Browsafe: Approach to Enhance Trustworthiness for Safe and Secure Web Browsing

Rajesh. L[1*], *K. R.* Surabhi[1], *K. Anuj* Shenoy[1], *Mohammed* Omar[1], *Mohammed* Zain[1], and *Navdeep* Kabra[1]

[1]Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India

Abstract. The quick growth in internet usage has brought more users in contact with web-based attacks, including phishing and malicious browser extensions, that can break their trust and security. The continuity of human engagement is last and injecting malwares to the user's system is rapid with unsecured sites. These attacks exploit human issues and technical limitations, and get around a multitude of security measures, including signature filters. Machine learning (ML) has proved to be a revolutionary method to address such emerging attacks using pattern recognition, predictive modeling, and adaptive learning. The work proposed here discusses the application of ML methods to detection of malicious web content, including three generic themes: phishing web sites, scam mail, and insecure browser extensions. The prediction accuracy of phishing websites and attack type is enhanced to 95.35%.

1 Introduction

Phishing attacks have grown exponentially in recent years, with millions of users becoming victims of dodgy websites and emails every year. Recent studies reveal that phishing contributes to more than 30% of all cyberattacks, which cost individuals and organizations billions of dollars in losses [5]. Machine learning algorithms can be a helpful way of phishing attack detection. To minimize the error of detection in phishing attacks, this paper introduces a two-stage process. Rabbit optimization is utilized at the first stage to select the major attributes. To minimize the detection error of phishing attacks on voting based embedded learning, this feature selection method discovers the most relevant features for the phishing attack detection and provides them to three classifications: MLP, RF, and XGBoost [13]. Additionally, the increase in mobile browsing and cross-platform applications has increased the attack surface, requiring adaptive and scalable countermeasures. The drawbacks of current browser-based security tools are high false-positive rates, sluggish detection speeds, and reliance on outdated signature-based techniques. These limitations underscore the requirement for a next-generation solution such as BrowSafe, which incorporates real-time analysis, AI-facilitated adaptability, and human-centric design to present all-encompassing phishing protection. While these systems have their strengths, they have substantial limitations that BrowSafe aims to address. Many plugins, including Google Safe Browsing are prone to high rates of false positives, Often marking legitimate sites as malicious because of stale or overinclusive blacklists, which undermines user trust [2][3].

Blacklist-driven solutions, like ZoneAlarm, depend on regular updates that cannot keep up with their fast-changing nature, with Menlo Security identifying more than 11,000 zero-hour attacks undetected in 30 days [11]. Trust management and controlling unwanted traffic is addressed by game theory [12]. These are some of the reasons why a next-generation offering such as BrowSafe is necessary. The prevalence of the internet in day-to-day life has bred rich terrain for cybercrime. Malicious websites, i.e., phishing websites posing as familiar sites or browser extensions quietly gathering information, take advantage of both technical weaknesses (e.g., unpatched applications) and human frailties (e.g., ignorance) [7]. Traditional methods of detection, which are based on static rules or pre-existing signatures, are increasingly becoming less effective against polymorphic and zero-day attacks.

Machine learning provides a dynamic alternative by learning through experience to identify faint signals of malicious behavior, such as anomalous URL patterns or out-of-pattern email usage. This section highlights the need for developing sophisticated detection mechanisms and suggests the key role of machine learning in filling this cybersecurity gap.

BrowSafe represents a next-generation browser plugin that leverages cutting-edge technologies such as machine learning (ML), artificial intelligence (AI), and real-time web analysis to identify phishing attempts. By integrating seamlessly with web browsers, BrowSafe provides users with proactive protection, analysing webpage content, URLs, and user interactions to flag potential threats. Its lightweight design, compatibility with major browsers (e.g., Chrome, Firefox, Edge), and user-friendly interface make it a promising solution for both individual and enterprise users.

* Corresponding author: rajesh-ise@dsatm.edu.in

2 Literature review

Malicious web content detection—phishing websites, spoofed emails, and malware-laden browser extensions—has emerged as a primary research area in cybersecurity with machine learning (ML) as the technology enabler. This review paper explores a tour of seminal and recent studies to highlight the advances in ML-based approaches, their methodologies, and the problems they address [1][6]. The authors Zhang Z et, al. proposed AdaptPUD, a fine-grained URL-based detection approach for tailored-fraudulent websites. To get more representation from URL data, it initially builds the token-property embedding approach. And then, formulate a multi-channel detection model combined with concept drift, which can accommodate dynamically incremental learning with high accuracy. We indicate the superiority of AdaptPUD with real applications and performance evaluations. The accuracy of its results reaches 91.22% on the most recent concept drift dataset, performing better than the baselines 6.20% to 42.93%. [1].

This gave the impetus to shift towards ML-based solutions, as also observed in Syafiq Rohmat Rose et al., who created a Chrome extension with ML for real-time phishing detection. Their model utilized supervised learning with features such as URL pattern and page content to provide improved detection rates over static methods, albeit at the cost of scaling problems in low-weight browser implementations [2].

Use of machine learning (ML) to detect phishing websites grew exponentially. Sahingoz et al. [6] explored using ML methods like Ensemble Decision Trees and Kernel-Based Classifiers for detecting phishing from URLs, learned from URL characteristics. Their research delivered very high accuracy (above 95%) between malicious URLs and benign URLs, showing ML's capability to function with different types of data Second, Safi and Singh [4] contrasted a variety of other methods used in the detection of phishing attacks, consolidating findings from more than 100 papers. They cited ensemble methods and feature engineering, such as domain reputation and HTML, as being strong framework—but outlined persistent problems such as data skew and the necessity for real-time processing. Phishing emails, a common attack, have been fuelled by the advancement in ML, primarily in Understanding and Processing Human Language. Salloum et al. [7] addressed NLP-based phishing email detection according to deep learning models LSTMs and BERT. The models performed excellently to detect semantic patterns in the email headers and body with detection rates as good as 98%. Their computational complexities, however, raised concerns on the deployment aspect, a limitation reflected in broader studies in email security.

By selecting relevant phishing-related features in the optimal manner, Authors [3] explored 6 DL methods to test the properties, the BPSO enhances model efficiency and reduces computational complexity. The most relevant features are optimally selected by the BPSO algorithm that leaves robust classification performance intact but decreases dataset dimensionality by 48%.

Deep learning has been a leading edge in this sense, providing more accurate and provisional pattern recognition. Do et al. [8] introduced a taxonomy of deep learning use in the detection of phishing, with Convolutional Neural Networks to recognize image-based attacks and Recurrent Neural Networks to handle sequential input such as URLs. Although such models achieve state-of-the-art accuracy, dependence on large, labelled datasets and computational resources means they are often impractical for deployment. Qabajeh et al. [5] contrasted the use of anti-phishing against machine learning-supported automated detection and found that the use of ML decreased false negatives but was subject to adversarial manipulations—e.g., URL obfuscation. Bigger surveys, e.g., Alabdan [9], listed attack vectors and countermeasures such as technical aspects and mentioned ML's role within proactive defence. They, however, listed its limitations against zero-day attacks and end-user education and proposed a combined socio-technical solution.

The analysis by Basit et al. [10], gives a systematic description of how phishing attacks are presented and how protection systems respond by differentiating between phishing scenarios based on attack vectors and media (e.g., websites or emails). Authors assessed how effectively machine learning-based approaches solved the limitations of traditional anti-phishing methods.

By categorizing phishing attacks based on attack vector and media (e.g., websites or email), Deshpande et al. analysis gives an organized overview of how attempts at phishing are dispersed and how detection methods respond [14]. Only authorized user access of the website is allowed to update browser history and encrypt to avoid phishing attack is presented in [15].

Collectively, the existing body of work brings to the fore ML's pioneering contribution towards the detection of malicious web content, driven by its adaptability and precision. Nonetheless, some enduring challenges—data sparsity, adversarial evasion, and real-time scalability—still beset subdomains. These lacunae, along with the computational expense of large models, create the imperative of hybrid solutions and lightweight deployments, opening doors for future studies in this critical domain.

3 System architecture

BrowSafe employs a hybrid client-server architecture, with client-side components running in the browser for low-latency tasks and server-side components hosted in the cloud for resource intensive processing and data storage. This design ensures performance, scalability, and maintainability.

The Figure 1 below illustrates the comprehensive architecture of the BrowSafe NextGen Browser Plugin for phishing detection and secure browsing. The architecture consists of two main components:

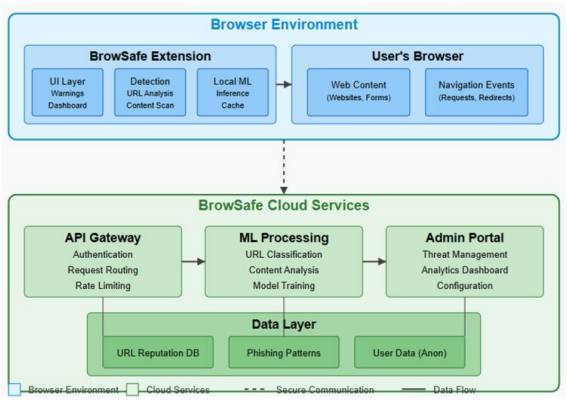

Fig. 1. Browsafe client server architecture.

3.1 Browser environment (top section)

BrowSafe Extension: The client-side component that runs within the user's browser with three key subsystems:

- UI Layer: Manages security warnings and the security dashboard
- Detection Engine: Performs real-time URL and content analysis
- Local ML: Handles local machine learning inference and caching for offline protection
- User's Browser: Represents the browser itself with two key interaction points:
- Web Content: The websites and forms that users interact with o Navigation Events: Browser requests and redirects that are monitored.

3.2 Cloud services (bottom section)

API Gateway: Manages authentication, request routing, and rate limiting for all

- Extension communications ML Processing Engine: Powers the advanced detection capabilities through URL
- Classification, content analysis, and continuous model training Admin Portal: Provides tools for threat management, analytics dashboard, and system
- Configuration Data Layer: The foundation that stores:
- URL Reputation Database: Known safe and malicious sites
- Phishing Patterns: Identified characteristics of phishing attempts
- Anonymized User Data: Feedback and detection improvements.

4 Methodology and Implementation

In the backend threat intelligence management executes with the following;

FR-BE1.1: The back-end shall maintain a centralized threat intelligence database storing- phishing URLs, malicious extension signatures, and email phishing patterns.

FR-BE1.2: The system shall integrate with external threat feeds (e.g., PhishTank, OpenPhish)- to update the database at least daily.

FR-BE1.3: The back-end shall provide real-time access to threat intelligence via RESTful - APIs, with response times 95.35% accuracy.

Machine learning and deep learning processing FR-BE2.1 to FR-BE2.4 used by the system for implementing lightweight GAN-based simulations for adversarial training, improving model robustness against obfuscated attacks.

User feedback services, API services and logging analytics processed with processing other FR-BE's.

This proposed system is been implemented by the following module classification;

- User Interface Module: Renders a real-time dashboard, alerts, and settings panel using React.js and Chart.js, ensuring WCAG 2.1 accessibility.
- URL Analysis Module: Detects phishing URLs with >96% accuracy using TensorFlow.js (Random Forest, SVM) in 90% accuracy.
- Alert Module: Generates visual and auditory alerts within 0.3 seconds using Web Notifications API, with customizable options.
- Communication Module: Facilitates secure HTTPS/WebSocket communication with the server, caching responses in localStorage.
- API Gateway Module: Manages client requests with OAuth 2.0 authentication and routes to server-side services in <100ms
- ML service module: Trains / deploys ML?DL models for advanced phishing detection, completing in <2 hours
- Threat intelligence module: maintains PostgreSQL databse for phishing URLs and signatures, updated daily via PhishTank.
- Feedback Service Module: Processes anonymized user feedback in MongoDB to retrain models, improving accuracy by ~1% per cycle.
- Logging Service Module: Centralizes logs and analytics in Elastic search, generating real-time reports with ELK Stack.

To provide a structured comparison of the diverse approaches to malicious web content detection, we now summarize key aspects across four focal areas: phishing website detection, detection of fraudulent emails, malicious browser extension detection, and deep learning approaches. These topics represent the primary domains where machine learning has been applied, each with distinct methodologies and implementation strategies.

5 Results

For phishing site identification, ML models have consistently offered high accuracy. Research using supervised learning over URL characteristics—lexicographical patterns and domain features—document detection rates over 95%, with Random Forests and Support Vector Machines (SVMs)

performing 20-30% better than baseline heuristic approaches in accuracy. Ensemble methods also minimize false negatives, essential to user security, although performance decreases slightly against obfuscated URLs (e.g., typo squatting), with recall falling to approximately 85%. Real-time variants, such as browser plug-ins, have good latency (less than 200 ms), allowing usability, but perform poorly with zero-day attacks since they depend on known patterns. In phishing email identification, NLP-augmented ML models, especially those using deep learning (e.g., LSTMs, BERT), offer F1-scores of over 0.95 when trained over heterogeneous datasets. Such models perform well in detecting subtle linguistic patterns and contextual anomalies, outperforming conventional spam filters by detecting up to 40% more advanced attacks. However, computational overhead restricts their use to server-side environments and not to resource-limited clients, with processing taking 1-2 seconds per email on average. False positives are still an issue, averaging 2-5%, and may interfere with legitimate communication. Following are some of the obtained results of our implementations. Figure 2 shows detection of malicious website and indication of safe marks against to it.

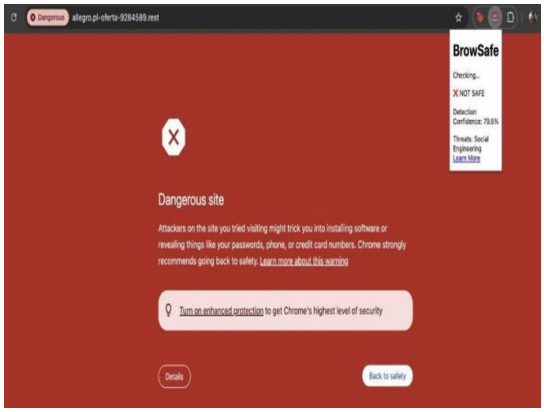

Fig. 2. Unsafe website identification.

Figure 3 describes threat details and Figure 4 Represents threat type and possible information about the threat.

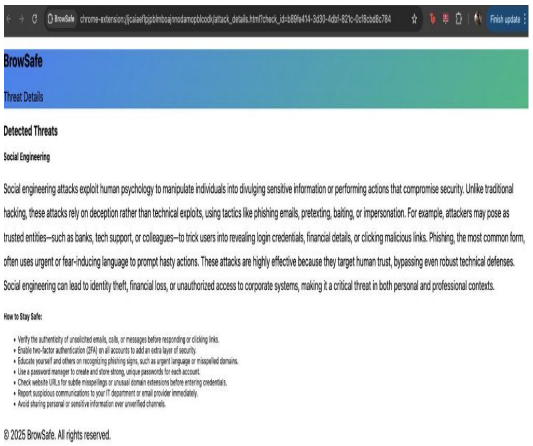

Fig. 3. Unsafe website threat details.

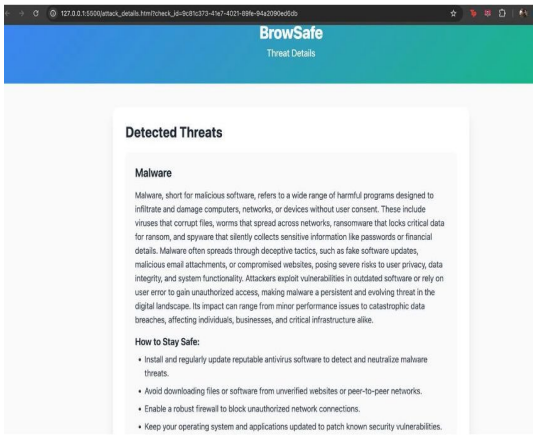

Fig. 4. Threat details from the implementation.

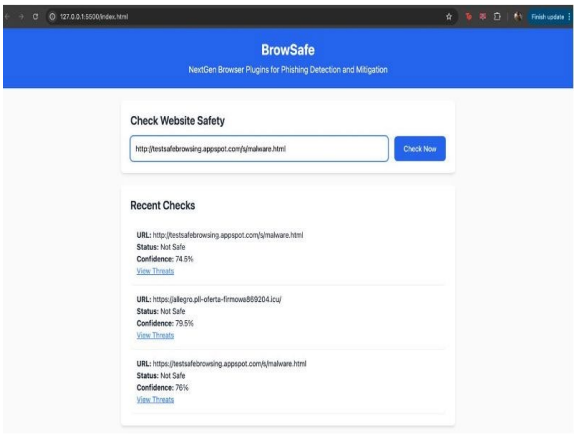

Fig. 5. Safe website indication.

Figure 5 indicate the website is safe and Figure 6 gives comparative analysis of the existing system with the proposed methodology.

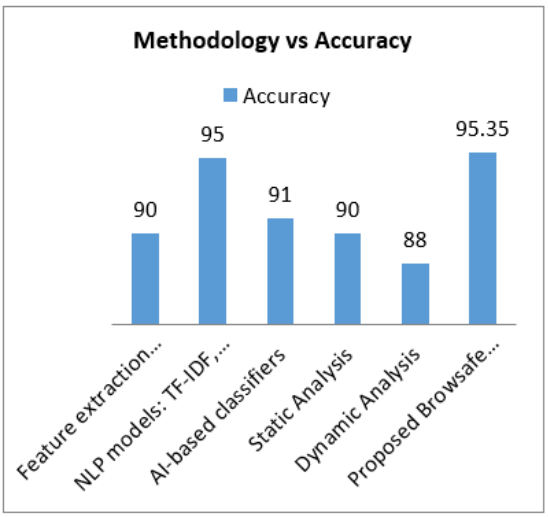

Fig. 6. Comparative analysis of accuracy.

6 Conclusion

Machine learning has grown up over the years malicious web content detection, with adaptive measures for phishing, malicious extensions, and harmful URLs. With the understanding that ongoing challenges still exists in the form of adversarial attacks and resource consumption, Machine Learning approaches helps in

detecting malicious web content with adaptive measures for phishing, malicious extensions, and harmful URLs. The proposed Browsafe approach identifies safe and unsafe websites with an accuracy of 95.35%, but it also provides threat information and website score more efficiently and helps users to proceed or not to proceed with the those websites. In future innovations yet to come—unifying automation and human instinct—promising a more secure web landscape remain in the pipeline subject to continuous research and practical application.

References

1. Z. Zhang, J. Wu, N. Lu, W. Shi, Z. Liu, AdaptPUD: An accurate URL-based detection approach against tailored deceptive phishing websites. Comput. Netw. **265**, 111303 (2025). https://doi.org/10.1016/j.comnet.2025.111303
2. M.A. Rose, N. Basir, N.F. Heng, N.J. Zaizi, M.M. Saudi, Phishing detection and prevention using Chrome extension, 2022 10th International Symposium on Digital Forensics and Security (ISDFS), IEEE, Istanbul, Turkey, June 22 (2022), 1-6
3. E.S. El-Kenawy, M.M. Eid, H.L. Hussein, A.M. Osman, A.M. Elshewey. Optimized deep learning model using binary particle swarm optimization for phishing attack detection: A comparative study, Mesopotamian J. CyberSecurity. **5**, 685-703 (2025). https://doi.org/10.58496/MJCS/2025/041
4. A. Safi, S. Singh. A systematic literature review on phishing website detection techniques, J. King Saud Univ. Comput. Inf. Sci. **35**, 590-611(2023). https://doi.org/10.1016/j.jksuci.2023.01.004
5. I. Qabajeh, F. Thabtah, F. Chiclana, A recent review of conventional vs. automated cybersecurity anti-phishing techniques, Comput. Sci. Rev. **29**, 44-55 (2018). https://doi.org/10.1016/j.cosrev.2018.05.003
6. O.K. Sahingoz, E. Buber, O. Demir, B. Diri. Machine learning based phishing detection from URLs, Expert Syst. Appl. **117**, 345-357 (2019). https://doi.org/10.1016/j.eswa.2018.09.029
7. S. Salloum, T. Gaber, S. Vadera, K. Shaalan, A systematic literature review on phishing email detection using natural language processing techniques. IEEE Access **10**, 65703-65727 (2022). https://doi.org/10.1109/ACCESS.2022.3183083
8. N.Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, H. Fujita, Deep learning for phishing detection: Taxonomy, current challenges and future directions. IEEE Access **10**, 36429-36463(2022). https://doi.org/10.1109/ACCESS.2022.3151903
9. R. Alabdan, Phishing attacks survey: Types, vectors, and technical approaches. Future Internet **12**, 168 (2020). https://doi.org/10.3390/fi12100168
10. A. Basit, M. Zafar, X. Liu, A.R. Javed, Z. Jalil, K. Kifayat, A comprehensive survey of AI-enabled phishing attacks detection techniques. Telecommun. Syst. **76**, 139-154 (2021). https://doi.org/10.1007/s11235-020-00733-2
11. K.L. Chiew, K.S. Yong, C.L. Tan, A survey of phishing attacks: Their types, vectors and technical approaches. Expert Syst. Appl. **106**, 1-20 (2018). https://doi.org/10.1016/j.eswa.2018.03.050
12. Y. Shen, Z. Yan, R. Kantola, Analysis on the acceptance of global trust management for unwanted traffic control based on game theory. Comput. Secur. **47**, 3-25 (2014). https://doi.org/10.1016/j.cose.2014.03.010
13. L. Shahba, A. Heidary-Sharifabad, M. Mollahoseini Ardakani, Detection of fake web pages and phishing attacks with rabbit optimization algorithm. J. Supercomput. **81**, 313 (2025). https://doi.org/10.1007/s11227-024-06658-w
14. K.V. Deshpande, J.T. Singh, A systematic review on website phishing attack detection for online users. Int. J. Image Graph, 2750013 (2025). https://www.worldscientific.com/doi/abs/10.1142/S0219467827500136
15. V. Madhavan, G. P. Anand, S. Sridhar, Safe URL Detection with Privacy Using Machine Learning and Cryptography Techniques. 2025 3rd International Conference on Disruptive Technologies (ICDT), IEEE, Greater Noida, India, May 13 (2025), 321-326