Advances in Computer Engineering and Science (ACES)
(2026)
Volume -1, Issues – 1

A Hybrid Phishing Website Detection Framework Using URL Feature Analysis and Visual Intelligence

P.Deepika
Department of Computer Science and Applications,
Vivekanandha College of Arts and Sciences for Women (Autonomous),
Elayampalayam,Tiruchengode
Email: deepikapalanisamy80@gmail.com

Mrs. S. Nathiya
Department of Computer Science and Applications,
Vivekanandha College of Arts and Sciences for Women (Autonomous),
Elayampalayam,Tiruchengode
Email: s.nathiyamsc@gmail.com

Abstract: Phishing attacks have become one of the most significant cyber security threats in recent years, targeting users by creating fraudulent websites that closely resemble legitimate platforms. Traditional detection methods such as blacklist-based systems and URL analysis are insufficient to detect newly emerging and visually deceptive phishing websites. This paper proposes a Hybrid AI-Based Phishing Website Detection Framework that integrates both URL feature analysis and visual intelligence. The system extracts structural features from URLs and analyses them using machine learning techniques, while webpage screenshots are processed using a Convolutional Neural Network (CNN) to identify phishing patterns. A hybrid decision engine combines the outputs of both models to generate the final classification. The proposed system is implemented as a browser extension that enables real-time phishing detection and provides risk scores along with explanations. Experimental results show that the hybrid model achieves an accuracy of 94.3%, outperforming traditional single-method approaches. This approach enhances cyber security by providing a more reliable and intelligent phishing detection system

Keywords: Phishing Detection, Machine learning, Deep learning, Convolutional Neural Network (CNN), URL Feature Analysis, Visual Intelligence, Cyber Security, Hybrid Model, Browser Extension, Real-time Detection.

I. INTRODUCTION

The rapid growth of internet technology has transformed the way individuals and organizations communicate, conduct business, and access information. Online platform such as banking systems, e-commerce websites, and social media applications has become an essential part of daily life. However, this increased reliance on online services has also led to a rise in cyber threats, among which phishing attacks are one of the most dangerous. Phishing is a type of cyber-attack in which attackers create fake websites or send fraudulent emails that mimic legitimate platforms in order to deceive users into revealing sensitive information such as usernames, password, and financial details. These attacks often lead to serious consequence such as identify theft, financial loss, and data breaches.

Traditional phishing detection methods primarily rely on blacklist database and URL based analysis techniques. While blacklist systems can detect known phishing websites, they fail to identify newly created websites that have not yet been added to the database. similarly, URL-Based detection methods analyse features such as domain length and suspicious keywords, but attackers can easily bypass these methods by creating carefully crafted URLs that appear legitimate.

In addition, modern phishing websites are designed to visually legitimate websites by copying layouts, logs, and design elements. This makes it difficult for traditional detection systems to identify phishing attacks based solely on URL features. Therefore, there is a need for a more advanced detection system that combine both structural and visual analysis. This paper proposes a hybrid approach that integrates machine learning and deep learning techniques to improve phishing detection accuracy and reliability.

II. LITERATURE SURVEY

Phishing detection has been an active area of research in cyber security due to the increasing number of online fraud attacks. Several techniques have been proposed by researchers to identify phishing websites, including blacklist-based detection, URL features analysis, content-based analysis, and machine learning approaches. One of the earliest and most commonly used methods for phishing detection is the blacklist-based approach. In this method, a database of known phishing websites is maintained, and any visited website is checked against this list. If a match is found, the website is flagged as phishing. Although this method is simple and effective for detecting known phishing sites, it has a major limitation. Newly created phishing websites cannot be detected until they are added to the blacklist, which creates a delay in detection. Another widely used techniques is URL-based features analysis. In this approach, machine learning algorithms are used to analyse the characteristics of a URL, such as its length, number of subdomains, use of special character, and presence of suspicious keywords. Researchers have shown that these features can help distinguish phishing UELs from legitimate ones. However, attackers have become more sophisticated and can create URLs that closely resemble legitimate domains, reducing the effectiveness of this method

Content based detection methods analyse the HTML structure and textual content of webpages. These methods examine elements such as forms, scripts, and hyperlinks to identify suspicious patterns. While content-based approaches can provide better detection than simple URL analysis, they are often computationally expensive and may not be suitable for real-time detection systems. In recent years, deep learning techniques have been applied to phishing detection, especially for visual analysis of webpages. Convolutional Neural

Networks (CNNs) have been widely used to analyse webpage screenshots and identify visual similarities between phishing websites and legitimate websites. These models can detect patterns such as fake login pages, logo imitation, and layout similarities. However, visual analysis alone may not be sufficient, as some phishing websites may not visually resemble legitimate sites

Some advanced systems combine multiple detection techniques to improve accuracy. Hybrid approaches that integrate URL analysis, content analysis, and visual analysis have shown promising results. These systems leverage the strengths of different methods to overcome the limitations of individual approaches. However, many existing hybrid systems are complex and do not provide real-time detection capabilities. From the literature review, it is clear that no single method is sufficient to effectively detect all types of phishing attacks. Blacklist systems fail to detect new attacks, URL –based methods can be bypassed, and visual analysis alone may not be reliable. Therefore, there is a need for a hybrid detection system that combines multiple techniques to achieve higher accuracy and reliability. This paper addresses these limitations by proposing a hybrid phishing detection framework that integrates URL feature analysis and CNN-based visual intelligence. The proposed system aims to provide accurate and real-time detection of phishing websites while improving overall system performance.

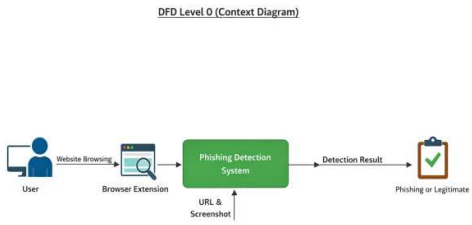

Figure 1 : Phishing Detection System

III. PROPOSED SYSTEM

The proposed system is a hybrid AI-Based Phishing Website Detection Framework designed to overcome the limitations of traditional phishing detection methods. The system combines both machine learning and deep learning techniques to analyse websites using URL features and visual intelligence. The architecture of the proposed system consists of multiple modules that work together to detect phishing websites in real time. These modules include the Browser Extension Module, URL Features Extraction Module, Visual Analysis Module, Hybrid Decision Engine, and Result Display Module.

The Browser Extension Model acts as the interface between the user and the system. When a user visits a website, the extension automatically captures the webpage URL and takes a screenshot of the page. This information is then sent to the backend server for further analysis. The URL Feature Extraction Module analyse the structure of the website URL. It extracts important features such as URL length, number of subdomains, presence of special characters, and suspicious keyword. These features are passed to a machine learning model, which predicts whether the URL is phishing or legitimate.

The Visual Analysis Module processes the captured webpage screenshot using a Convolutional Neural Network (CNN). The CNN model is trained to identify visual patterns commonly found in phishing websites, such as fake login forms websites that cannot be identified using URL features alone. The Hybrid Decision Engine plays a crucial role in combining the outputs of both the URL analysis model and the CNN model. It calculates a final risk score based on the predictions from both models. If the combined score exceeds a predefined threshold, the website is classified as phishing; otherwise, it is considered legitimate. The Result Display Module presents the final output to the through the browser extension interface. The system displays the classification result along with a risk score and explanation of suspicious features detected. This helps users understand the reason behind the classification.

The proposed system offers several advantages over traditional methods. It improves detection accuracy by combining multiple techniques, provides real-time phishing detection, and enhances user awareness through explainable outputs. This hybrid approach ensures better performance and reliability in identifying phishing websites.

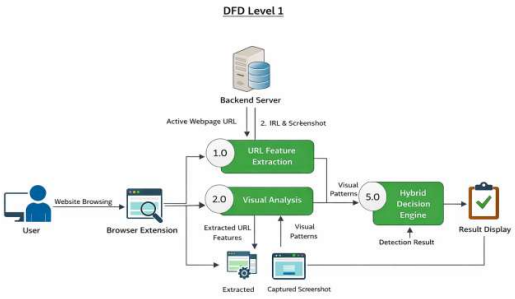

Figure 2 : Website Browsing

IV. METHODOLOGY

The methodology of the proposed system involves the integration of machine learning and deep learning techniques to detect phishing websites. The system processes both URL features and webpage screenshots to make accurate predictions.

4.1 URL Features Extraction

The first step in the methodology is extracting features from the website. URLs often contain patterns that can indicate whether a website is phishing or legitimate. The system extracts several important features, including:

- Length of the URL
- Number of subdomains
- Presence of special character such as "@", "-", "_"
- Use of IP address instead of domain name
- Presence of suspicious keyword such as "login", "verify", "secure", and "update"

These features are converted into numerical values and used as input for a machine learning model. The model analyses these features and predicts whether the URL is suspicious.

4.2 CNN-Based Visual Analysis

The second step involves analysing the visual appearance of the webpage using a Convolutional Neural Network (CNN). The CNN model processes the screenshot of the webpage to identify visual patterns that are commonly found in phishing websites.

The architecture of the CNN model consists of multiple layers:

- **Input Layer:** Receives the webpage image resized to a fixed dimension (e.g., 224*224 pixels)
- **Convolutional Layer:** Apply filters to extract important visual features such as edges, shapes, and textures.
- **ReLU Activation Function:** Introduces non-linearity to improve learning capability
- **Pooling Layer:** Reduce the spatial dimensions of feature maps while preserving important information
- **Flatten Layer:** Converts feature maps into a one-dimensional vector
- **Fully Connected Layer:** Performs classification based on extracted features
- **Output Layer**: Produces the final predictions (phishing or legitimate)

The CNN model is trained using labelled datasets containing phishing and legitimate website screenshots. This allows the model to learn visual differences between phishing and genuine websites.

4.3 Hybrid Decision Engine

The first step in the methodology is combining the outputs of the URL analysis model and the CNN model using a hybrid decision engine. The purpose of this module is to improve detection accuracy by considering both structural and visual features.

The final scores calculated using a weighted formula:

Final Score=0.5* URL Score + 0.5* CNN Score

If the final score is greater than or equal to a threshold value (e.g., 50), the website is classified as phishing. Otherwise, it is classified as legitimate.

This hybrid approach ensures that even if one model fails to detect a phishing website, the other can model can compensate for it. As a result, the overall accuracy and reliability of the system are significantly improved

4.4 Advantages of Methodology

- Combines multiple detection techniques
- Improves accuracy and reduces false positives
- Enables real-time phishing detection
- Provides explainable results for users

IV. IMPLEMENTATION

The proposed phishing detection system is implemented using a combination of programming languages, frameworks, and tools to ensure efficient and real-time performance. The system is designed with a modular architecture that integrates both frontend and backend components. The backend of the system is developed using Python, which provides strong support for machine learning and deep learning libraries. The URL feature analysis is implemented using machine learning algorithms, while the visual analysis is performed using a Convolutional Neural Network (CNN) built using TensorFlow AND Keras frameworks. The dataset used for training the models includes both phishing and legitimate website data. URL datasets are collected from publicly available sources such as Phish Tank and Kaggle repositories. Webpage screenshots are collected and labled accordingly for training the CNN model. The dataset is divided into training and testing sets, typically in a 70:30 ratio.

The CNN model is trained using labelled images of phishing and legitimate websites. Image preprocessing techniques such as resizing, normalization, and augmentation are applied to improve model performance. The trained model is saved and loaded during runtime for prediction. The frontend of the system is implemented as a browser extension using HTML, CSS, and JavaScript. The extension captures the current webpage URL and takes a screenshot of the webpage. This data is then sent to the backend server using API calls. Flask is used as the backend web framework to handle communication between the browser extension and the machine learning models. The backend processes the received data and returns the prediction result to the extension in real time.

The system workflow is as follows: when a user visits a website, the extension captures the URL and screenshot, sends them to the backend server, where both the URL analysis model and CNN model process the data. The hybrid decision engine combines the results and sends the final classification back to the extension, which displays it to the user. The implementation ensures that the system operates efficiently with minimal delay, making it suitable for real-time phishing detection application.

V. RESULT AND ANALYSIS

The performance of the proposed hybrid phishing detection system is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. The evaluation is conducted using a dataset consisting of both phishing and legitimate websites. The URL-based model and CNN-based model are first evaluated individually, followed by the evaluation of the hybrid model. The results show that the hybrid approach significantly outperforms the individual models

Performance Metrics

- **Accuracy:** Measures the overall correctness of the model

- **Precision:** Measures the correctness of positive predictions
- **Recall:** Measures the ability to detect phishing websites
- **F1-Score:** Harmonic mean of precision and recall

VI. Experimental Results

The results clearly indicate that the hybrid model provides better performance compared to individual models. The URL-based model performs well in detecting structural anomalies but fails to identify visually deceptive phishing websites. On the other hand, the CNN model can detect visual similarities but may miss structural patterns.

Table 1 : Comparison of Models

MODEL	ACCURACY
URL-Based Model	85.6%
CNN Model	88.2%
Hybrid Model	94.3%

By combining both approaches, the hybrid model achieves higher accuracy and reduces false positives and false negatives. The system also demonstrates good performance in detecting newly created phishing websites. The real-time implementation using a browser extension further enhances the usability of the system, making it suitable for practical deployment

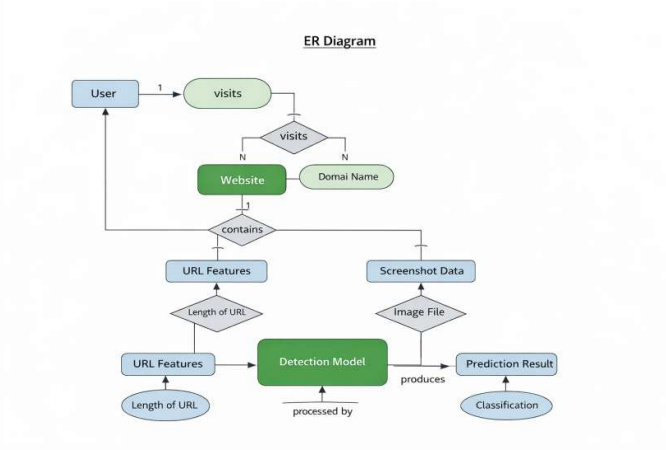

Figure 3 : ER Diagram

VII. CONCLUSION

This paper presented a Hybrid AI-Based Phishing Website Detection Framework that integrates URL features analysis and CNN-based visual intelligence. The proposed system effectively addresses the limitation of traditional phishing detection methods by combining structural and visual analysis techniques. The system is capable of detecting phishing websites in real time using a browser extension, providing users with immediate feedback and risk assessment. The hybrid decision engine improves the overall detection accuracy and reliability of the system. Experimental results demonstrate that the proposed model achieves higher accuracy compared to individual detection approaches. The system also reduces false positives and improves detection of visually deceptive phishing websites. In future work, the system can enhanced by incorporating advanced deep learning architectures such as ResNet or Transformer-based models. Additionally, larger datasets can be used to further improve the performance and generalization of the model.

VIII. REFERENCE

1. A. Jain and B. Gupta, “Phishing detection: analysis of visual similarity-based approaches,” *Security and Communication Networks*, 2017.

2. R. Verma and K. Dyer, “On the character of phishing URLs,” *Proceedings of ACM Conference*, 2015.

3. PhishTank Dataset, “https://www.phishtank.com”

4. Kaggle Dataset, “Phishing Website Dataset”

5. S. Garera et al., “A framework for detection of phishing attacks,” *ACM Workshop*, 2007.

6. J. Ma, L. Saul, S. Savage, and G. Voelker, “Beyond blacklists: Learning to detect malicious web sites from suspicious URLs,” *Proceedings of ACM SIGKDD*, 2009.

7. M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” *IEEE Communications Surveys & Tutorials*, 2013.

8. D. Canali, M. Cova, G. Vigna, and C. Kruegel, “Prophiler: A fast filter for the large-scale detection of malicious web pages,” *International World Wide Web Conference*, 2011.

9. A. Le, A. Markopoulou, and M. Faloutsos, “PhishDef: URL names say it all,” *IEEE INFOCOM*, 2011.

10. S. Marchal, J. Francois, R. State, and T. Engel, “PhishScore: Phishing detection using URL features,” *IEEE International Conference on Communications*, 2014.

11. A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” *NeurIPS*, 2012.

12. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *ICLR*, 2015.

13. C. Szegedy et al., “Going deeper with convolutions,” *IEEE CVPR*, 2015.

14. H. Abutair and A. Belghith, “Using case-based reasoning for phishing detection,” *IEEE Conference*, 2017.

15. S. Ding, S. Ward, and X. Wang, “A hybrid approach for phishing detection using machine learning and deep learning,” *IEEE Access*, 2019.

16. Y. Rao and S. Pais, “Detection of phishing websites using neural networks,” *IEEE Conference*, 2019.

17. PhishTank, “Phishing Website Dataset,” Available: https://www.phishtank.com