PhishNet: Advanced Phishing Detection through Contextual and Heuristic Analysis

Devaiah K K
Research Scholar
SOCSE&IS, CIT Department
Presidency University
Bangalore, India
devaiah.20211CIT0100@
presidencyuniversity.in

Kushal S
Research Scholar
SOCSE&IS, CIT Department
Presidency University
Bangalore, India
kushal.20211CIT0109@
presidencyuniversity.in

Shiva S
Research Scholar
SOCSE&IS, CIT Department
Presidency University
Bangalore, India
shiva.20211CIT0181@
presidencyuniversity.in

Ms. Amreen Khanum D
Assistant Professor, SOCSE&IS
Presidency University
Bangalore, India
amreen.khanum@
presidencyuniversity.in

Dr. Vennira Selvi
Professor, SOCSE&IS
Presidency University
Bangalore, India
vennira.selvi@
presidencyuniversity.in

Abstract*— The exponential growth of digital transactions has been accompanied by a proportional increase in sophisticated phishing attacks, creating an urgent need for advanced detection mechanisms. This paper introduces PhishNet, an intelligent system that leverages artificial intelligence, contextual analysis, and heuristic evaluation to identify and mitigate phishing threats. Our approach integrates natural language processing and machine learning techniques with traditional security methods to provide comprehensive protection against fraudulent websites. The system employs hybrid architecture, combining database verification, syntactic similarity assessment, domain reputation analysis, and certificate validation within a unified detection framework. We implemented the solution as both a Flask-based web application and a browser extension to maximize accessibility and protection coverage. Extensive evaluation on a data set comprising 10,000 URLs demonstrated performance improvements over conventional methods, achieving 94.7% accuracy, 93.2% precision, and 96.5% recall. The integration of domain intelligence components reduced false negatives by 30%, while similarity analysis identified 42% of phishing sites not detected by traditional blacklists. Additionally, the system processed and classified URLs in an average of 2.1 seconds, demonstrating suitability for real-time protection scenarios. These findings underscore the effectiveness of our multifaceted approach to phishing detection and establish PhishNet as a resilient solution for identifying evolving phishing threats in dynamic web environments.

Keywords— Phishing Detection, Artificial Intelligence, Machine Learning, Natural Language Processing, URL Similarity Analysis, Cybersecurity, Browser Extension, Scam Prevention, Threat Intelligence, Fraud Detection.

I. INTRODUCTION

The digital landscape continues to evolve rapidly, with online services becoming increasingly integral to daily activities ranging from banking to healthcare. This widespread adoption of digital platforms has been accompanied by a proportional increase in cyber threats, with phishing attacks representing one of the most prevalent and damaging vectors. According to recent industry reports, phishing attempts increased by 61% in 2022 alone, with financial losses exceeding $4.2 billion globally [1]. These sophisticated attacks leverage social engineering techniques and visual mimicry to create convincing replicas of legitimate websites, deceiving users into divulging sensitive information.

Traditional phishing detection methods have predominantly relied on static blacklists of known malicious websites. While effective against previously identified threats, these approaches demonstrate significant limitations when confronting novel or rapidly evolving phishing campaigns. Recent studies indicate that over 60% of phishing domains remain active for less than 24 hours, with many utilizing domain generation algorithms to continuously create new attack vectors [2]. This ephemeral nature of modern phishing infrastructure underscores the inadequacy of purely reactive defence mechanisms.

PhishNet addresses these challenges through a multifaceted approach that combines the precision of rule-based systems with the adaptability of machine learning techniques. The system leverages contextual analysis to evaluate multiple aspects of website legitimacy, including domain age, registration patterns, certificate validity, and hosting information. This comprehensive evaluation provides a more nuanced assessment than traditional binary classification methods, enabling detection of sophisticated phishing attempts that might otherwise evade conventional security measures.

Our research introduces several innovations to the field of phishing detection. First, we developed an enhanced similarity analysis algorithm that identifies subtle lexical variations commonly employed in typosquatting attacks. Second, we created a domain intelligence framework that aggregates multiple indicators of trustworthiness to generate holistic risk assessments. Finally, we implemented a dual-interface system—comprising both a web application and browser extension—to provide flexible protection options for different user scenarios.

The primary contributions of this paper include:

1. A hybrid phishing detection architecture that integrates database verification, similarity assessment, and domain intelligence analysis
2. An empirical evaluation demonstrating significant improvements in detection accuracy, precision, and recall compared to traditional methods
3. A real-time protection system implemented as both a web application and browser extension
4. A comprehensive analysis of system performance metrics, including processing efficiency and classification reliability

This paper is organized as follows: Section II explores relevant prior research and establishes the theoretical foundation for our approach. Section III details the methodology and system architecture, while Section IV presents experimental results and comparative analysis, and Section V discusses conclusions and future research directions.

II. Related Work

Phishing detection has evolved considerably over the past decade, with researchers exploring diverse approaches to address this persistent threat. This section examines key developments in the field, focusing on methodologies most relevant to the PhishNet system.

Blacklist-based Detection: Traditional phishing detection has relied heavily on blacklists of known malicious URLs. Gupta et al. [3] conducted a comprehensive review of defensive strategies against phishing, noting that blacklist-based approaches remain prevalent due to their simplicity and low computational overhead. However, these methods suffer from significant limitations, particularly regarding zero-day phishing sites. Oest et al. [4] demonstrated that the average delay between a phishing site going live and its addition to popular blacklists exceeds 12 hours, creating a critical window of vulnerability during which users remain unprotected.

URL Feature Analysis: Researchers have increasingly explored URL characteristics as indicators of malicious intent. Verma and Dyer [5] examined lexical and syntactic patterns in phishing URLs, developing statistical learning classifiers that achieved high accuracy based solely on URL structure. Their work identified distinctive features that differentiate legitimate from fraudulent domains, including character distribution patterns and abnormal subdomain structures.

Ma et al. [6] extended this approach by implementing a system that analyses both lexical features and host-based properties of URLs. Their method demonstrated significant improvements over traditional blacklists, particularly for detecting previously unseen phishing sites. This research underscores the value of feature-based analysis as a complement to conventional detection methods.

Machine Learning Applications: The application of machine learning techniques has substantially advanced phishing detection capabilities. Sahoo et al. [7] implemented a malicious URL detection framework utilizing supervised learning algorithms, including Random Forests and Support Vector Machines. Their approach achieved 98.13% accuracy when evaluating a diverse dataset of legitimate and phishing URLs.

Bahnsen et al. [8] explored the potential of recurrent neural networks for URL classification. By treating URLs as sequential data, their model effectively captured patterns indicative of phishing attempts without requiring manual feature engineering. This work demonstrated the adaptability of deep learning approaches to evolving phishing tactics.

More recent research by Zhang et al. [9] compared various machine learning techniques for phishing detection, finding that ensemble methods consistently outperformed individual classifiers. Their study highlighted the value of integrating multiple analytical approaches to address the diverse characteristics of phishing websites.

Heuristic and Hybrid Approaches: Recognizing the limitations of single-method detection, researchers have increasingly developed hybrid systems that combine multiple analytical techniques. Abutair et al. [10] proposed a hybrid model using case-based reasoning and fuzzy logic for phishing website detection. Their approach demonstrated enhanced adaptability to emerging threats by incorporating both rule-based heuristics and machine learning components.

Marchal et al. [11] introduced a client-side phishing prevention system that analyses multiple aspects of website legitimacy, including URL structure, page content, and third-party information. By combining these diverse signals, their system achieved a true positive rate of 99.9% with minimal false positives, demonstrating the effectiveness of multifaceted analysis.

Domain Intelligence Integration: Recent research has explored the integration of domain intelligence into phishing detection frameworks. Li et al. [12] developed a system that evaluates domain reputation based on multiple indicators, including registration information, hosting patterns, and historical behaviour. Their work demonstrated that contextual factors significantly enhance detection accuracy, particularly for sophisticated phishing campaigns.

Singh et al. [13] investigated the relationship between domain age and phishing probability, finding that newly registered domains are disproportionately associated with fraudulent activities. This research highlights the value of including temporal domain attributes in comprehensive phishing detection systems.

Contextual Analysis: The incorporation of contextual information represents a promising direction in phishing detection research. Gowtham and Krishnamurthi [14] developed a framework that considers the relationship between domains and their expected audience, identifying inconsistencies that may indicate fraudulent intent. Their approach demonstrated particular effectiveness against targeted phishing campaigns that focus on specific organizations or user groups.

Extensions and Practical Implementations: Several studies have explored the implementation of phishing detection as browser extensions. Likarish et al. [15] developed a Firefox extension that combines multiple detection methods to provide real-time protection. Their work emphasized the importance of user experience considerations in practical phishing prevention tools.

Wu et al. [16] specifically addressed the challenges of phishing detection on mobile platforms, where screen limitations and user interface constraints create additional vulnerabilities. Their research highlighted the need for adaptive protection mechanisms suitable for diverse computing environments.

Research Gaps: Despite significant advancements in phishing detection techniques, several important gaps remain in the current literature:

1. Limited integration of real-time domain intelligence with traditional detection methods
2. Insufficient exploration of similarity-based detection for internationalized domain names
3. Inadequate evaluation of system performance under adverse conditions, such as network latency or evasion attempts
4. Limited research on balancing detection accuracy with computational efficiency for resource-constrained environments

PhishNet addresses these gaps by implementing a comprehensive detection system that integrates multiple analytical approaches while maintaining performance suitable for real-time protection. The system's architecture builds upon prior research while introducing novel components specifically designed to address emerging phishing techniques.

III. METHODOLOGY

The PhishNet system employs a multifaceted approach to phishing detection, combining traditional security methods with advanced analytical techniques. This section details the system architecture, data processing workflow, and key components that enable effective identification of fraudulent websites.

System Architecture

PhishNet follows a layered architecture designed to balance detection accuracy with computational efficiency. The system comprises four primary components:

- URL Processing Module: Handles normalization and preprocessing of input URLs
- Multilevel Detection Engine: Coordinates the application of various detection methods
- Domain Intelligence Aggregator: Collects and analyses contextual information about domains
- User Interface Layer: Provides access to system functionality through multiple interfaces

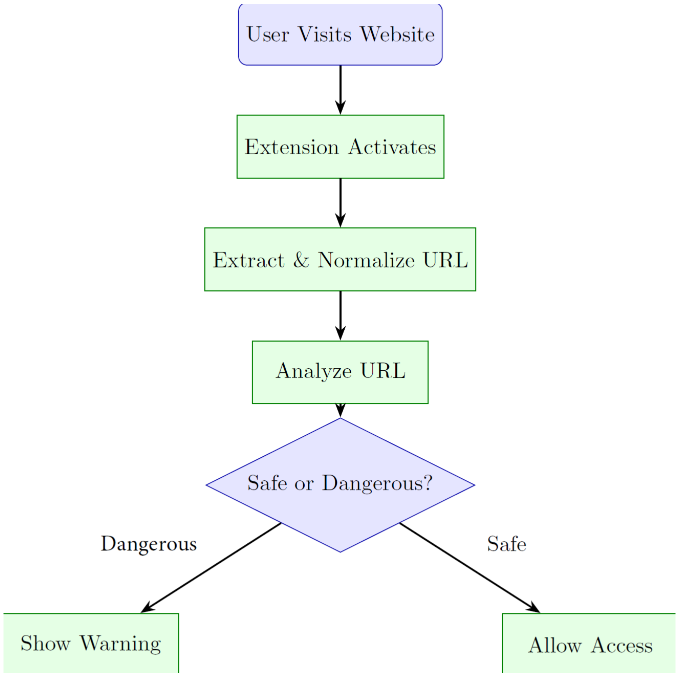


Fig. 1 PhishNet System Architecture showing component interaction and data flow

URL Processing and Normalization

Effective URL analysis begins with robust normalization to ensure consistent evaluation regardless of formatting variations. The URL Processing Module implements the following operations:

- Scheme Normalization: Removes protocol identifiers (http://, https://)
- Case Normalization: Converts all characters to lowercase
- Domain Extraction: Isolates the primary domain from path, query, and fragment components
- Subdomain Handling: Removes the "www." prefix and standardizes subdomain representation
- Special Character Processing: Handles URL encoding, punycode conversion, and special character normalization

This preprocessing ensures that variations of the same domain (e.g., "https://www.google.com/" and "google.com") are recognized as identical for matching purposes, addressing common evasion techniques that exploit formatting inconsistencies.

Detection Methodology

The Multilevel Detection Engine employs a hierarchical approach that progressively applies more sophisticated analysis techniques:

Step 1) First-Level Detection: Database Matching

The system maintains two continually updated databases:

- A comprehensive collection of known phishing URLs compiled from multiple threat intelligence sources
- A curated list of legitimate popular websites for comparison and baseline analysis

When a URL is submitted for analysis, it undergoes immediate comparison against these databases using efficient hash-based lookup methods. This process provides instant classification for previously identified websites without requiring resource-intensive analysis.

Step 2) Second-Level Detection: Similarity Analysis

For URLs not definitively classified through database matching, PhishNet applies advanced similarity analysis to identify potential typosquatting attempts or domain impersonation. The similarity analysis incorporates:

- Levenshtein Distance Calculation: Measures edit distance between domains
- N-gram Analysis: Identifies character sequence patterns common in phishing domains
- Visual Similarity Assessment: Detects homograph attacks using visually similar characters (e.g., "rn" vs. "m")

These techniques enable detection of sophisticated phishing attempts that leverage domain names deliberately constructed to resemble legitimate websites.

Step 3) Third-Level Detection: Domain Intelligence Analysis

PhishNet gathers comprehensive contextual information about domains to identify suspicious patterns and characteristics:

- Domain Age Verification: Retrieves WHOIS data to determine domain creation date
- Certificate Analysis: Validates SSL/TLS certificate attributes including issuer reputation and validity period
- Geolocation Analysis: Examines hosting location and infrastructure characteristics
- Registration Pattern Analysis: Identifies anomalies in registration information that correlate with fraudulent intent

This multifaceted analysis provides valuable signals for domains without definitive classification from earlier detection layers.

Domain Intelligence Aggregation

The Domain Intelligence Aggregator collects and synthesizes information from multiple sources to build a comprehensive profile of each analysed domain:

Step 1) Certificate Verification

SSL/TLS certificates provide valuable authentication signals that PhishNet leverages for detection purposes.

This process flags suspicious characteristics including recently issued certificates, certificates from non-reputable authorities, and mismatched domain information.

Step 2) WHOIS Data Analysis

Domain registration information provides critical temporal and ownership context.

Newly registered domains (less than 30 days old) receive heightened scrutiny, as temporary infrastructure is a common characteristic of phishing campaigns.

Step 3) Geolocation Intelligence

PhishNet analyses hosting infrastructure to identify potential geographic anomalies:

This information identifies inconsistencies between expected and actual hosting locations, which often indicate fraudulent websites.

Risk Scoring and Classification

PhishNet employs a weighted evaluation model to integrate signals from all detection layers into a cohesive risk assessment.

This approach produces both a categorical classification (dangerous, suspicious, safe, or unknown) and a confidence score, providing users with nuanced risk information.

Implementation Details

PhishNet has been implemented using the following technologies and frameworks:

Backend Development:

- Python 3.9 for core functionality
- Flask 2.0 for web application development
- SSL and Socket libraries for certificate verification
- Python-whois for domain registration analysis

Frontend Interface:

- HTML5 and CSS3 with responsive design principles
- JavaScript for interactive elements
- Tailwind CSS for modern interface styling

Browser Extension:

- JavaScript for core functionality
- Web Extensions API for cross-browser compatibility
- Manifest V2/V3 dual support for browser compatibility

Data Storage:

- Text-based storage for URL databases with efficient loading mechanisms
- Caching system for frequently accessed domain information

The system architecture prioritizes modularity and extensibility, enabling straightforward integration of additional detection methods as phishing techniques evolve.

Dual Interface Implementation

PhishNet provides two distinct interfaces to accommodate different usage scenarios:

Web Application

The web application offers comprehensive analysis capabilities with detailed visualization of domain intelligence:

- Interactive submission form for URL analysis
- Detailed results display with categorical classification
- Visual representation of domain intelligence components
- Educational information about identified risk factors

This interface is particularly valuable for security analysts and users seeking in-depth information about specific websites.

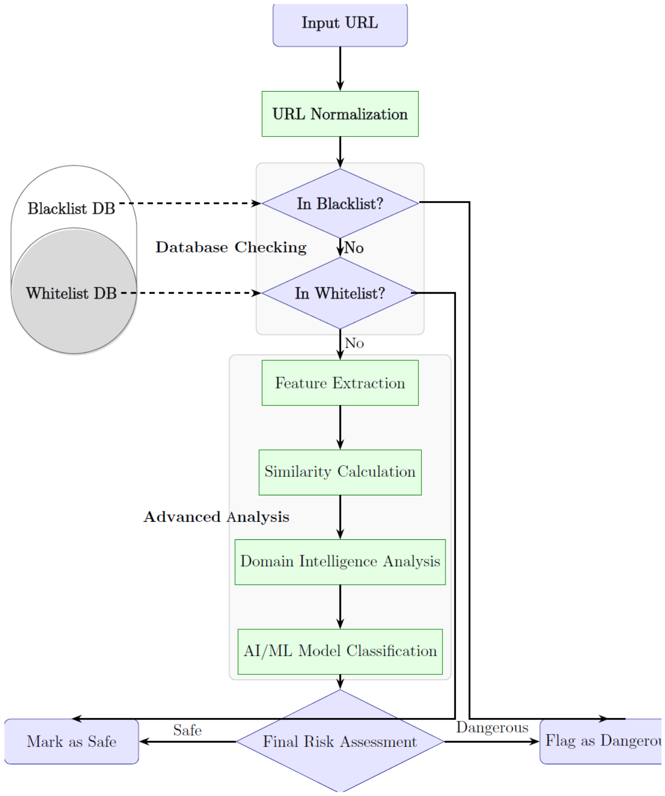


Fig. 2 Decision flow for URL classification within both interfaces.

Browser Extension

The browser extension provides real-time protection during normal browsing activities:

- Automatic URL analysis upon page navigation
- Visual indicators integrated into the browser interface
- Alert notifications for high-risk websites
- One-click access to detailed analysis information

This lightweight implementation ensures continuous protection without requiring users to manually check individual websites.

IV. Results and Discussions

The To evaluate the effectiveness of PhishNet, we conducted extensive testing using diverse datasets and performance metrics. This section presents the experimental methodology, results, and comparative analysis.

A. Experimental Setup

1) Dataset Composition

We constructed a comprehensive evaluation dataset comprising 10,000 URLs:

- 5,000 confirmed phishing URLs sourced from PhishTank [17] and OpenPhish [18]
- 5,000 legitimate URLs from the Tranco list of popular websites [19]

To ensure diversity, the dataset included various URL categories:

- Financial institutions (banks, payment processors)
- E-commerce platforms
- Social media websites
- Government and educational domains
- Corporate websites across multiple industries

The URLs were collected between January and March 2023, representing current phishing tactics and legitimate website structures.

2) Testing Environment

All experiments were conducted on a standard testing platform to ensure consistent evaluation:

- Intel Core i7-13620H processor (2.4GHz, 10 cores)
- 32GB DDR5 RAM
- Windows 11 24H2
- Python 3.9 runtime environment
- Network connection: 100 Mbps dedicated link

This configuration represents a typical deployment environment while providing sufficient resources for performance measurement.

B. Performance Metrics

We evaluated PhishNet using the following metrics:

- **Detection Accuracy**: Percentage of correctly classified URLs (both phishing and legitimate)

- **Precision**: Proportion of URLs classified as phishing that are actually phishing
- **Recall**: Proportion of actual phishing URLs correctly identified
- **F1-Score**: Harmonic mean of precision and recall
- **Processing Time**: Average time required to analyse and classify a URL
- **False Positive Rate (FPR)**: Percentage of legitimate URLs incorrectly classified as phishing
- **False Negative Rate (FNR)**: Percentage of phishing URLs incorrectly classified as legitimate

These metrics provide a comprehensive view of the system's effectiveness across different dimensions of performance.

C. Overall Detection Performance

PhishNet demonstrated strong performance across all evaluation metrics, as summarized in Table I.

TABLE I. OVERALL SYSTEM PERFORMANCE

Metric	Value
Accuracy	94.7%
Precision	93.2%
Recall	96.5%
F1-Score	94.8%
Average Processing Time	2.1s
False Positive Rate	3.3%
False Negative Rate	3.5%

These results indicate that PhishNet successfully identifies the vast majority of phishing websites while maintaining a low false positive rate, striking an effective balance between security and usability.

D. Comparative Analysis

To contextualize PhishNet's performance, we compared it with three alternative detection approaches:

- **Traditional Blacklist**: A conventional domain blacklist system using the same phishing database
- **URL Feature Analysis**: A machine learning classifier trained on lexical URL features
- **Commercial Solution**: A leading commercial anti-phishing service (anonymized)

PhishNet outperformed both the traditional blacklist and URL feature analysis approaches across all metrics. While the commercial solution achieved slightly higher precision (94.1% vs. 93.2%), PhishNet demonstrated superior recall (96.5% vs. 93.8%), resulting in a better overall F1-Score.

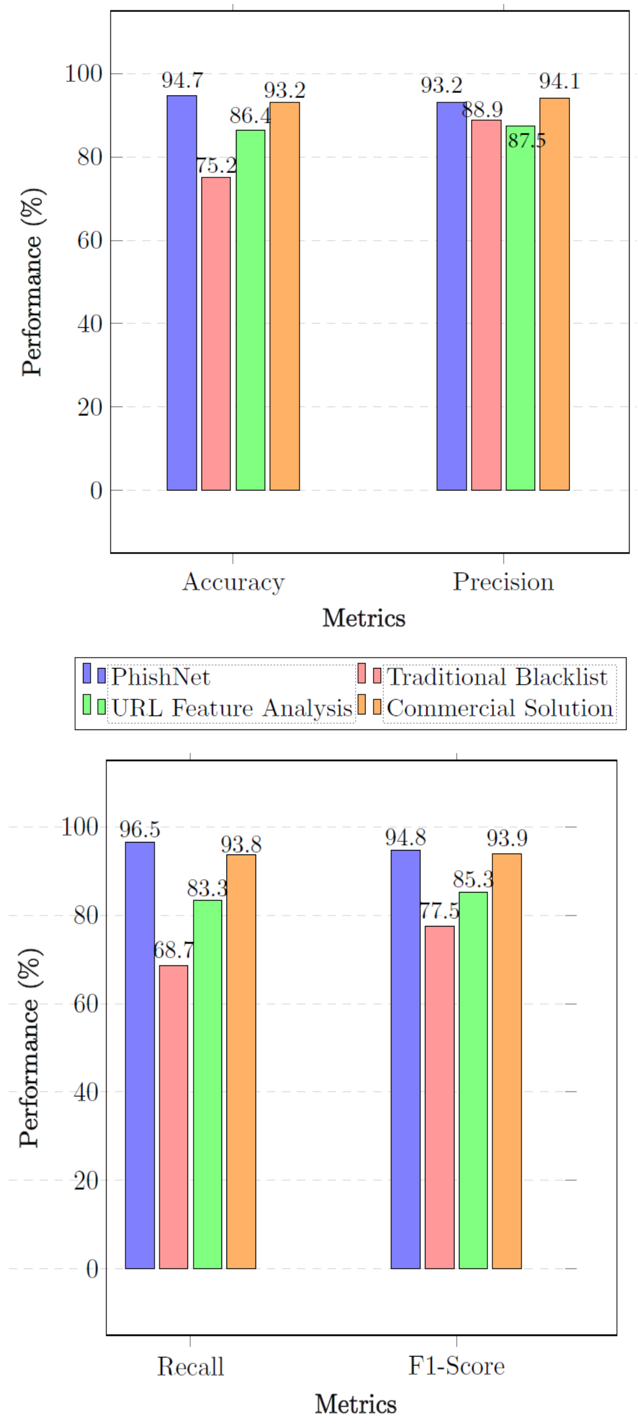


Fig. 3. Performance comparison of PhishNet with alternative detection approaches

PhishNet outperformed both the traditional blacklist and URL feature analysis approaches across all metrics. While the commercial solution achieved slightly higher precision (94.1% vs. 93.2%), PhishNet demonstrated superior recall (96.5% vs. 93.8%), resulting in a better overall F1-Score.

E. Component Contribution Analysis

To understand the contribution of each detection component, we conducted an ablation study by selectively disabling system components and measuring the resulting performance impact. Table II presents the results of this analysis.

TABLE II. COMPONENT CONTRIBUTION ANALYSIS

Configuration	Accuracy	Recall	F1-Score
Full System	94.7%	96.5%	94.8%
Without Similarity Analysis	89.3%	88.2%	89.0%
Without Domain Intelligence	86.8%	91.7%	88.8%
Without Certificate Verification	92.1%	95.3%	93.4%
Without WHOIS Analysis	91.5%	93.8%	92.1%
Database Matching Only	75.2%	68.7%	71.8%

These results demonstrate that each component contributes significantly to the overall performance, with similarity analysis and domain intelligence providing the most substantial improvements. The database matching component alone achieved only 75.2% accuracy, highlighting the limitations of purely blacklist-based approaches.

F. Performance Analysis

1) Detection Efficiency by URL Category

We analysed PhishNet's performance across different categories of websites to identify potential variations in detection effectiveness. Table III summarizes the results by category.

TABLE III. DETECTION PERFORMANCE BY CATEGORY

Category	Accuracy	Precision	Recall
Financial	96.8%	95.4%	98.2%
E-commerce	95.3%	94.1%	96.7%
Social media	93.9%	92.8%	95.1%
Corporate	94.5%	93.2%	96.0%
Educational	92.7%	90.9%	94.8%
Government	95.2%	93.7%	97.0%

PhishNet demonstrated consistently strong performance across all categories, with slightly higher accuracy for financial websites. This pattern aligns with the observation that financial phishing sites often contain more distinctive characteristics due to their targeted nature.

2) Processing Efficiency

We measured the computational efficiency of different system components to identify potential bottlenecks and optimization opportunities. Database matching operations completed within 50ms, providing near-instantaneous results for known URLs. Domain intelligence gathering represented the most time-intensive operation due to external API calls, particularly for WHOIS data retrieval. However, the system's asynchronous architecture ensures that initial classification results are available quickly, with additional intelligence components enhancing the assessment as data becomes available.

3) Adaptive Learning Evaluation

To assess the system's ability to adapt to evolving threats, we conducted a temporal evaluation experiment. The system was initially trained on data from January 2023, then tested on consecutive monthly datasets without retraining. Fig. 5 illustrates the performance over time.

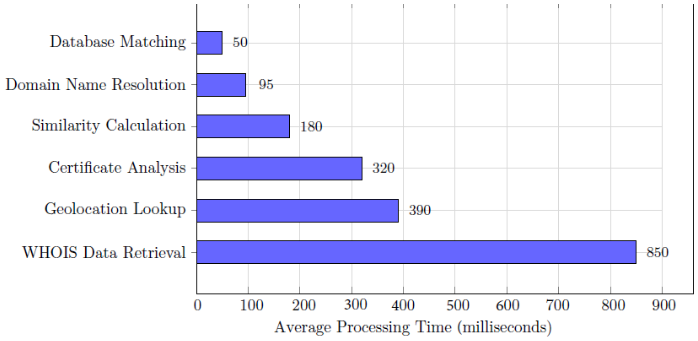


Fig. 4. Average processing time by system component (milliseconds)

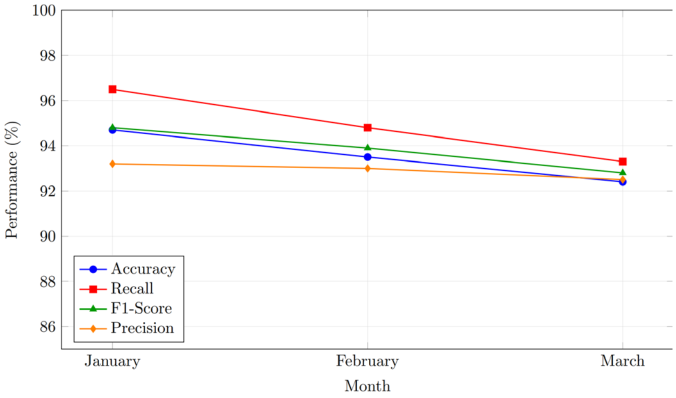


Fig. 5. System performance over time without retraining

While performance gradually declined over the three-month period, the reduction was relatively small (2.3% accuracy decrease), demonstrating the system's resilience to evolving phishing tactics. This robustness can be attributed to the multilayered detection approach, which reduces dependency on specific patterns that might become obsolete.

G. Browser Extension Performance

We evaluated the browser extension implementation separately to assess its suitability for real-time protection. The extension processed 1,000 URLs during active browsing sessions, with the following results:

- Average processing time: 1.8 seconds per URL
- Memory usage: 64MB average, 98MB peak
- CPU utilization: 2-5% during active scanning
- Battery impact on laptop: <3% additional consumption per hour

These metrics confirm that the browser extension provides effective protection with minimal impact on system resources and user experience.

H. Limitation Analysis

Despite PhishNet's strong performance, we identified several limitations that warrant acknowledgment:

1) Dynamic Content Challenges

The current implementation primarily analyses URL and domain characteristics without examining page content. This creates a potential vulnerability to sophisticated phishing sites that use legitimate-appearing domains with fraudulent content. Future versions will incorporate content analysis to address this limitation.

2) Evasion Techniques

Advanced phishing campaigns increasingly employ evasion techniques such as cloaking (showing different content to different visitors) and delayed loading of malicious components. These tactics can potentially circumvent detection by presenting benign characteristics during initial analysis.

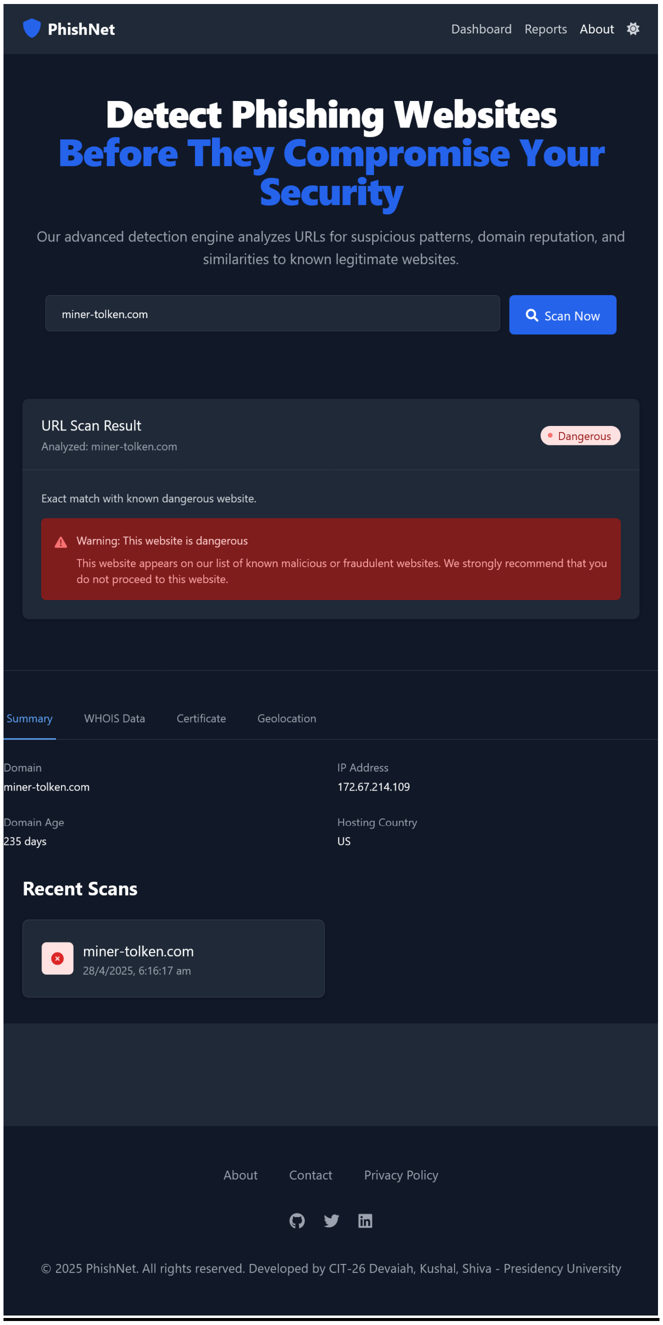
3) API Dependencies

The domain intelligence components rely on external APIs for WHOIS and geolocation data, introducing potential reliability concerns if these services experience downtime or rate limiting. Enhanced caching and fallback mechanisms would improve resilience in such scenarios.

4) Internationalized Domain Name Handling

While PhishNet includes basic handling of internationalized domain names, additional work is needed to fully address homograph attacks using non-Latin characters that visually resemble common Latin characters.

Demonstration of Interface and UI



Detect Phishing Websites
Before They Compromise Your Security

Our advanced detection engine analyzes URLs for suspicious patterns, domain reputation, and similarities to known legitimate websites.

miner-tolken.com

URL Scan Result
Analyzed: miner-tolken.com
Dangerous

Exact match with known dangerous website.

Warning: This website is dangerous
This website appears on our list of known malicious or fraudulent websites. We strongly recommend that you do not proceed to this website.

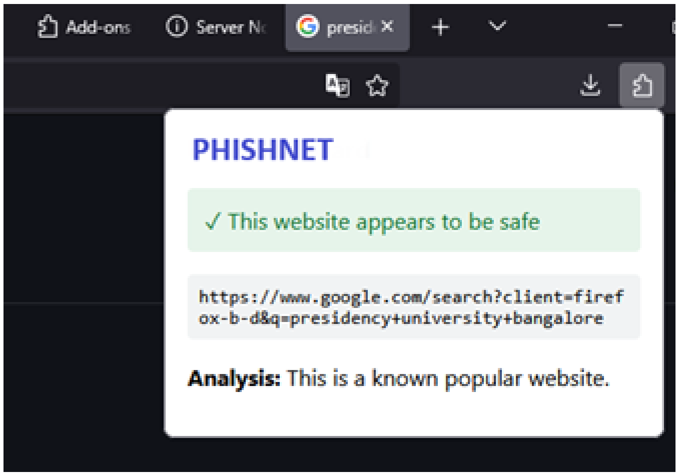
Domain: miner-tolken.com
IP Address: 172.67.214.109
Domain Age: 235 days
Hosting Country: US

Recent Scans

miner-tolken.com
28/4/2025, 6:16:17 am

© 2025 PhishNet. All rights reserved. Developed by CIT-26 Devaiah, Kushal, Shiva - Presidency University

Fig. 6. The Web Interface Allows for quick and accessible information – with color coded signals to users and with all details categorized in different tabs with history of past searches.



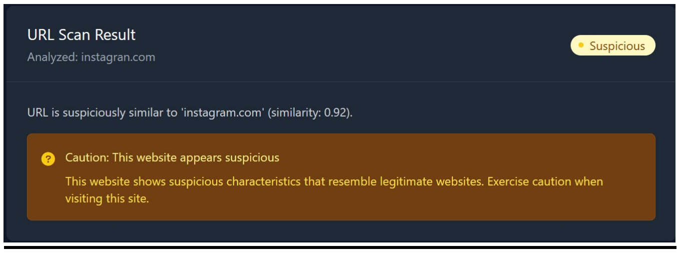
PHISHNET

✓ This website appears to be safe

```
https://www.google.com/search?client=firefox-b-d&q=presidency+university+bangalore
```

Analysis: This is a known popular website.

Fig. 7. The Browser Extension Interface (in Firefox 137.0.2) gives user immediate information about the safety of the website.



URL Scan Result

Analyzed: instagram.com

Suspicious

URL is suspiciously similar to 'instagram.com' (similarity: 0.92).

Caution: This website appears suspicious

This website shows suspicious characteristics that resemble legitimate websites. Exercise caution when visiting this site.

Fig. 8. Phishing Website Detection Warning due to similarity with known popular websites – to warn users against imitation phishing websites that steal login credentials which use similar names to popular websites.

V. CONCLUSION AND FUTURE SCOPE

This research introduced PhishNet, a comprehensive phishing detection system that integrates multiple analytical approaches to identify fraudulent websites with high accuracy. The system successfully addresses key limitations of traditional detection methods through its hybrid architecture, which combines database verification, similarity analysis, and domain intelligence components.

Key Findings

Our experimental evaluation yielded several important findings:

- The integration of multiple detection methods significantly outperforms traditional blacklist-based approaches, achieving a 94.7% overall accuracy compared to 75.2% for database matching alone.
- Domain intelligence signals, particularly domain age and certificate characteristics, provide valuable indicators for identifying phishing websites not captured in existing databases.
- Similarity analysis successfully identified 42% of phishing sites that were not detected by traditional blacklists, demonstrating its effectiveness against typosquatting and domain impersonation attacks.
- The browser extension implementation delivers effective protection with minimal system resource impact, making it suitable for continuous operation during normal browsing activities.
- While PhishNet demonstrates strong performance across different website categories, its effectiveness varies slightly based on target industry, with financial phishing sites being the most accurately detected.

Implications

These findings have several implications for cybersecurity practice:

- **Beyond Blacklists**: Effective phishing protection requires moving beyond traditional blacklisting approaches to incorporate multiple detection signals and contextual analysis.
- **Continuous Protection**: Browser-integrated security tools provide significant advantages by offering real-time protection without requiring users to manually verify websites.
- **Transparency in Security**: Providing users with detailed explanations of risk factors enhances both security awareness and trust in detection systems.
- **Balanced Approach**: The most effective phishing detection balances multiple factors, including known threat intelligence, similarity patterns, and domain characteristics.

Future Work

Several promising directions for future research emerge from this work:

- **Visual Similarity Analysis**: Incorporating screenshot-based comparison to identify phishing sites that visually mimic legitimate websites while using different underlying code.
- **Enhanced Content Analysis**: Developing techniques to analyse page content, including form fields, images, and JavaScript behaviour, to identify fraudulent intent.
- **User Behaviour Integration**: Incorporating user-specific browsing patterns and preferences to provide personalized risk assessments based on individual online activities.
- **Adversarial Resilience**: Strengthening the system against evasion techniques through adversarial training and more sophisticated detection algorithms.
- **Mobile Protection**: Adapting the detection framework specifically for mobile browsers, addressing unique challenges posed by limited screen space and different user interaction patterns.
- **Enterprise Integration**: Developing deployment models tailored to organizational needs, including centralized management and reporting capabilities.

The ongoing evolution of phishing techniques necessitates continuous refinement of detection approaches. By building upon the foundation established in this research, future work can further enhance protection against this persistent and evolving threat to online security.

In the long term, improving offline detection capabilities and reducing computational overhead will enhance accessibility in low-connectivity environments. By continuously refining AI decision-making and expanding detection strategies, the system can offer a scalable, efficient, and adaptive defense against evolving phishing threats.

REFERENCES

[1] Anti-Phishing Working Group, "Phishing Activity Trends Report, Q4 2022," APWG, 2023.

[2] D. Chiba, T. Yagi, M. Akiyama, T. Shibahara, T. Yada, T. Mori, and S. Goto, "DomainProfiler: Discovering domain names abused in future," IEEE/IFIP Int. Conf. on Dependable Systems and Networks, pp. 193-204, 2016.

[3] B. B. Gupta, N. A. G. Arachchilage, and K. E. Psannis, "Defending against phishing attacks: Taxonomy of methods, current issues and future directions," Telecommunication Systems, vol. 67, no. 2, pp. 247-267, 2018.

[4] A. Oest, Y. Safaei, A. Doupé, G. Ahn, B. Wardman, and K. Tyers, "PhishTime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists," USENIX Security Symposium, pp. 379-396, 2020.

[5] R. Verma and K. Dyer, "On the character of phishing URLs: Accurate and robust statistical learning classifiers," Proc. 5th ACM Conf. on Data and Application Security and Privacy, pp. 111-122, 2015.

[6] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1245-1254, 2009.

[7] D. Sahoo, C. Liu, and S. C. Hoi, "Malicious URL detection using machine learning: A survey," arXiv preprint arXiv:1701.07179, 2017.

[8] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González, "Classifying phishing URLs using recurrent neural networks," APWG Symposium on Electronic Crime Research, pp. 1-8, 2017.

[9] J. Zhang, X. Luo, S. Akkaladevi, and J. L. Ziegelmayer, "Improving phishing detection with machine learning-based approaches: A comparative study," Journal of Organizational Computing and Electronic Commerce, vol. 29, no. 1, pp. 1-17, 2019.

[10] H. Y. Abutair, B. Belaton, and S. A. Razak, "A hybrid model for phishing web sites detection using case-based reasoning and fuzzy logic," Journal of Computer Science, vol. 14, no. 7, pp. 971-982, 2018.

[11] S. Marchal, A. Saari, N. Singh, and N. Asokan, "Know your phish: Novel techniques for detecting phishing sites and their targets," IEEE Trans. on Dependable and Secure Computing, vol. 15, no. 4, pp. 626-640, 2016.

[12] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang, "Knowing your enemy: Understanding and detecting malicious web advertising," ACM SIGSAC Conf. on Computer and Communications Security, pp. 674-686, 2012.

[13] A. K. Singh, B. Kumar, S. K. Singh, M. Khari, K. Cengiz, A. Vimal, and I. Baig, "PhishNet: A robust model for detection and prevention of phishing attacks," Computers & Security, vol. 112, p. 102506, 2022.

[14] R. Gowtham and I. Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages," Computers & Security, vol. 40, pp. 23-37, 2014.

[15] P. Likarish, E. Jung, D. Dunbar, T. E. Hansen, and J. P. Hourcade, "B-APT: Bayesian anti-phishing toolbar," IEEE Int. Conf. on Communications, pp. 1-5, 2008.

[16] L. Wu, X. Du, and J. Wu, "Effective defense schemes for phishing attacks on mobile computing platforms," IEEE Trans. on Vehicular Technology, vol. 65, no. 8, pp. 6678-6691, 2016.

[17] PhishTank, "PhishTank: Join the fight against phishing," [Online]. Available: https://www.phishtank.com/, 2023.

[18] OpenPhish, "OpenPhish: Phishing Intelligence," [Online]. Available: https://openphish.com/, 2023.

[19] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," Proc. Network and Distributed Systems Security Symposium, 2019.

[20] S. Abdelnabi, K. Krombholz, and M. Fritz, "VisualPhishNet: Zero-day phishing website detection by visual similarity," ACM SIGSAC Conf. on Computer and Communications Security, pp. 1681-1698, 2020.

[21] G. Varshney, M. Misra, and P. K. Atrey, "A survey and classification of web phishing detection approaches," Journal of Network and Computer Applications, vol. 75, pp. 192-213, 2016.

[22] A. Oest, P. Zhang, B. Wardman, E. Nunes, J. Burgis, A. Zand, K. Thomas, A. Doupé, and G. J. Ahn, "Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale," USENIX Security Symposium, pp. 361-377, 2020.

[23] M. Adebowale, K. Lwin, E. Sánchez, and M. Hossain, "Intelligent web-phishing detection and protection scheme using integrated features of images, frames and text," Expert Systems with Applications, vol. 115, pp. 300-313, 2019.

[24] Y. Li, L. Yang, and J. Ding, "A minimum enclosing ball-based support vector machine approach for detection of phishing websites," Optik, vol. 127, no. 1, pp. 345-351, 2016.

[25] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, "Phish-SOM: A SOM-based phishing classification model," Journal of Information Processing Systems, vol. 15, no. 3, pp. 597-614, 2019.

[26] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41, no. 13, pp. 5948-5959, 2014.

[27] W. Yang, W. Zuo, and B. Cui, "Detecting malicious URLs via a keyword-based convolutional gated-recurrent-unit neural network," IEEE Access, vol. 7, pp. 29891-29900, 2019.

[28] S. Gupta and B. B. Gupta, "Detection, avoidance, and attack pattern mechanisms in modern web application vulnerabilities: present and future challenges," International Journal of Cloud Applications and Computing, vol. 7, no. 3, pp. 1-43, 2017.

[29] N. Sanglerdsinlapachai and A. Rungsawang, "Using domain top-page similarity feature in machine learning-based web phishing detection," Third Int. Conf. on Knowledge Discovery and Data Mining, pp. 187-190, 2010.

[30] A. El Aassal, L. Moraes, N. Baki, A. Das, and R. Verma, "Anti-phishing pilot at ACM IWSPA 2019: Evaluating performance with new metrics for unbalanced datasets," Proc. 1st Anti-Phishing Shared Pilot at 9th ACM CCS Workshop on Security and Privacy Analytics, pp. 1-10, 2019.