TruthGuard: Verifying Website Authenticity

Anugunj Barange[1], Tanishka Pandagre[2], Parul Pawar[3], Pooja Makode[4], Khushi Waghmode[5], Prof. Nilesh Mishra[6]

[1, 2, 3, 4, 5]B.Tech. Scholars, [6]Assistant Professor
Department of Artificial Intelligence and Data Science
Shri Balaji Institute of Technology and Management
Betul, RGPV University (M.P.), India

Abstract

With the Internet and improvement of intensive merging of social life, the Internet looks different that people are learning and jobs, in the meantime, we need to open to increase serious security attacks. Ways to identify different network threats, especially not first seen attacks, is a primary question that needs to be seen immediately. URL for phishing spots aims to collect personal information such as user identification, password and online money -related exchanges. The web uses websites that are visually and semantic as authentic websites. Since most of the customers go online to get the administration provided by the authorities and money -related organizations, there has been a significant increase in phishing hazards and attacks over the years. As technology increases, the phish methods have begun to make severe progress, and it should be avoided by using anti-phishing techniques to detect phishing. Machine learning is an official unit that can be used to target the phish attack. This study develops and creates a model that can guess if a URL connection is valid or phishing

People for cyber security are now looking for reliable and stable identity techniques to detect phishing sites. By removing and evaluating many aspects of authentic and phishing URL, the project uses machine learning techniques to detect phishing URL.

Finally, the study provided a model for phishing and URL classification in a valid URL. It will be very valuable to identify the phish attacks by certifying any link that is delivered to them to prove its validity.

Keywords: Security Attacks, Personal Information, Phishing Hazards, Anti-Phishing Techniques, Machine Learning, Cyber Security, Valid URL

1. 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. 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 blacklist method for phishing website detection blocks known malicious URLs but has drawbacks like limited coverage of new sites, zero-hour phishing attack, resource intensiveness, false positives, and scalability issues. It is effective when combined with heuristic analysis and machine learning for robust protection. Heuristic methods for phishing website detection have several drawbacks: they can produce false positives and false negatives, they require constant updates and are complex to implement that can cause performance overhead. Their effectiveness depends heavily on the quality of the rules and algorithms used.

To overcome the above drawbacks, many security researchers are 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.

2. LITERATURE SURVEY

Reference	**Paper Title**	**Journal Name and Year of Publication**	**Methodology**	**Techniques Used**	**Key Findings**	**Limitations**
Sakhare et al. (2024)	*Phishing Website Detection Using Advanced Machine Learning Techniques*	*International Journal of Intelligent Systems and Applications in Engineering (IJISAE),* 2024	Machine Learning-based Phishing Website detection	Advanced ML models (Random Forest, Decision Tree, SVM, Neural Networks)	The study demonstrates that **ML models effectively classify phishing websites** with high accuracy.	Lacks real-time implementation and comparison with recent AI models.
Ramesh et al. (2023)	*Phishing Detection System using Random Forest Algorithm*	*International Journal for Research Trends and Innovation,* 2023	Phishing Detection Using Random Forest Algorithm	Feature-based classification using **Random Forest**	Achieves **high accuracy in phishing detection** by analyzing website characteristics.	Does not explore **deep learning approaches** or adversarial attacks.
Mahajan & Siddavatam (2018)	*Phishing Website Detection using Machine Learning Algorithms*	*International Journal of Computer Applications (IJCA),* 2018	Machine Learning Algorithms for Phishing Detection	Decision Tree, SVM, Logistic Regression	Proposes an **ML-based approach** to differentiate phishing from legitimate websites.	Uses **older ML models**; lacks comparison with deep learning-based techniques.

3. SYSTEM ARCHITECTURE

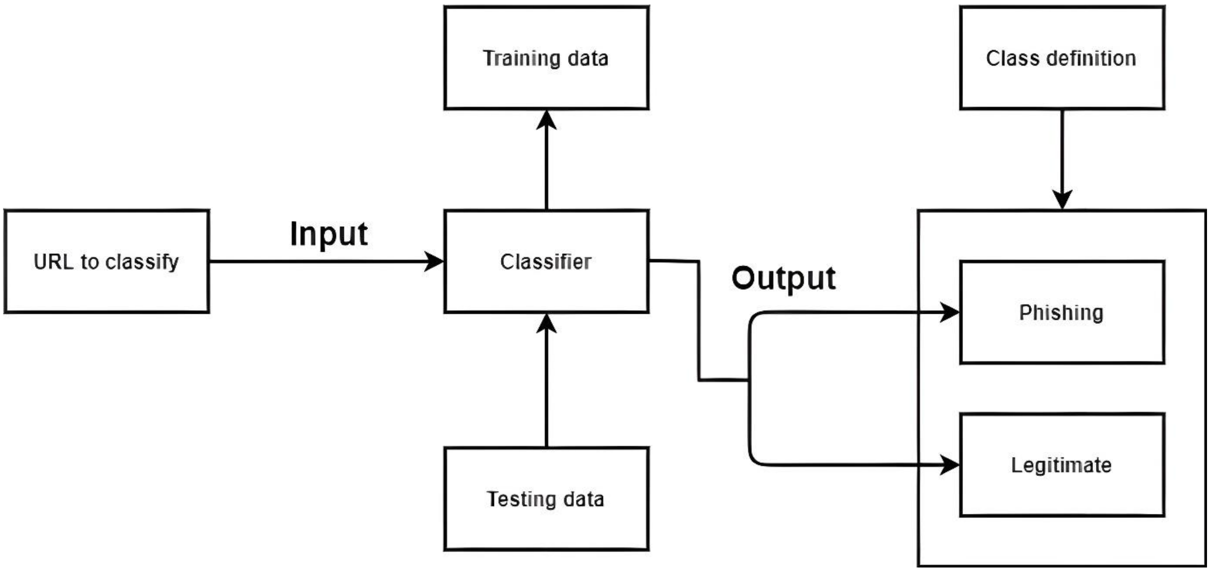


Fig: System Architecture

4. METHODOLOGY

The methodology used to achieve the earlier stated objectives is explained below. The dataset collection consists of phishing and legitimate URLs which were obtained from open-source platforms. The dataset was then pre-processed that is cleaned up from any abnormality such as missing data to avoid data imbalance. Afterward, expository data analysis was done on the dataset to explore and summarize the dataset. Once the dataset was free from all anomalies, website content-based features were extracted from the dataset to get accurate features to train and test the model. An extensive review was done on existing works of literature and machine learning models on detecting phishing websites to best decide the classification models to solve the problem of detecting phishing websites.

Hence, Series of these machine learning classification models such as Decision Tree, Support Vector Machine, Multilayer perceptions, Auto encoder Neural Network and Random Forest was deployed on the dataset to distinguish between phishing and legitimate URLs. The best model with high training accuracy out of all the deployed models was selected then integrated into a developed web application. Thus, a user can enter a URL link on the web application to predict if it is phishing or legitimate.

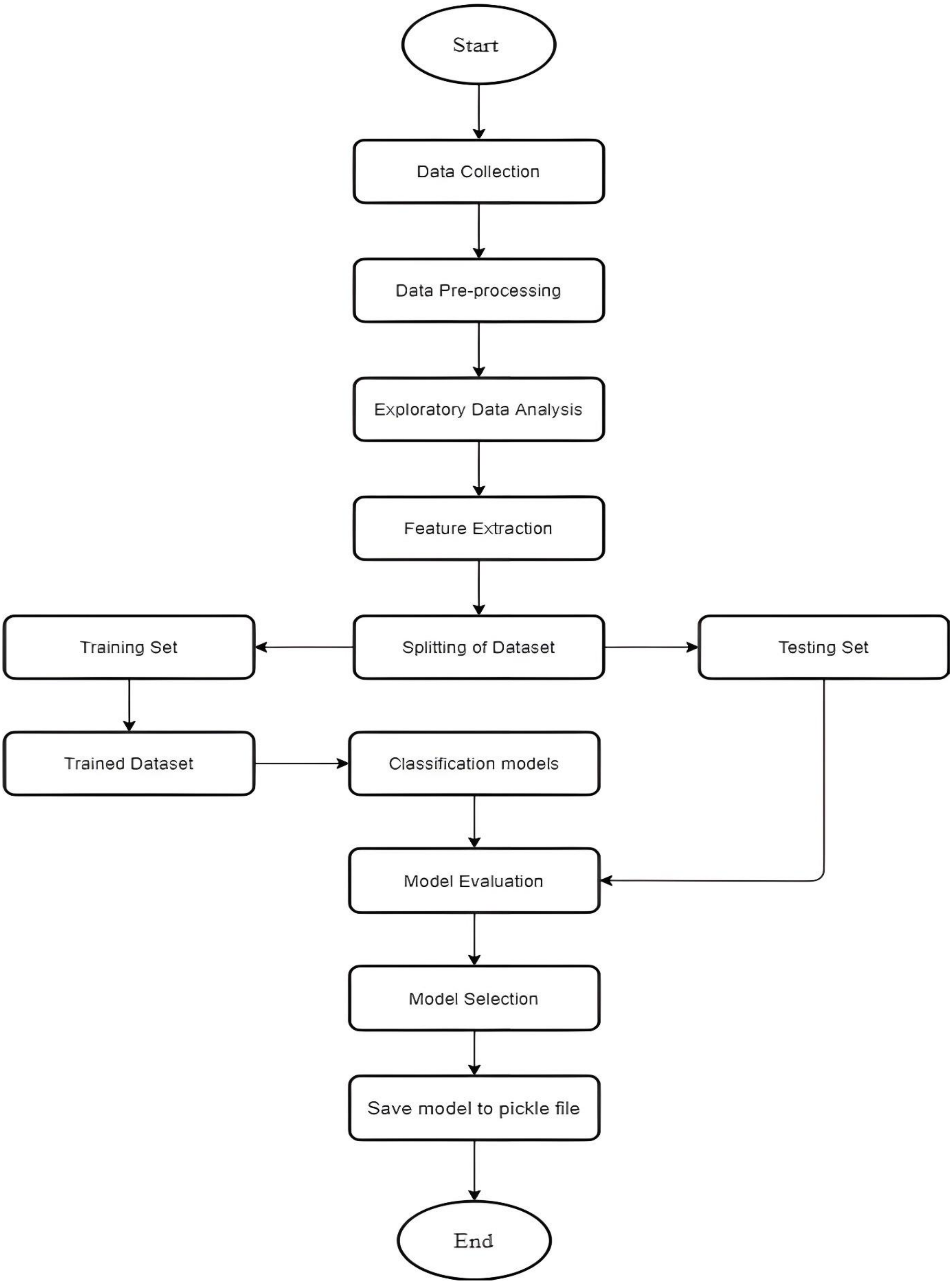
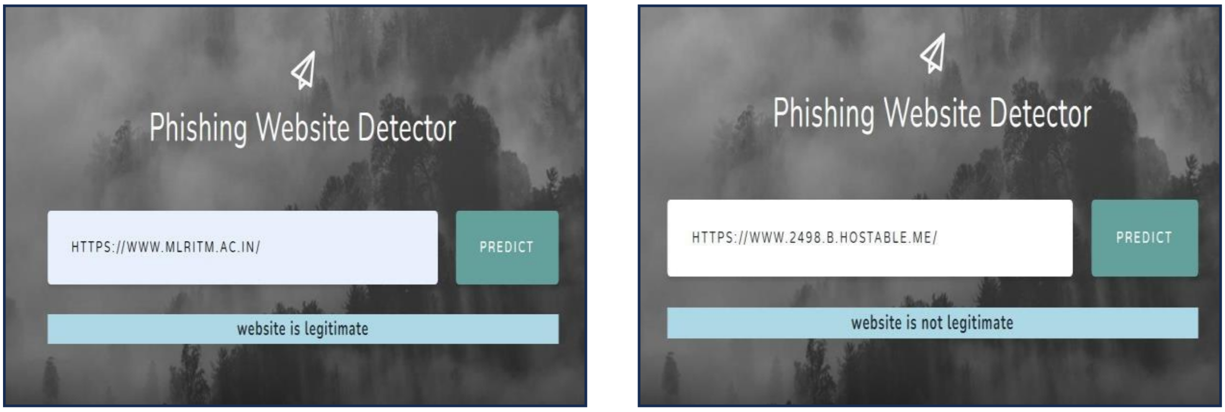


Fig: Flowchart of the proposed system

5. RESULT



HTTPS://WWW.MLRITM.AC.IN/

website is legitimate

HTTPS://WWW.2498.B.HOSTABLE.ME/

website is not legitimate

Fig: Interface

6. CONCLUSION

The demonstration of phishing is turning into an advanced danger to this quickly developing universe of innovation. Today, every nation is focusing on cashless exchanges, business online, tickets that are paperless and so on to update with the growing world. Yet phishing is turning into an impediment to this advancement. Individuals are not feeling web is dependable now. It is conceivable to utilize AI to get information and assemble extraordinary information items. A lay person, completely unconscious of how to recognize a security danger shall never invite the danger of making money related exchanges on the web. Phishers are focusing on installment industry and cloud benefits the most.

The project means to investigate this region by indicating an utilization instance of recognizing phishing sites utilizing ML. It aimed to build a phishing detection mechanism using machine learning tools and techniques which is efficient, accurate and cost effective. The project was written in Python. The proposed method used machine learning classifiers to achieve this and a comparative study of the algorithms was made. A good accuracy score was also achieved.

REFERENCES

1. Dr. Nitin N. Sakhare, Jyoti L. Bangare, Dr. Radhika G. Purandare, Disha S. Wankhede, Pooja Dehankar, "Phishing Website Detection Using Advanced Machine Learning Techniques", International Journal of Intelligent Systems and Applications in Engineering (IJISAE), 12(12s), 329–346, 2024.
2. Ahmed Abdeen Hamed,Malgorzata Zachara-Szymanska and Xindong Wu, "Safeguarding authenticity for mitigating the harms of generative AI: Issues, research agenda, and policies for detection, fact-checking, and ethical AI",ScienceDirect, Volume 27, Issue 2, 16 February 2024.
3. Dr.G. Ramesh, R.B. Lokitha, R.R. Monisha, N.S. Neha, "Phishing Detection System using Random Forest Algorithm", International Journal for Research Trends and Innovation, Volume 8, Issue 4, 2023.
4. Rishikesh Mahajan, Irfan Siddavatam, "Phishing Website Detection using Machine Learning Algorithms", International Journal of Computer Applications (IJCA),Volume 181 – No. 23, October 2018.
5. Jayakrishnan Ashok, Pankaj Badoni, "Web Content Authentication: A Machine Learning Approach to Identify Fake and Authentic Web Pages on Internet", July 2021.

6. Suresh Kumar Krishnamoorthy, Sasikala Thankappan, "A Novel Method to Authenticate in Website Using CAPTCHA-Based Validation", Journal of Security and Privacy, 2016.
7. M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks", International Journal of Computer Science and Information Security, 2010.
8. Manar H. Alalfi, James R. Cordy, Thomas R. Dean, "A Survey of Analysis Models and Methods in Website Verification and Testing", 2007.