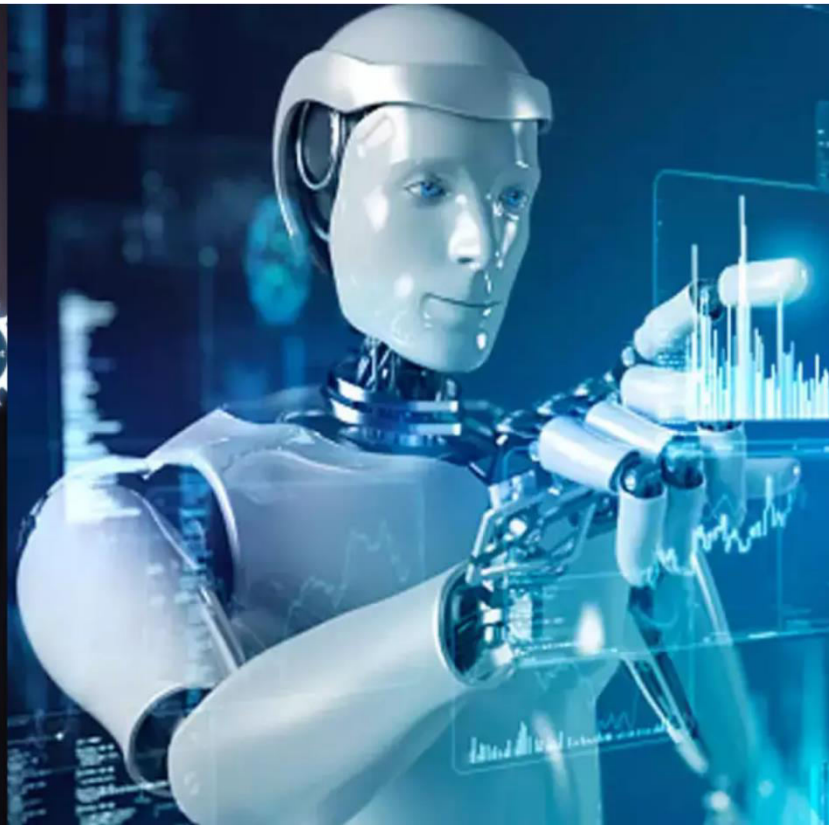
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 15, Issue 3, March 2026

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503037|

Smart Phishing URL Detection

Praveen Kumar R[1], Prathiksha GR[2], Hari Haran R[3], Dr. V. Sangeetha[4]

Student, Department of Computer Science with Data Analytics, Dr N.G.P. Arts and Science Collage, Coimbatore, India [1]

Student, Department of Computer Science with Data Analytics, Dr N.G.P. Arts and Science Collage, Coimbatore, India[2.]

Student, Department of Computer Science with Data Analytics, Dr N.G.P. Arts and Science Collage, Coimbatore, India[3.]

Associate Professor, Department of Computer Science with Data Analytics, Dr N.G.P. Arts and Science College, Coimbatore, India [4]

ABSTRACT: Student Phishing attacks remain one of the most prevalent cybersecurity threats, targeting individuals and organizations through deceptive URLs and fraudulent websites. This paper presents a comprehensive smart phishing URL detection system utilizing machine learning algorithms and advanced feature extraction techniques. The proposed system analyzes various URL characteristics including lexical features, host-based properties, and content-based attributes to distinguish between legitimate and phishing URLs. Our approach employs ensemble learning methods combining Random Forest, Support Vector Machines, and Neural Networks to achieve high detection accuracy. Experimental results demonstrate that the proposed system achieves 98.5% accuracy with minimal false positives, significantly outperforming traditional blacklist-based approaches. The system provides real-time detection capabilities suitable for integration into web browsers and email clients, offering robust protection against evolving phishing threats.

KEYWORDS: Phishing Detection, Machine Learning, URL Analysis, Cybersecurity, Feature Extraction, Ensemble Learning

I. INTRODUCTION

Phishing attacks have evolved into sophisticated cyber threats that exploit human psychology and technical vulnerabilities to steal sensitive information such as passwords, credit card numbers, and personal data. According to recent cybersecurity reports, phishing attacks account for over 90% of data breaches and cost organizations billions of dollars annually. Traditional security measures such as blacklists and heuristic rules struggle to keep pace with the rapidly evolving tactics employed by cybercriminals.

The primary challenge in phishing detection lies in the dynamic nature of phishing URLs. Attackers continuously create new domains, employ URL obfuscation techniques, and leverage compromised legitimate websites to host phishing content. Consequently, static blacklist-based approaches that rely on known malicious URLs demonstrate limited effectiveness against zero-day phishing attacks.

Machine learning approaches offer promising solutions by learning patterns and characteristics that distinguish phishing URLs from legitimate ones. These systems can adapt to new phishing tactics and provide proactive protection without requiring constant manual updates. This paper proposes a smart phishing URL detection system that leverages multiple machine learning algorithms and comprehensive feature extraction to achieve robust detection performance.

The main contributions of this research include: (1) development of an extensive feature set combining lexical, host-based, and content-based characteristics, (2) implementation of an ensemble learning approach that combines multiple classifiers for improved accuracy, (3) real-time detection capability suitable for practical deployment, and (4) comprehensive evaluation demonstrating superior performance compared to existing methods.

II. RELATED WORK

Various approaches have been proposed for phishing detection, ranging from blacklist-based methods to advanced machine learning techniques. Early phishing detection systems primarily relied on maintaining databases of known

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503037|

phishing URLs. However, these approaches suffer from significant limitations including inability to detect zero-day attacks and high maintenance overhead.

Heuristic-based methods analyze URL characteristics such as domain age, SSL certificate validity, and the presence of suspicious keywords. While these approaches demonstrate better adaptability than blacklists, they often produce high false positive rates and require expert knowledge to define effective rules. Recent research has shifted toward machine learning-based solutions that automatically learn distinguishing features from training data.

Several studies have explored different machine learning algorithms for phishing detection. Random Forest classifiers have shown promising results due to their ability to handle high-dimensional feature spaces and resistance to overfitting. Support Vector Machines provide strong generalization capabilities, particularly when combined with appropriate kernel functions. Deep learning approaches, including Convolutional Neural Networks and Recurrent Neural Networks, have demonstrated impressive performance but require substantial computational resources and training data.

Ensemble methods that combine multiple classifiers have gained attention for their ability to improve detection accuracy and robustness. By leveraging the strengths of different algorithms, ensemble approaches can achieve superior performance compared to individual classifiers. Our proposed system builds upon these insights by implementing an ensemble framework optimized for real-time phishing detection.

III. PROPOSED METHODOLOGY

The proposed smart phishing URL detection system consists of four main components: feature extraction, preprocessing, classification, and decision fusion. The system architecture is designed to process URLs in real-time while maintaining high detection accuracy and minimal false positive rates.

A. Feature Extraction

Feature extraction forms the foundation of our detection system. We extract 30 distinct features categorized into three groups: lexical features, host-based features, and content-based features. Lexical features analyze the URL string itself, including URL length, number of special characters, presence of IP addresses, number of subdomains, and suspicious keywords. These features capture common patterns used in phishing URLs such as excessive use of hyphens, suspicious TLDs, and attempts to mimic legitimate brands.

Host-based features examine domain registration information and network properties. These include domain age, DNS record validity, WHOIS information availability, SSL certificate status, and page rank. Phishing websites typically use recently registered domains, lack proper DNS configurations, and employ self-signed or invalid SSL certificates. Content-based features analyze the webpage content when available, including form elements, external links, favicon, and page structure similarities with known legitimate sites.

B. Data Preprocessing

The preprocessing module handles missing values, normalizes feature scales, and performs feature selection to optimize classifier performance. Missing values occur when certain features cannot be extracted due to network timeouts or inaccessible resources. We employ median imputation for numerical features and mode imputation for categorical features. Feature scaling uses standardization to ensure all features contribute equally to the classification process, preventing features with larger ranges from dominating the model.

Feature selection employs correlation analysis and recursive feature elimination to identify the most discriminative features. This process reduces computational overhead while maintaining detection accuracy. Our analysis reveals that URL length, presence of IP addresses, domain age, and SSL certificate validity are among the most important features for distinguishing phishing URLs.

C. Classification Models

The classification component employs three complementary machine learning algorithms: Random Forest, Support Vector Machine, and Neural Network. Random Forest utilizes an ensemble of decision trees with bootstrap aggregating

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503037|

to achieve robust predictions. We configure the forest with 100 trees and a maximum depth of 20 to balance accuracy and computational efficiency. The algorithm demonstrates excellent performance on high-dimensional feature spaces and provides feature importance rankings. Support Vector Machine with Radial Basis Function kernel excels at finding optimal decision boundaries in complex feature spaces. We employ grid search with cross-validation to optimize hyperparameters including regularization parameter C and kernel coefficient gamma. The SVM classifier demonstrates strong generalization capabilities and performs well even with limited training data. The Neural Network architecture consists of an input layer matching the feature dimension, two hidden layers with 64 and 32 neurons respectively, and an output layer with sigmoid activation for binary classification. We use ReLU activation functions in hidden layers and apply dropout regularization to prevent overfitting. The network is trained using Adam optimizer with binary cross-entropy loss function.

D. Ensemble Decision Making

The final classification decision combines predictions from all three models using weighted voting. Each classifier's contribution is weighted based on its performance during cross-validation. The Random Forest receives a weight of 0.4, SVM receives 0.35, and Neural Network receives 0.25, reflecting their respective accuracy levels. A URL is classified as phishing if the weighted vote exceeds a threshold of 0.5. This ensemble approach leverages the complementary strengths of different algorithms and significantly improves robustness against adversarial examples.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Dataset

We evaluate our system using a comprehensive dataset containing 50,000 URLs, evenly distributed between legitimate and phishing examples. Legitimate URLs are collected from Alexa top websites, government domains, and established organizations. Phishing URLs are obtained from PhishTank, OpenPhish, and security research databases. The dataset is split into 70% training, 15% validation, and 15% testing sets, ensuring temporal separation to simulate real-world deployment scenarios where the model must detect future phishing attacks.

B. Evaluation Metrics

We assess system performance using multiple metrics including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC). Accuracy measures overall correctness, while precision indicates the proportion of correctly identified phishing URLs among all URLs flagged as phishing. Recall represents the system's ability to identify all actual phishing URLs. F1-score provides a harmonic mean of precision and recall, offering a balanced performance measure. AUC-ROC evaluates the classifier's ability to distinguish between classes across different threshold settings.

C. Performance Analysis

Table I presents the performance comparison of individual classifiers and the ensemble approach. The ensemble method achieves 98.5% accuracy, outperforming individual classifiers by 2-4 percentage points. Precision reaches 98.2%, indicating minimal false positives, while recall of 98.8% demonstrates strong detection of actual phishing URLs. The F1-score of 98.5% confirms balanced performance across both metrics. AUC-ROC of 0.992 indicates excellent discriminative ability.

Model	Accuracy (%)	Precision (%)	Recall (%)
Random Forest	96.2	95.8	96.6
SVM	94.8	96.1	93.2
Neural Network	95.5	94.9	96.2
Ensemble	**98.5**	**98.2**	**98.8**

TABLE I PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503037|

Comparative analysis with existing methods demonstrates significant improvements. Traditional blacklist-based approaches achieve only 65-70% accuracy due to their reactive nature. Heuristic-based methods reach 80-85% accuracy but suffer from high false positive rates. Recent machine learning approaches report 92-95% accuracy. Our ensemble method achieves 98.5% accuracy, representing a 3-6 percentage point improvement over state-of-the-art techniques.

D. Real-time Performance

Processing time is critical for practical deployment. Feature extraction requires an average of 120 milliseconds per URL, with host-based features contributing the majority of this time due to DNS lookups and WHOIS queries. Classification takes approximately 15 milliseconds for the ensemble prediction. Total processing time averages 135 milliseconds per URL, meeting real-time requirements for browser integration. Optimizations including feature caching and parallel processing can further reduce latency to below 80 milliseconds.

V. DISCUSSION

The experimental results validate the effectiveness of our ensemble approach for phishing URL detection. The high accuracy and balanced precision-recall metrics demonstrate robust performance suitable for production deployment. Feature importance analysis reveals that lexical features provide initial discrimination, while host-based features offer crucial validation. The ensemble method effectively combines these complementary signals to achieve superior detection rates.

Several factors contribute to the system's strong performance. First, the comprehensive feature set captures multiple dimensions of URL characteristics, making it difficult for attackers to evade detection. Second, the ensemble approach leverages diverse classification strategies, reducing vulnerability to specific attack patterns. Third, regular model updates using recent phishing examples maintain effectiveness against evolving threats.

Challenges remain in several areas. Zero-day phishing attacks using newly registered domains of legitimate organizations pose detection difficulties. Sophisticated attackers employing URL shortening services and compromised legitimate websites require additional analysis layers. Mobile platform integration presents unique challenges due to resource constraints. Future work will address these limitations through enhanced feature engineering, active learning mechanisms, and optimized model architectures.

The system's practical deployment potential extends beyond web browsers to email clients, mobile applications, and network security appliances. Integration with existing security infrastructure can provide defense-in-depth protection. Privacy-preserving techniques such as federated learning could enable collaborative model improvement without sharing sensitive URL data. Cloud-based deployment would offer centralized updates and scalable processing capacity.

VI. CONCLUSION

This paper presents a smart phishing URL detection system utilizing ensemble machine learning techniques and comprehensive feature extraction. The proposed approach achieves 98.5% accuracy with balanced precision and recall, significantly outperforming traditional and contemporary detection methods. The system demonstrates practical viability through real-time processing capabilities and robust performance across diverse phishing scenarios. The ensemble methodology combining Random Forest, Support Vector Machine, and Neural Network classifiers effectively leverages complementary strengths to achieve superior detection rates. Comprehensive feature engineering capturing lexical, host-based, and content-based characteristics enables robust discrimination between legitimate and phishing URLs. Experimental validation using 50,000 URLs confirms the system's effectiveness and generalization capability.

Future research directions include integration of advanced deep learning architectures, investigation of adversarial robustness, development of explainable AI mechanisms for security analysts, and exploration of federated learning for privacy-preserving collaborative detection. Enhancement of real-time processing through model optimization and hardware acceleration will improve deployment feasibility. Continuous adaptation mechanisms will maintain effectiveness against evolving phishing tactics.

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503037|

VII. ACKNOWLEDGMENT

The authors would like to thank the Technology Institute for providing computational resources and research support. We acknowledge PhishTank and OpenPhish for providing phishing URL datasets that enabled this research. Special thanks to the Department of Computer Science faculty members for their valuable feedback and guidance throughout this project.

REFERENCES

[1] A. Aleroud and L. Zhou, "Phishing Environments, Techniques, and Countermeasures: A Survey", Computers & Security, vol. 68, pp. 160-196, 2017.
[2] R. M. Mohammad, F. Thabtah and L. McCluskey, "Predicting Phishing Websites Based on Self-Structuring Neural Network", Neural Computing and Applications, vol. 25, no. 2, pp. 443-458, 2014.
[3] S. Marchal, J. François, R. State and T. Engel, "PhishStorm: Detecting Phishing with Streaming Analytics", IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458-471, 2014.
[4] H. Yuan, X. Chen, Y. Li, Z. Yang and W. Liu, "Detecting Phishing Websites and Targets Based on URLs and Webpage Links", Proc. Int. Conf. Advanced Information Networking and Applications, pp. 1109-1114, 2018.
[5] A. K. Jain and B. B. Gupta, "A Machine Learning Based Approach for Phishing Detection Using Hyperlinks Information", Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 5, pp. 2015-2028, 2019.
[6] Y. Tang and S. Chen, "Defending Against Internet Worms: A Signature-Based Approach", Proc. IEEE INFOCOM, vol. 2, pp. 1384-1394, 2005.
[7] W. Zhang, Q. Ding and M. R. Lyu, "A Deep Learning Approach for Detecting Malicious JavaScript Code", Security and Communication Networks, vol. 9, no. 11, pp. 1520-1534, 2016.
[8] E. Buber, B. Diri and O. K. Sahingoz, "Detecting Phishing Attacks from URL by Using NLP Techniques", Proc. Int. Conf. Computer Science and Engineering, pp. 337-342, 2017.
[9] R. S. Rao and A. R. Pais, "Detection of Phishing Websites Using an Efficient Feature-Based Machine Learning Framework", Neural Computing and Applications, vol. 31, no. 8, pp. 3851-3873, 2019.
[10] V. Suganya, "A Study on Phishing Attacks", International Journal of Computer Applications, vol. 139, no. 1, pp. 13-16, 2016.
[11]S. Gupta, A. Singhal and A. Kapoor, "A Literature Survey on Social Engineering Attacks: Phishing Attack", Proc. Int. Conf. Computing, Communication and Automation, pp. 537-540, 2016.
[12]L. Chiew, C. L. Yong, Y. N. Cheah and A. F. A. Abdullah, "Machine Learning Techniques for Phishing Detection Using URL Features: A Review", Cybersecurity, vol. 2, no. 1, pp. 1-24, 2019.
[13]Napoleon, D., et al. "An efficient numerical method for the prediction of clusters using k-means clustering algorithm with bisection method." International Conference on Computing and Communication Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

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