RAMAIAH INSTITUTE OF TECHNOLOGY, BANGALORE – 560054
(Autonomous Institute, Affiliated to VTU)

Department of Computer Science & Engineering

CS44: Data Communication and Networking

Report On

Phishing URL Detection Using Random Forests Algorithm

Shamanth Hiremath	1MS22CS128
Sanchit Vijay	1MS22CS122
Trijal Shinde	1MS22CS153

Under the Guidance

Manjula L
Assistant Professor

Ramaiah Institute of Technology
(Autonomous Institute, Affiliated to VTU)
MSR Nagar, MSRIT Post, Bangalore-560054

April 2024-July 2024

RAMAIAH INSTITUTE OF TECHNOLOGY, BANGALORE – 560054

(Autonomous Institute, Affiliated to VTU)

Department of Computer Science & Engineering

<u>Evaluation Report</u>

Team Member Details		
Sl. No.	USN	Name
1.	Shamanth Hiremath	1MS22CS128
2.	Sanchit Vijay	1MS22CS122
3.	Trijal Shinde	1MS22CS153

SL No.	Component	Maximum Marks	Marks Obtained
1	Simulation of attack - Demo	10	
2	Report	10	
Total Marks		20	

Signature of the Student **Signature of the Faculty**

TABLE OF CONTENTS

1. Introduction

Phishing attacks are a prevalent cybersecurity threat that exploits human psychology and trust in digital communications. Attackers masquerade as legitimate entities, such as banks or reputable organizations, to deceive users into disclosing sensitive information like passwords, credit card numbers, or personal data. These attacks often rely on sophisticated social engineering techniques, using emails, websites, or messages that appear authentic to trick individuals into taking actions that benefit the attackers.

The consequences of falling victim to phishing can be severe, leading to financial loss, identity theft, or unauthorized access to sensitive accounts. Businesses and individuals alike face risks ranging from compromised data integrity to reputational damage. Phishing techniques continue to evolve, adapting to technological advancements and user behavior, making detection and prevention increasingly challenging.

Detecting phishing URLs is crucial for mitigating these threats. This project employs a RandomForest classifier, a robust machine learning algorithm capable of processing a wide range of URL features. These features include URL structure, domain characteristics, SSL validity, and behavioral indicators like pop-ups or abnormal redirects. RandomForest's ensemble of decision trees aggregates predictions to classify URLs as either phishing or legitimate, providing a high level of accuracy and robustness against evolving attack methods.

By combining thorough feature extraction with advanced machine learning techniques, this project aims to enhance online security by effectively identifying and mitigating phishing threats. Future developments include real-time URL checking APIs, improved feature extraction methods, and browser extensions for seamless user protection against phishing attacks.

2. Literature Survey

Phishing is a significant threat in the digital age, where attackers create fraudulent websites to deceive users and steal sensitive information. The use of machine learning techniques to detect phishing websites has garnered significant attention. This literature survey explores the methodologies and findings of various studies focused on phishing detection using machine learning and feature selection methods

Detection Techniques:

- **Domain-based Features**: Shirazi explores unbiased phishing detection using domain characteristics [3].
- **Hybrid SVM and KNN**: Altaher combines SVM and KNN for classification [4].
- **Content Consistency**: Chen et al. check content alignment for phishing detection [5].
- **Associative Classification**: Abdelhamid et al. use data mining for rule discovery [6].
- **Rule-Based Methods**: Moghimi and Varjani propose rule-based detection techniques [7].
- **Fuzzy Data Mining**: Aburrous et al. employ fuzzy logic for uncertain data [8].
- **Heuristic Approaches**: Solanki et al. and Lee et al. use heuristic indicators for detection [10], [11].
- **Logo Utilization**: Chiew et al. leverage website logos for authenticity checks [9].

Machine Learning Techniques:

- **URL Analysis**: Basnet and Doleck use ML to detect phishing URLs [12].
- **Efficiency**: Gu et al. develop efficient phishing detection methods [13].
- **Ensemble Feature Selection**: Chiew et al. propose hybrid ensemble frameworks [14].
- **Feature Selection**: Zabihimayvan and Doran employ fuzzy rough set theory for feature selection [15].

Experimental Findings:

- Studies evaluate ML algorithms like J48, Random Forest, and MLP using feature selection methods like InfoGain and ReliefF [17].
- Performance metrics emphasize accuracy improvements and computational efficiency [18].
- Optimal feature sets reduce from 48 to 20, enhancing model efficiency without compromising accuracy [22].

3. Methodology

Feature Extraction Process:

When a URL is input into the system, it undergoes comprehensive feature extraction to analyze various characteristics. Some of these features include:

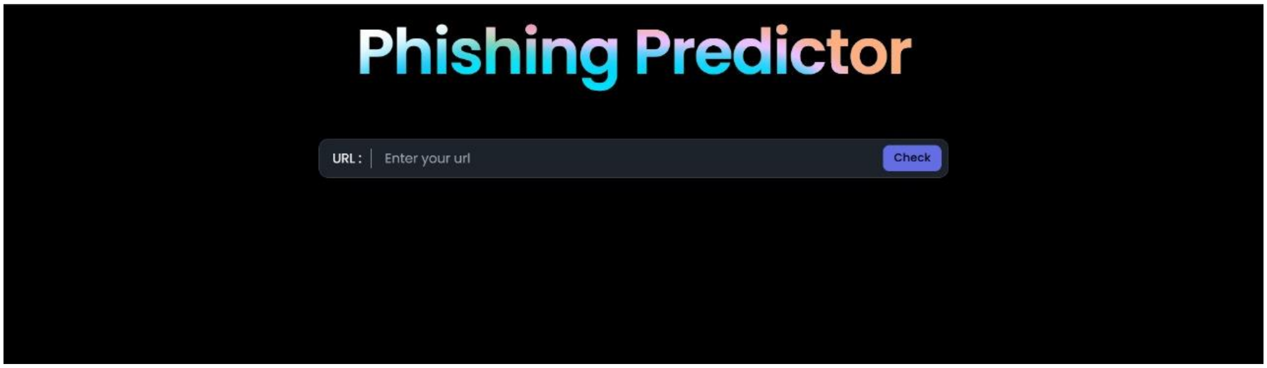
- **URL Type:** Checks if the URL contains an IP address, its length, and whether it's shortened.
- **Special Characters:** Identifies the presence of "@" symbol and double slashes in the URL path.
- **Domain Analysis:** Examines prefix/suffix domains, subdomains, SSL certificate validity, and domain registration length.
- **Web Page Attributes:** Determines the presence of a favicon, unusual port numbers, and "https" token.
- **Web Page Elements:** Checks for external URL requests, anchor elements, and the number of links within HTML tags.
- **Security Indicators:** Evaluates suspicious form handling, email submission requirements, abnormal URL redirection, and features like mouse-over and right-click protection, pop-up windows, and iframes.
- **Domain Metrics:** Includes domain age, DNS records, estimated web traffic, Google PageRank, indexing status, external links, and various statistical feature.

Prediction Using RandomForest:

The RandomForest classifier used in this project is trained on a feature vector comprising 32 extracted features from each URL input. RandomForest operates by constructing multiple decision trees during the training phase, where each tree is trained on a random subset of the training data and a random subset of the features. This ensemble approach allows each tree to independently classify URLs based on their feature vectors. During the classification phase, the final prediction is determined through a majority voting mechanism across all decision trees. This approach not only enhances the robustness of the model against overfitting, a common challenge in phishing detection due to the dynamic nature of attacks, but also provides insights into the importance of different URL characteristics in identifying phishing attempts. By combining thorough feature extraction with the ensemble learning capabilities of RandomForest, this project aims to deliver accurate and reliable phishing URL detection, contributing significantly to online security measures for users.

Graphical User Interface (GUI) for URL Classification:

As part of this project, a user-friendly GUI has been developed to enhance usability and accessibility. The GUI empowers users by providing an intuitive platform where they can input any URL of interest. Upon submission, the system promptly processes the URL through the feature extraction pipeline and applies the RandomForest model for classification. Users receive immediate feedback indicating whether the URL is categorized as phishing or legitimate. This interactive interface not only simplifies the process of URL verification but also ensures users can make informed decisions about online safety with ease and confidence.



4. Results and Discussion

Our study focused on developing and evaluating a RandomForest classifier for detecting phishing URLs. Through rigorous experimentation and validation, we achieved a significant accuracy of 97.31% in distinguishing between legitimate and phishing websites. This high accuracy demonstrates the robustness and reliability of our approach in effectively identifying malicious URLs.

The RandomForest model was trained on a comprehensive feature set extracted from URLs, including characteristics such as domain age, presence of HTTPS, URL length, and various security indicators. The ensemble nature of RandomForest, utilizing multiple decision trees trained on different subsets of features, contributed to its ability to handle the complex feature space of URL attributes.

In practical terms, achieving an accuracy of 97.31% means that our classifier correctly identified phishing URLs in the vast majority of cases, thereby enhancing online security for users. This result is particularly significant given the evolving nature of phishing attacks, where malicious actors continuously adapt their strategies to evade detection.

However, it's worth noting that while our approach is robust, feature extraction for URLs can be time-consuming. In some cases, the process may fail to generate all desired features, impacting the classifier's performance. This challenge highlights opportunities for future optimization in feature extraction techniques to streamline and enhance the efficiency of our phishing detection system.

Furthermore, our approach not only prioritizes accuracy but also considers interpretability and scalability. By providing insights into feature importance and classification decisions, our model empowers users and security professionals to better understand and mitigate online threats.

The success of our RandomForest classifier underscores its effectiveness as a robust solution for phishing URL detection. Moving forward, future research could explore enhancements such as integrating real-time data feeds and expanding the feature set to further improve detection capabilities in dynamic online environment

```
RANDOM FOREST MODEL SPECS:
Accuracy of the model: 97.31789746464618 %
Best Accuracy = 0.9731789746464618
Best parameters = {'criterion': 'entropy', 'max_features': 'log2',
'n_estimators': 100}

Cross-validation score: 0.9693641373263504
Confusion Matrix:
[       True  False
    +ve [1186   63]
    -ve [26   1489]
]
```

5. Conclusion

Phishing attacks represent persistent threats in the digital landscape, exploiting human trust and technological vulnerabilities to compromise sensitive information. This report has explored various forms of phishing, including email, spear phishing, smishing, and vishing, highlighting their sophisticated tactics and detrimental impacts on individuals and organizations.

Our study focused on developing and evaluating a RandomForest classifier for detecting phishing URLs, achieving a commendable accuracy of 97.31%. This approach leveraged advanced machine learning techniques to analyze comprehensive URL features, including domain age, HTTPS presence, and URL structure. The robust performance of our classifier underscores its efficacy in distinguishing between legitimate and malicious URLs, thereby enhancing online security measures.

While technological solutions like machine learning models are pivotal in phishing detection, effective defense strategies must also integrate user education and awareness initiatives. Educating users to recognize phishing indicators and adopt safe online practices is crucial in fortifying defenses against evolving cyber threats.

Looking ahead, continuous research and collaboration across sectors are essential to refine detection mechanisms and adapt to emerging phishing tactics. By prioritizing proactive cybersecurity measures and fostering a culture of vigilance, stakeholders can collectively mitigate the impact of phishing attacks and bolster trust in digital interactions.

6. References:

[1]. Anti-Phishing Working Group. Phishing Activity Trends Report 4th Quarter. 2018. [Online]. Available: https://docs.apwg.org

[2]. Microsoft Security Intelligence Report. Volume 24. 2019. [Online]. Available: https://www.microsoft.com/security

[3]. Hossein Shirazi. Unbiased phishing detection using domain name based features. PhD thesis, Colorado State University, Libraries.

[4]. Altyeb Altaher. Phishing websites classification using hybrid SVM and KNN approach. *International Journal of Advanced Computer Science and Applications*, 8(6):90–95, 2017. https://doi.org/10.14569/ijacsa.2017.080611

[5]. Yi-Shin Chen, Yi-Hsuan Yu, Huei-Sin Liu, and Pang-Chieh Wang. Detect phishing by checking content consistency. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, pages 109–119. IEEE, 2014. https://doi.org/10.1109/iri.2014.7051880

[6]. Neda Abdelhamid, Aladdin Ayesh, and Fadi Thabtah. Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41(13):5948–5959, 2014. https://doi.org/10.1016/j.eswa.2014.03.019

[7]. Mahmood Moghimi and Ali Yazdian Varjani. New rule-based phishing detection method. *Expert Systems with Applications*, 53:231–242, 2016. https://doi.org/10.1016/j.eswa.2016.01.028

[8]. Maher Aburrous, M Alamgir Hossain, Keshav Dahal, and Fadi Thabtah. Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert Systems with Applications*, 37(12):7913–7921, 2010. https://doi.org/10.1016/j.eswa.2010.04.044

[9]. Kang Leng Chiew, Ee Hung Chang, Wei King Tiong, et al. Utilisation of website logo for phishing detection. *Computers & Security*, 54:16–26, 2015. https://doi.org/10.1016/j.cose.2015.07.006

[10]. Jaydeep Solanki and Rupesh G Vaishnav. Website phishing detection using heuristic-based approach. In *Proceedings of the third international conference on advances in computing, electronics and electrical technology*, 2015.

[11]. Jin-Lee Lee, Dong-Hyun Kim, and Lee Chang-Hoon. Heuristic-based approach for phishing site detection using URL features. In *Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology-CEET*, pages 131–135, 2015. https://doi.org/10.15224/978-1-63248-056-9-84

[12]. Ram B Basnet and Tenzin Doleck. Towards developing a tool to detect phishing URLs: a machine learning approach. In *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, pages 220–223. IEEE, 2015. https://doi.org/10.1109/cict.2015.63

[13]. Xiaoqing Gu, Hongyuan Wang, and Tongguang Ni. An efficient approach to detecting phishing web. *Journal of Computational Information Systems*, 9(14):5553–5560, 2013.

[14]. Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin SC Yong, and Wei King Tiong. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, 484:153–166, 2019. https://doi.org/10.1016/j.ins.2019.01.064

[15]. Mahdieh Zabihimayvan and Derek Doran. Fuzzy rough set feature selection to enhance phishing attack detection. *arXiv preprint arXiv:1903.05675*, 2019. https://doi.org/10.1109/fuzz-ieee.2019.8858884

[16]. Adwan Yasin and Abdelmunem Abuhasan. An intelligent classification model for phishing email detection. *arXiv preprint arXiv:1608.02196*, 2016.

[17]. Mohammad Almseidin, Maen Alzubi, Szilveszter Kovacs, and Mouhammd Alkasassbeh. Evaluation of machine learning algorithms for intrusion detection system. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000277–000282. IEEE, 2017. https://doi.org/10.1109/sisy.2017.8080566

[18]. Mouhammad Alkasassbeh and Mohammad Almseidin. Machine learning methods for network intrusion detection. *Icccnt 2018 - The 20TH International Conference On Computing, Communication And Networking Technologies*, 2018.

[19]. Ibrahim Obeidat, Nabhan Hamadneh, Mouhammd Alkasassbeh, Mohammad Almseidin, and Mazen AlZubi. Intensive pre-processing of kdd cup 99 for network intrusion classification using machine learning techniques. 2019. https://doi.org/10.3991/ijim.v13i01.9679

[20]. Mouhammd Alkasassbeh, Ghazi Al-Naymat, AB Hassanat, and Mohammad Almseidin. Detecting distributed denial of service attacks using data mining techniques. *International Journal of Advanced Computer Science and Applications*, 7(1):436–445, 2016. https://doi.org/10.14569/ijacsa.2016.070159

[21]. Mouhammad Alkasassbeh. An empirical evaluation for the intrusion detection features based on machine learning and feature selection methods. *Journal of Theoretical and Applied Information Technology*, 95(22), 2017.

[22]. Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: introduction and review. *Journal of Biomedical Informatics*, 2018. https://doi.org/10.1016/j.jbi.2018.07.014

[23]. Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2018.