International Journal of Future Engineering Innovations

A Heuristic Based Approach for Detection of Phishing Websites from URL's Using Machine Learning

Gonugunta Sri Sai Amrutha [1*], Pasupuleti Bhavana [2], Thanniru kanaka Purna Chandra Rao [3], Devarakonda Rajasekhar [4]
[1-4] Kalllam Haranadhareddy Institute of Technology, Chowdavaram, Guntur, Andhra Pradesh, India

* Corresponding Author: **Gonugunta Sri Sai Amrutha**

Article Info

ISSN (online): 3049-1215
Volume: 02
Issue: 02
March-April 2025
Received: 06-02-2025
Accepted: 02-03-2025
Page No: 51-55

Abstract
Phishing attacks pose a significant threat to online security, exploiting unsuspecting users through deceptive websites that mimic legitimate platforms. This project presents a heuristic-based approach for detecting phishing websites using different ML techniques. The system is developed in Python and utilizes a robust dataset consisting of 11,055 URLs with 30 key characteristics to determine their legitimacy. Various machine learning models were evaluated, with LR demonstrating optimal performance, achieving a training accuracy of 95.00% and a validation accuracy of 92.00%. The Proposed system provides real-time analysis of URLs and enhances security by identifying Phishing attempts across multiple online platforms such as emails, SMS, social media, and websites. The project also includes an email spam detection module, further Strengthening protection against cyber threats. The developed model offers an efficient, accurate, and scalable solution to mitigate phishing attacks, contributing to a safer online environment.

DOI: https://doi.org/10.54660/IJFEI.2025.2.2.51-55

Keywords: Phishing Detection, Machine Learning, URL Analysis, Cybersecurity, Logistic Regression, Spam Detection, Heuristic Approach, Real-time Analysis

1. Introduction

The internet has become an essential part of our daily lives, but with its Growth comes an increasing number of cyber threats—one of the most dangerous being Phishing attacks. Cyber criminals create fake websites that look identical to real ones, tricking users into entering their personal information, such as Passwords and banking details. These attacks are constantly evolving, making them harder to detect with traditional security measures.

Our project aims to tackle this issue using machine learning. By analyzing the unique characteristics of website URLs, we train a model to distinguish between legitimate and phishing sites. We used a dataset of 11,055 URLs, extracting 30 key features to train different machine learning models. After testing multiple approaches, LR was found to be the most effective, achieving 95% training accuracy and 92% validation accuracy.

To further enhance security, we also developed an email spam detection system to help users identify suspicious emails that may contain phishing links. With real-time detection and a user-friendly interface, our system provides a reliable and scalable solution to protect users from online fraud. This research highlights the power of ML in Cybersecurity and how it can be used to stay ahead of cyber threats.

2. Related work

Phishing detection has been extensively studied, with traditional methods like Blacklists and rule-based systems proving inadequate against evolving threats. To address this, Researchers have explored machine learning and heuristic-based techniques for more accurate detection.

Several studies focus on URL-based phishing detection, analyzing factors such as URL length, special characters, and domain attributes (Mohammad *et al*., 2019; Alazab *et al*., 2015). Machine learning models like Decision Trees, Random Forests, and

SVMs have also been applied, with ensemble methods improving classification accuracy (Jain *et al*., 2018) [8]. More recent work incorporates deep learning approaches, such as Recurrent Neural Networks (RNNs), to detect novel Phishing patterns (Sahingoz *et al*., 2020) [5].

Hybrid approaches combining URL and content analysis have shown promise in improving accuracy and reducing false positives (Abdelhamid *et al*., 2020). Real-time detection methods, including browser extensions and user behavior monitoring, have also been explored for immediate threat identification (Bertino *et al*., 2019; Ranjan *et al*., 2022).

Our approach builds on these studies by integrating machine learning-based URL analysis with real-time phishing detection. Using a dataset of 11,055 URLs and 30 features, our Logistic Regression model achieves 95% training accuracy and 92% validation accuracy. Additionally, we incorporate an email spam detection module to enhance Cybersecurity against phishing threats.

3. Proposed Methodology

To combat Phishing attacks, we developed a ML-based system that analyzes website URLs to determine their legitimacy.

- **Data collection & preprocessing**
  We gathered 11,055 URLs, analyzing 30 key features like domain age, URL length, and HTTPS usage. The dataset was cleaned and split into 80% training and 20% testing to ensure accurate evaluation.
- **Model training & selection**
  Various machine learning models, including Logistic Regression, Random Forest, and Decision Tree, were tested. Logistic Regression performed best, achieving 95% training accuracy and 92% validation accuracy without overfitting.
- **Real-time detection & email spam filtering**
  Users can input a URL for instant classification. Additionally, we integrated an email spam detection system using Natural Language Processing (NLP) to identify phishing attempts in emails.
- **User awareness & engagement**
  To enhance security awareness, users can share phishing experiences, helping others recognize real-world scams.

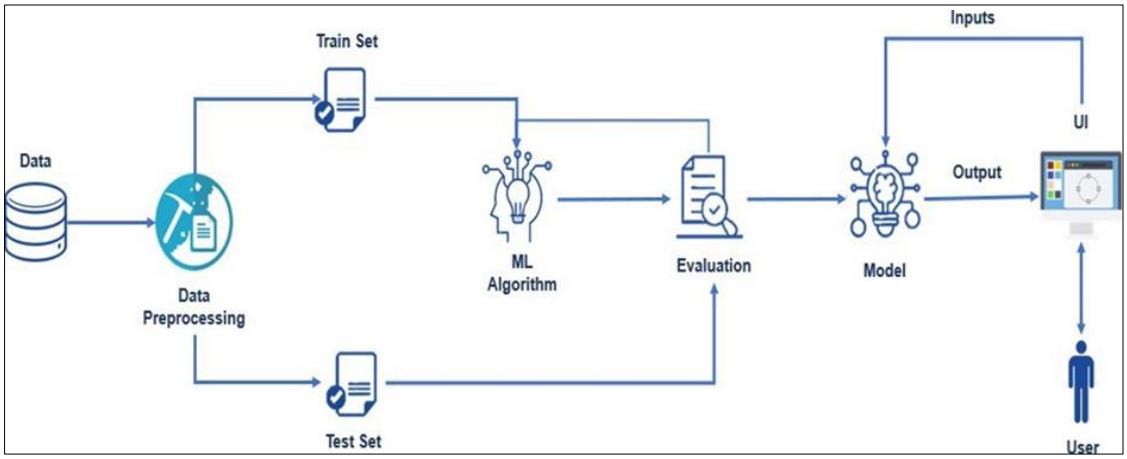

Fig 1: Machine Learning-Based Phishing Detection System Flow

This system follows a structured approach to detect phishing websites using ML (machine learning). It involves data collection, preprocessing, model training, evaluation, deployment, and real-time user interaction to classify URLs as phishing or legitimate.

Step 1: Data Collection

- Gather raw data, including phishing and legitimate URLs.
- Store data in a structured format for processing.

Step 2: Data Preprocessing

- Clean the data and remove inconsistencies.
- Extract 30 key features (e.g., domain age, HTTPS usage, URL length).
- Split the dataset into training (80%) and testing (20%) sets.

Step 3: Model Training

- Train the model using the training dataset.
- Apply machine learning algorithms such as Logistic Regression, Decision Tree, or Random Forest.

Step 4: Model Evaluation

- Test the model using the test dataset.
- Measure accuracy, precision, recall, and F1-score.
- Optimize the model for better performance.

Step 5: Model Deployment

- Deploy the trained model into a real-world application.
- Integrate it into a Flask-based web interface for user interaction.

Step 6: User Input & Detection

- The user enters a URL into the system.
- The model analyzes the input and predicts if it is phishing or legitimate.

Step 7: Output & Feedback

- The system displays the result to the user through a UI (User Interface).
- Users can report false detections to improve the model's accuracy

4. Result and Discussion

We used phishing URL's Data Set from Kaggle to train our proposed model. We split the dataset as 80% training and 20% as testing. The following metrics are considered to evaluate the performance of our ML model.

- Accuracy
- Precision
- Recall
- F1-Score

Furthermore, we compared the proposed model with existing machine learning models, including Random Forest, Decision Tree, K-Nearest Neighbors, and Logistic Regression, to analyze its performance.

Table 1: Comparative analysis of machine learning models

Model	Accuracy (%)	Precision	Recall	Fl-Score	ROC-AUC
Logistic Regression	92.0	0.91	0.90	0.91	0.94
Random Forest	91.	0.89	0.92	0.91	0.93
Decision Tree	89.	0.86	0.87	0.86	0.90
KNN	83.	0.80	0.85	0.82	0.87

Our result justify that the proposed model performs better than the existing machine learning based models

5. Experimental Results

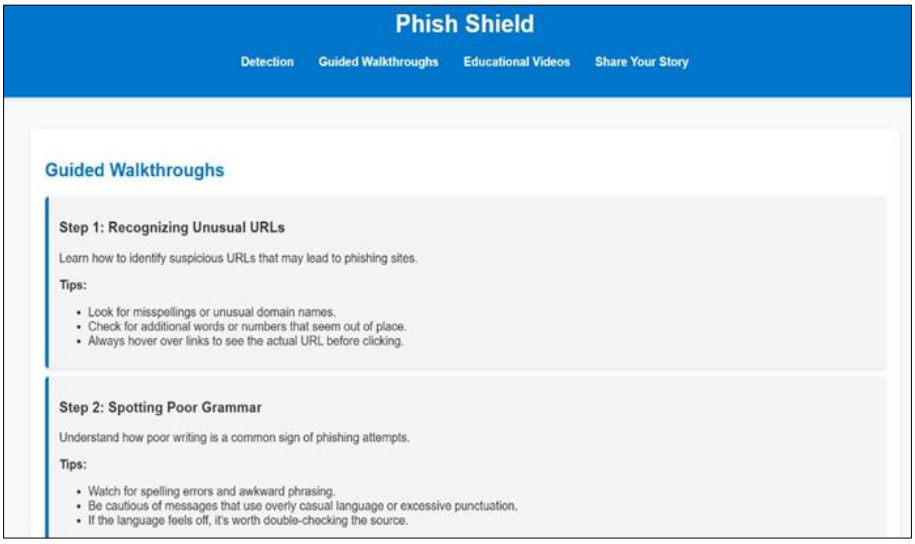

Fig 2: Phish Shield platform

This screen displays the Phish Shield platform, which provides step-by-step guidance on identifying phishing attempts. It highlights key phishing indicators such as unusual URLs and poor grammar, helping users recognize and avoid potential threats.

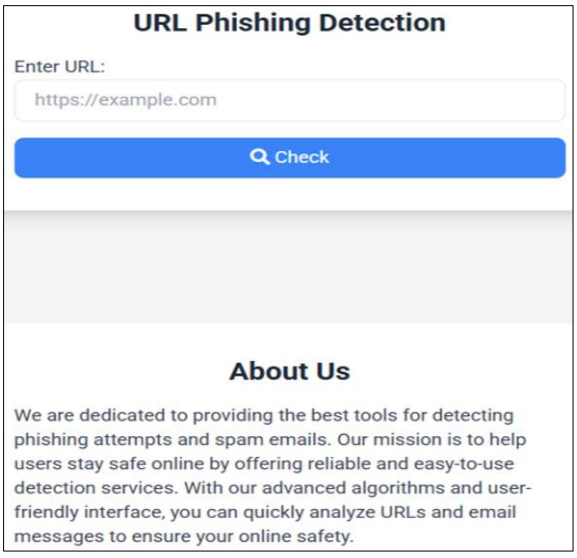

Fig 3: Prediction Page

This image displays a phishing detection tool where users can input a URL to check for phishing threats. The platform uses advanced algorithms to ensure online safety by detecting fraudulent websites.

Fig 4: URL Phishing Detection Input

Fig 5: Phishing Detection Result

The user enters a URL into the phishing detection tool to verify its authenticity. The system confirms that the entered URL is safe and belongs to a legitimate website.

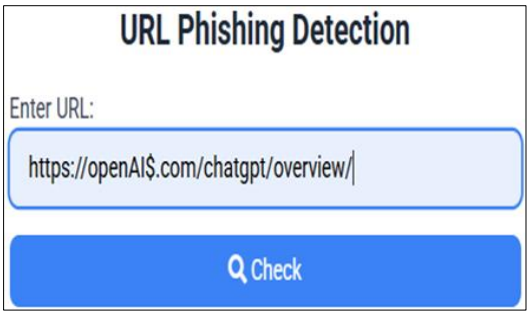

Fig 6: URL Phishing Detection Input

Fig 7: Phishing Detection Result

The user enters a URL into the phishing detection system for verification. The system detects the URL as suspicious, advising caution before proceeding.

6. Conclusion

We used a machine learning model, Logistic Regression which gave a testing accuracy of 92% without overfitting of the model, unlike the other models which are too reliant on data in the dataset our model.Our project also includes an email spam detection system which predicts the text authenticity. This helps users analyze he truthfulness of email text they receive from unknown sources. Our project lets users give their personal phishing experience and view others experience of phishing attempts. This lets the users aware of the phishing attack's types, attacker's approaches to perform phishing. This also lets the admin analyze new approaches

and integrate those solutions to better the performance of the website. In conclusion, our project a heuiristic based approach for detection of phishing websites fromurls using machine learning shows a reliable, high performance, accurate and real-time detection of phishing websites.

7. References

1. RSA. RSA fraud report. Available from: https://go.rsa.com/l/797543/2020-07-08/3njln/797543/48525/RSA_Fraud_Report_Q1_2020.pdf (Accessed 14 January 2021).
2. APWG. Phishing Attack Trends Reports, 24, November 2020. Available from: https://docs.apwg.org/reports/apwg_trends_report_q3_2020.pdf (Accessed 14 January 2021).
3. Dhamija R, Tygar JD, Hearst M. Why phishing works. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 22–27 April 2006; 581–590. (2006).
4. Jain AK, Gupta BB. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP J. on Info. Security. 2016;9:1–11. DOI: https://doi.org/10.1186/s13635-016-0034-3.
5. Sahingoz OK, Buber E, Demir O, Diri B. Machine learning based phishing detection from URLs. Expert Syst. Appl. 2019;117:345–357. (2019).
6. Haruta S, Asahina H, Sasase I. Visual similarity-based phishing detection scheme using image and CSS with target website finder. 978-1-5090-5019-2/17/$31.00 ©2017 IEEE. (2017).
7. Cook DL, Gurbani VK, Daniluk M. Phishwish: A stateless phishing filter using minimal rules. In: Tsudik G, editor. Financial Cryptography and Data Security; 324. Berlin, Heidelberg: Springer-Verlag; 2008.
8. Jain AK, Gupta BB. A machine learning-based approach for phishing detection using hyperlinks information. J. Ambient. Intell. Humaniz. Comput. DOI: https://doi.org/10.1007/s12652-018-0798-z. (2018).
9. Li Y, Yang Z, Chen X, Yuan H, Liu W. A stacking model using URL and HTML features for phishing webpage detection. Futur. Gener. Comput. Syst. 2019;94:27–39. (2019).
10. Xiang G, Hong J, Rose CP, Cranor LC. CANTINA+: A feature-rich machine learning framework for detecting phishing websites. ACM Trans. Inf. Syst. Secur. 2011;14(2):1–28. DOI: https://doi.org/10.1145/2019599.2019606.