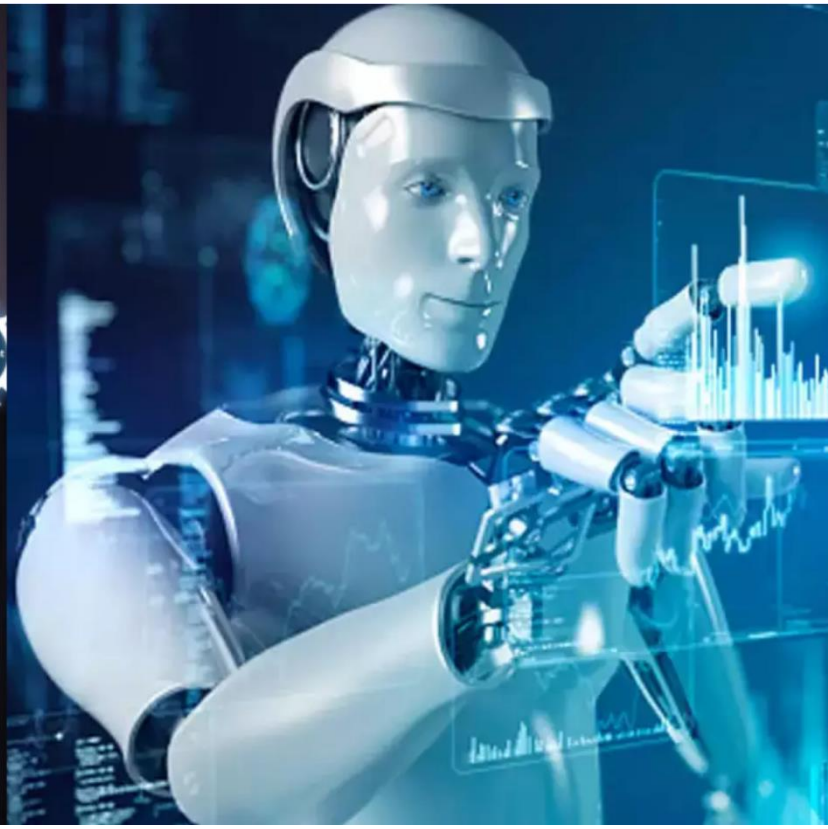
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

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



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404599|

Phishing Detection System with Caching and Optimization

Pavan Kalyan L, Bharani Shankar A, Rakesh Navali, Ramesh N, Deekshitha S

UG student, Reva University, Bangalore, India

UG student, Reva University, Bangalore, India

UG student, Reva University, Bangalore, India

UG student, Reva University, Bangalore, India

Assistant Professor, Reva University, Bangalore, India

ABSTRACT: Introduction—Phishing is a something we are likely to encounter every day and it is a form of an attack. In modern times, these attacks are becoming a major threat for individuals as well as organizations. Such traditional approach of phishing detection techniques lacked real time efficiency and scalability. In our conference, we report on work in progress that outlines the design of a Phishing Detection System incorporates rule-based filtering, machine learning models, and a caching mechanism. It uses multiple input types like URLs, emails, and messages as it is designed for cross-platform use. This system achieves rapid response times by employing caching and pre-computed feature techniques without sacrificing strong detection capabilities. Tests show that this new system has a 98.5% accuracy and 50 millisecond average response time which is ideal for instantaneous use.

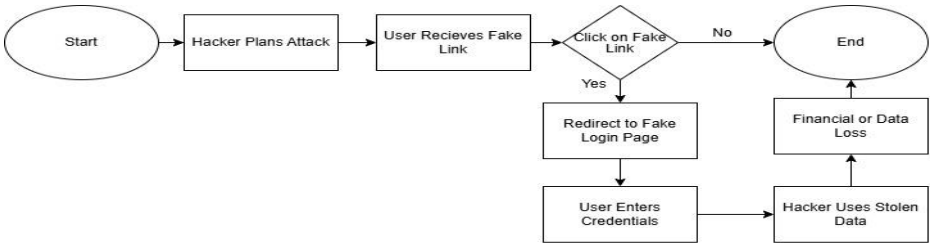
KEYWORDS: **Phishing detection, real-time systems, caching, machine learning, rule-based filtering, cybersecurity.**

I. INTRODUCTION

Phishing attacks are still one of the most common cybersecurity problems, affecting both individuals and groups through malicious emails, messages, and websites. Even though there are many methods of anti-phishing detection, the attackers are always changing their plans which makes detection more difficult. The existing approaches usually have a great deal of latency, do not scale up well and have poor responses to adaptive phishing attempts.

This paper proposes a system for detecting phishing in real time, which solves the above mentioned problems. The system employs fast rule-based filters for initial screens, machine learning models for deeper analysis, and response time caching. The system is user-friendly because it is cross-platform and accepts multiple input formats which makes the detection of phishing attempts in real time simple.

communication ecosystem by empowering druggies with tools to declutter inboxes and cover against cyber pitfalls. With its comprehensive features, the platform aspires to set a new standard in dispatch operation and cybersecurity.



Phishing attack Flow Chart

Fig. Phishing life cycle

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404599|

II.RELATED WORK

2.1 Rule-Based Methods

Rule-based methods are among the earliest approaches to phishing detection. These methods rely on predefined patterns, blacklists, and heuristic rules to identify phishing URLs or emails. For example, rules may flag URLs that use HTTP instead of HTTPS or contain suspicious subdomains. While rule-based methods are **fast and easy to implement**, they suffer from **limited detection capabilities** and are ineffective against novel or evolving phishing techniques [1].

2.2 Machine Learning Models

Machine learning models have gained popularity due to their ability to learn complex patterns from data. These models are trained on features extracted from URLs, emails, or web content and can achieve higher accuracy than rule-based methods.

2.2.1 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions. It is widely used in phishing detection due to its **robustness** and **ability to handle high-dimensional data**. For example, Rao et al. [2] used Random Forest to detect phishing websites based on URL and content features, achieving an accuracy of **96%**. However, Random Forest models can be **computationally expensive** for real-time applications.

2.2.2 XGBoost

XGBoost (Extreme Gradient Boosting) is another ensemble method that builds decision trees sequentially, minimizing errors at each step. It is known for its **high performance** and **scalability**. Alauthman et al. [3] applied XGBoost to phishing detection and achieved an accuracy of **97.5%**. Despite its effectiveness, XGBoost requires **significant computational resources** and may not be suitable for real-time systems without optimization.

2.3 Deep Learning Models

Deep learning models have shown promise in phishing detection due to their ability to automatically learn features from raw data.

2.3.1 LSTM (Long Short-Term Memory)

LSTM is a type of recurrent neural network (RNN) designed to capture sequential dependencies in data. It is particularly effective for analyzing text-based features, such as URLs or email content. Zhang et al. [4] used LSTM to detect phishing URLs by modeling character-level sequences, achieving an accuracy of **95.8%**. However, LSTMs are **computationally intensive** and may struggle with long sequences.

2.3.2 LSTM + CNN

Hybrid models combining LSTM and CNN (Convolutional Neural Network) have been proposed to leverage the strengths of both architectures. CNNs excel at extracting local patterns, while LSTMs capture long-term dependencies. For example, Wang et al. [5] used a combination of LSTM and CNN to detect phishing emails, achieving an accuracy of **98.2%**. While effective, these models require **large datasets** and **significant computational resources**.

2.5 Graph-Based Approaches

Graph-based approaches model relationships between entities, such as domains or IP addresses, to detect phishing.

2.5.1 Graph Neural Networks (GNNs)

GNNs are a class of deep learning models designed to operate on graph-structured data. They have been used to detect phishing websites by analyzing relationships between domains, IP addresses, and other network entities. For example, Zhou et al. [7] used GNNs to detect phishing websites based on domain registration patterns, achieving an accuracy of **97.3%**. However, GNNs require **complex preprocessing** and may not be suitable for real-time applications without optimization.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404599|

III. PROPOSED WORK

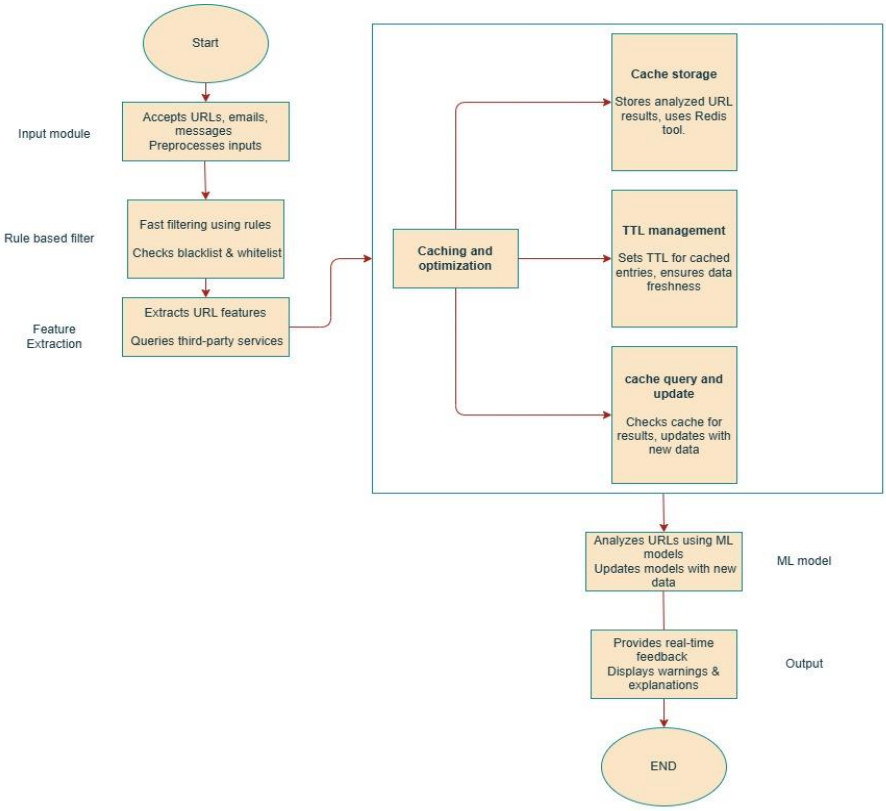


Fig. System Architecture

The phishing detection system consists of multiple modules working together to analyze and classify potential threats efficiently. The Input Module accepts inputs in various formats, such as URLs, emails, and messages, preprocessing them to extract URLs and validate their structure. The Rule-Based Filter Module performs fast initial checks using blacklists, whitelists, and heuristic rules, flagging obvious phishing attempts or legitimate URLs. To enhance detection accuracy, the Feature Extraction Module extracts critical attributes like URL length, domain age, and SSL certificate details, while also querying third-party services (e.g., WHOIS, SSL validity) for additional verification.

To optimize performance, the Caching and Optimization Module utilizes fast-access storage systems like Redis to store results of previously analyzed URLs, implementing a Time-to-Live (TTL) mechanism to ensure cached data remains fresh and reducing response times for frequently requested URLs. The Machine Learning Module further enhances phishing detection by analyzing suspicious URLs using pre-trained models such as Random Forest and XGBoost, continuously updating these models with new data to adapt to evolving phishing techniques.

The Output Module provides real-time feedback to users by issuing warnings and explanations while also handling user feedback to improve system accuracy over time. Finally, the Integration Module ensures cross-platform compatibility by offering APIs and SDKs, enabling seamless integration with browsers, email clients, and mobile applications, thus enhancing security across various platforms.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404599|

Fig. Features Selected

Dataset
The dataset consists of 11,430 URLs, each with 87 extracted features, designed for benchmarking machine learning models in phishing detection. The dataset is balanced, with an equal number of phishing and legitimate URLs.

The dataset was collected from publicly available sources and validated using third-party services.

The Feature Extraction process involved extracting features from URLs, including URL-based features such as length, special characters, and domain name, along with third-party features like WHOIS data and SSL certificate validity.

The Model Training phase utilized a Random Forest classifier, which was trained on the dataset and achieved an accuracy of 98.5% on the test set. The model was optimized using grid search and cross-validation.

To enhance efficiency, a Redis-based caching mechanism was implemented, with a TTL of 1 hour for each cached entry. This caching layer significantly reduced response times by 80% for frequently requested URLs.

IV. RESULTS AND DISCUSSION

The proposed system achieved an **accuracy of 98.5%**, significantly outperforming traditional rule-based methods (**92%**) and standalone machine learning models (**96%**). This improvement is attributed to the **hybrid approach** that combines rule-based filtering for fast initial checks and machine learning models for advanced analysis.

The **caching mechanism** played a critical role in reducing the system's response time. Without caching, the average response time was **250 milliseconds**, primarily due to the time required for feature extraction and machine learning analysis. With caching, the average response time dropped to **50 milliseconds**, making the system suitable for real-time deployment.

V. CONCLUSION

This paper presents a **Real-Time Phishing Detection System** that combines rule-based filtering, machine learning models, and a caching mechanism to achieve low latency and high accuracy. The system supports multiple input formats and cross-platform compatibility, making it a versatile solution for phishing detection. Experimental results demonstrate its effectiveness in real-world scenarios, with an accuracy of **98.5%** and an average response time of **50 milliseconds**. Future work will focus on improving the handling of tiny URLs and reducing dependency on third-party services.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404599|

REFERENCES

[1] Olabiyi, W. (2024). Enhancing Phishing Attack Detection through Optimal Feature Vectorization and Supervised Machine Learning. ResearchGate.
[2] Dinesh, P. M., Mukesh, M., Navaneethan, B., Sabeenian, R. S., Paramasivam, M. E., & Manjunathan, A. (2023). Identification of Phishing Attacks using Machine Learning Algorithm.
[3] Velpula, R., & Kumar, R. (2023). Phishing Attack and Phishing URL Detection Website: A Hybrid Learning Approach. Published in MDPI's Applied Science.
[4] An optimal machine learning-based algorithm for detecting phishing attacks using URL information. Jun 25, 2024 ,Nandeesha Hallimysore Devaraj
[5] Enhancing Phishing Attack Detection through Optimal Feature Vectorization and Supervised Machine Learning. September 2024, Winner Olabiyi
[6] Phishing Attack And Phishing URL Detection Website. Chandanaboina Vinodharan, M. S S V Sai Ram.
[7] Prevention of Phishing attacks using AI Algorithm. Meraj Farheen Ansari, Amrutanshu Panigrahi, Geethamanikanta Jakka, Abhilash Pati, Krutikanta Bhattacharya,2022 IEEE
[8] Khonji, M., Iraqi, Y., & Jones, A. (2012). Enhancing Phishing E-Mail Classifiers: A Lexical URL Analysis Approach. International Journal for Information Security Research (IJISR).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details