagi, and M. Itoh, ''Searching structural neighborhood of malicious URLs to improve blacklisting,'' in *Proc. IEEE/IPSJ Int. Symp. Appl. Internet*, Jul. 2011, pp. 1–10.

[19] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, ''PhishNet: Predictive blacklisting to detect phishing attacks,'' in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–5.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details