

Article Citation Format

Tajuddeen Mashkur Muhammad, Agwom Itserim Emmanuel, Fatimah Adamu-Fika, Kamaludeen Shehu Bature, Abdullahi Danladi Maiauduga, Aanuoluwapo Enyojo Baba-Onoja & Onyinye Vivian Okpoko (2025): A Feature-Driven Approach to Phishing URL Detection Using Machine Learning. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology. Vol. 13, No. 2. Pp 109-116.
www.isteams.net/digitaljournal
[dx.doi.org/10.22624/AIMS/DIGITAL/V13N2P9](https://doi.org/10.22624/AIMS/DIGITAL/V13N2P9)

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 26th April, 2025
Review Type: Blind Peer
Final Acceptance: 17th June, 2025

A Feature-Driven Approach to Phishing URL Detection Using Machine Learning

¹Tajuddeen Mashkur Muhammad, ²Agwom Itserim Emmanuel, ^{*3}Fatimah Adamu-Fika, ⁴Kamaludeen Shehu Bature, ⁵Abdullahi Danladi Maiauduga, ⁶Aanuoluwapo Enyojo Baba-Onoja & ⁷Onyinye Vivian Okpoko

^{1,2,3,4,5}Department of Cyber Security, Air Force Institute of Technology, Kaduna, Nigeria.

^{6,7}Department of Computer Science, Air Force Institute of Technology, Kaduna, Nigeria.

***Corresponding Author:** f.adamu-fika@afit.edu.ng

ABSTRACT

Phishing attacks continue to pose a major cybersecurity threat by exploiting malicious URLs to deceive users and steal sensitive information. This study proposes a feature-driven machine learning system for phishing URL detection, leveraging customised lexical, structural, and domain-based attributes. Using publicly available datasets from PhishTank and OpenPhish, three models—LightGBM, XGBoost, and Logistic Regression—were evaluated. Among them, LightGBM achieved the best performance, with 93.2% accuracy and a 93.2% F1-score. Feature importance analysis highlighted the central role of URL length, suspicious keywords, and domain-related features in distinguishing phishing URLs from benign ones. The findings demonstrate the effectiveness of tailored feature engineering in improving detection accuracy and provide practical insights for building scalable, real-time phishing prevention tools. Future work will extend this framework by incorporating more diverse datasets, exploring deep learning approaches, and validating deployment in real-world environments.

Keywords: Cybersecurity, Feature Extraction, LightGBM, Machine Learning, Malicious URLs, Phishing Detection.

1. INTRODUCTION

Phishing attacks remain one of the most persistent and sophisticated threats in cyberspace, targeting individuals and organisations to steal sensitive information such as login credentials, financial data, and personal details. These attacks commonly rely on malicious Uniform Resource Locators (URLs), often embedded in emails or websites, to deceive users. In 2023, phishing accounted for 36% of all data breaches in the United States, underscoring its prevalence and impact (IT Governance, 2023). The growing complexity of phishing URLs—including obfuscation techniques and multi-layered redirection chains—demands advanced and adaptive detection mechanisms (Zhang et al., 2022; Cloudflare, 2023).

Figure 1 provides a foundational understanding of phishing attacks.

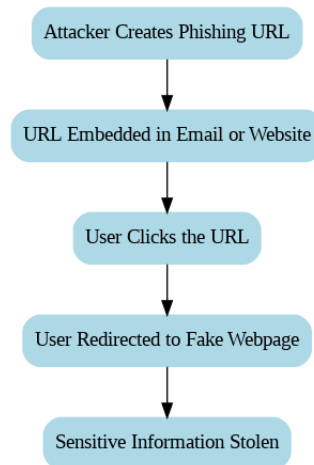


Figure 1: How phishing attacks work.

Traditional detection approaches such as blacklisting and heuristic-based systems have well-documented limitations. Blacklists, which rely on databases of previously identified malicious URLs, are easily evaded by the rapid creation of new or modified URLs (Zhang et al., 2022; Patel, 2018). Heuristic systems, meanwhile, apply predefined patterns but often fail to generalise to novel or highly obfuscated URLs (Patel, 2018; Zhang et al., 2016). These weaknesses highlight the need for detection methods that adapt to evolving phishing tactics (Liang et al., 2022; Zhang et al., 2016). Figure 2 illustrates the underlying behaviour of malicious URLs.

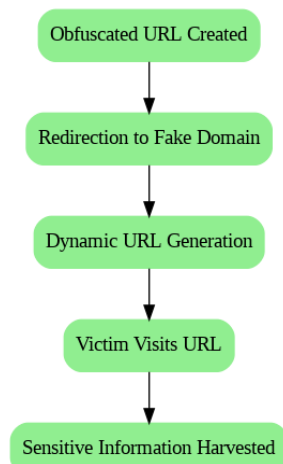


Figure 2: How malicious URLs operate.

Machine learning (ML) offers a powerful alternative by uncovering complex patterns in URL structures and behaviours that static methods often miss (Lavanya & Shanthi, 2023; Ghaleb et al., 2022). Unlike blacklists and heuristics, ML models can process large datasets and adapt to emerging phishing strategies (Zhang et al., 2022).

Recent research has demonstrated its effectiveness: Lavanya and Shanthi (2023) applied deep learning to lexical and host-based features, achieving high accuracy; Ghaleb, Shoufan, and Moustafa (2022) used ensemble learning to reduce false positives; and Liang, Chen, and Wang (2022) showed that domain-specific features such as URL length, character frequency, and domain age significantly improve classification performance. Collectively, these studies demonstrate the potential of ML-based approaches, but many still overlook the role of customised feature engineering and scalability for real-time applications.

Feature engineering is particularly critical for phishing detection. Incorporating domain- and context-specific features enhances accuracy and reduces false positives, yet many studies continue to rely on standard datasets with limited feature diversity (Liang et al., 2022; Zhang et al., 2016). Moreover, research has often prioritised comparing algorithms rather than addressing practical deployment issues such as real-time performance and integration into operational security systems (Liang et al., 2022; Zhang et al., 2016). Our earlier work (Muhammad et al., 2025) addressed part of this gap by focusing on behavioural features, specifically malicious redirection patterns, and demonstrating their significant contribution to detection accuracy. The present study extends those findings by integrating behavioural insights with a broader set of lexical, structural, and domain-based features. This expanded feature-driven approach aims to capture a fuller representation of phishing URL characteristics, enabling the development of a more versatile and effective detection framework.

Accordingly, this study develops and evaluates a machine learning-based phishing detection system that emphasises customised feature engineering and real-time applicability. Using datasets from PhishTank and OpenPhish (PhishTank, 2023; OpenPhish, 2023), the system extracts a diverse feature set and evaluates the performance of three models: LightGBM, XGBoost, and Logistic Regression. The key contribution of this work lies in advancing feature-driven approaches that improve detection accuracy while supporting real-time deployment, moving beyond algorithm-focused or offline evaluations found in prior studies.

2. METHODS AND MATERIALS

Dataset

The study utilised two publicly available datasets: PhishTank (PhishTank, 2023) and OpenPhish (OpenPhish, 2023). Both provide annotated collections of phishing and legitimate URLs suitable for machine learning research. The combined dataset contained 30,000 URLs: 15,000 phishing URLs and 15,000 legitimate URLs. Each URL was labelled as either malicious (phishing) or benign.

Feature Extraction

Feature engineering was central to improving model performance. Three categories of features were extracted from each URL (Liang et al., 2022; Zhang et al., 2022):

- Lexical Features
 - URL length
 - Number of special characters (e.g., /, ?, =, &)
 - Presence of suspicious keywords (e.g., “login,” “verify,” “secure”)
- Domain Features
 - Domain age (calculated using WHOIS data)
 - Top-level domain (TLD) type (e.g., .com, .net, .org)
- Structural Features
 - Number of subdomains

- Length of the hostname
- Use of an IP address instead of a domain name

These features were selected for their demonstrated relevance to phishing detection in prior studies.

Machine Learning Models

Three machine learning models were implemented and compared:

1. LightGBM – A gradient-boosting framework optimised for speed and efficiency on structured data.
2. XGBoost – A gradient-boosting algorithm with strong performance for classification tasks.
3. Logistic Regression – A linear baseline model, included for interpretability and comparison.

Figure 3 demonstrates how ML detects malicious URLs.

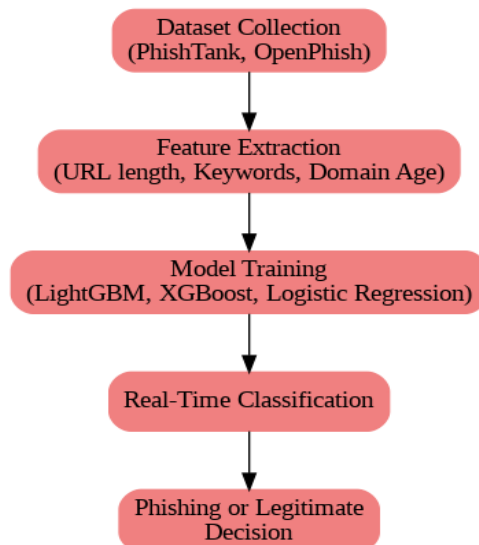


Figure 3: How ML detects malicious URLs.

Evaluation Metrics

Model performance was assessed using four metrics:

- Accuracy: Proportion of all URLs (phishing and legitimate) correctly classified.
- Precision: Proportion of URLs classified as phishing that were actually phishing.
- Recall (Sensitivity): Proportion of phishing URLs correctly identified.
- F1-Score: Balanced measure that combines precision and recall to indicate overall detection effectiveness.

Workflow

The workflow for system development, as shown on Figure 4, consisted of five steps:

1. Data Collection – URLs were gathered from PhishTank and OpenPhish and preprocessed to remove duplicates.
2. Feature Engineering – Lexical, structural, and domain features were extracted.
3. Data Splitting – The dataset was divided into training (70%) and testing (30%) subsets.
4. Model Training – LightGBM, XGBoost, and Logistic Regression were trained on the extracted features.

5. Evaluation – Models were evaluated on the test set using the selected metrics.

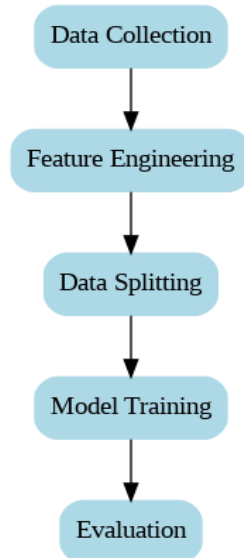


Figure 4: Workflow of the feature-driven phishing detection system.

3. RESULTS

The performance of the models was assessed on the testing dataset, which comprised 30% of the total data. Results are summarised in Table 1. The performance of the models was assessed on the testing dataset, which comprised 30% of the total data. Results are summarised in Table 1.

Table 1: ML Models Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
LightGBM	93.2%	92.8%	93.6%	93.2%
XGBoost	91.5%	91.0%	92.0%	91.5%
Logistic Regression	85.3%	84.5%	86.0%	85.2%

Key observations:

1. LightGBM achieved the highest accuracy (93.2%) and F1-score (93.2%).
2. XGBoost performed competitively at 91.5% accuracy.
3. Logistic Regression achieved 85.3% accuracy, performing lower than gradient-boosting models.

Feature importance analysis for LightGBM and XGBoost indicated the most influential attributes were:

- URL length
- Presence of suspicious keywords
- Domain age
- Number of subdomains

Figure 5 illustrates the feature ranking for LightGBM.

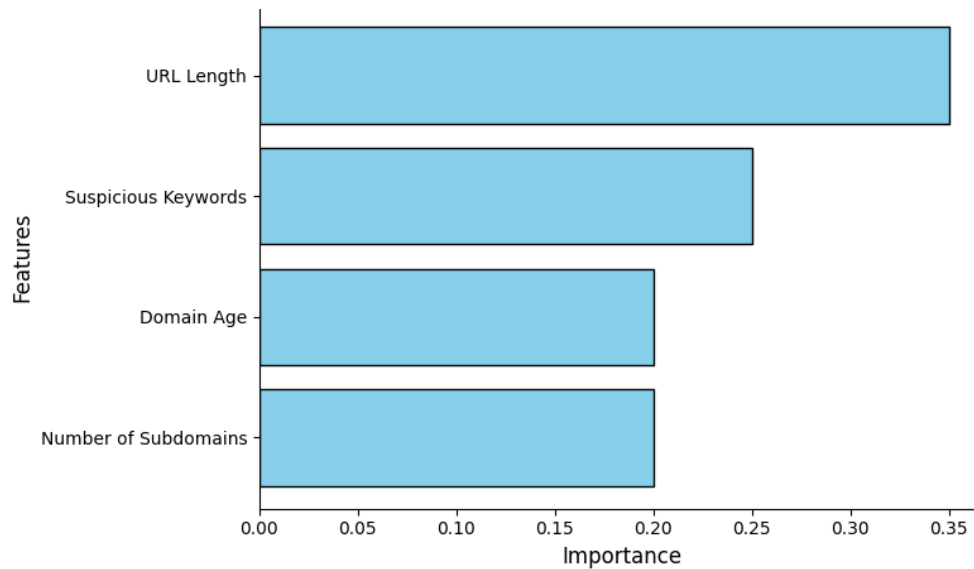


Figure 5: Feature importance ranking for LightGBM.

4. DISCUSSION

The results demonstrate that gradient-boosting models, particularly LightGBM, are highly effective for phishing URL detection. With an accuracy and F1-score of 93.2%, LightGBM outperformed both XGBoost and Logistic Regression, showing its ability to model complex interactions between features. These findings are consistent with prior work (Ghaleb et al., 2022), which also highlighted the strength of boosting methods for phishing detection tasks. The feature importance analysis further reinforces the role of feature engineering. URL length emerged as the most significant attribute, aligning with Liang et al. (2022), who noted that phishing URLs often use unusually long strings to obscure malicious intent. Similarly, suspicious keywords (e.g., “login,” “secure”) and domain age proved to be strong indicators of phishing behaviour. Together, these results underscore that carefully engineered features are central to achieving high detection performance.

This study advances beyond earlier research by achieving higher accuracy than ensemble methods reported by Ghaleb et al. (2022), and by focusing on custom feature extraction instead of relying solely on standard datasets. Additionally, by emphasising scalability and real-time applicability, the system addresses gaps noted by Zhang et al. (2022), who called for adaptive phishing detection frameworks. From a practical standpoint, the strong performance of LightGBM suggests its suitability for real-world deployment in browsers, email filters, and network security gateways. Its ability to scale efficiently also makes it well suited for handling large volumes of data in operational environments. Nevertheless, some limitations remain. The datasets used (PhishTank and OpenPhish) may not fully capture the global diversity of phishing attacks, and the study did not explore advanced deep learning models that could potentially further improve detection.

Future research should focus on:

1. Expanding dataset diversity through multi-source and real-time collection.
2. Exploring hybrid ML–DL approaches for enhanced performance.
3. Validating the system in real-time deployment environments.

5. CONCLUSION AND RECOMMENDATIONS

This study developed a machine learning-based system for phishing URL detection, leveraging customised feature engineering and datasets from PhishTank and OpenPhish. Three models were evaluated—LightGBM, XGBoost, and Logistic Regression—with LightGBM achieving the best performance (93.2% accuracy, 93.2% F1-score). The findings highlight the critical role of tailored features, particularly URL length and suspicious keywords, in distinguishing phishing URLs from legitimate ones.

Key contributions of this work include:

1. Development and evaluation of a feature-driven phishing detection framework combining lexical, domain, and structural attributes.
2. Identification of critical features (e.g., URL length, suspicious keywords, domain age) that significantly enhance detection accuracy.
3. Empirical validation of LightGBM as the most effective model, outperforming baseline methods.

These contributions demonstrate the promise of feature-driven ML approaches for real-world phishing prevention and lay the foundation for scalable, real-time deployment.

5.1 Recommendations

To advance phishing detection research and practice, we propose:

1. Expanding dataset diversity – Incorporating URLs from regional and targeted phishing campaigns to improve generalisability.
2. Exploring advanced techniques – Investigating hybrid ML-DL models (e.g., CNNs, transformers) to further boost performance.
3. Prototype deployment – Designing practical tools with user-friendly interfaces for browsers, email filters, and enterprise systems.
4. Updating feature sets – Continuously refining engineered features to address new and evolving phishing tactics.

5.2 Future Work

While this study focused on offline evaluation, the next stage will involve real-time deployment of the system. This will require building an adaptive architecture capable of processing live URL streams, testing performance in operational environments, and refining features for emerging attack strategies. The ultimate goal is to translate these promising offline results into a scalable, production-ready phishing detection tool that strengthens digital security at scale.

REFERENCES

- Cloudflare. (2023). *Introducing Cloudflare's 2023 phishing threats report*. Cloudflare. <https://blog.cloudflare.com/2023-phishing-report/>
- Ghaleb, T. A., Shoufan, A., & Moustafa, N. (2022). Phishing URL detection using ensemble learning. *Computers & Security*, 112, 102529. <https://doi.org/10.1016/j.cose.2021.102529>
- IT Governance. (2023). *51 must-know phishing statistics for 2023*. <https://www.itgovernance.co.uk/blog/51-must-know-phishing-statistics-for-2023>
- Lavanya, V., & Shanthi, R. (2023). Deep learning approaches for phishing URL detection: A comprehensive analysis. *Journal of Cybersecurity Studies*, 8(1), 45–58.

- Liang, H., Chen, Y., & Wang, Z. (2022). Feature-based phishing detection using self-paced wide and deep learning. *Information Systems Frontiers*, 24, 789–801. <https://doi.org/10.1007/s10796-021-10157-7>
- Muhammad, T. M., Emmanuel, A. I., Adamu-Fika, F., Bature, K. S., Bello, A., Okpoko, O. V., & Baba-Onoja, A. E. (2025). *Behavioural feature engineering for phishing detection: The case of redirection patterns*. Manuscript submitted for publication.
- OpenPhish. (2023). *Phishing intelligence platform*. <https://openphish.com>
- Patel, J. (2018). *Design and implementation of heuristic based phishing detection techniques* (Master's thesis, University of Victoria). University of Victoria. <https://dspace.library.uvic.ca/handle/1828/9880>
- PhishTank. (2023). *Verify suspected phishing URLs*. <https://phishtank.com>
- Zhang, W., Liu, J., & Zhou, X. (2022). Advanced phishing URL detection using machine learning techniques. *IEEE Access*, 10, 45678–45689. <https://doi.org/10.1109/ACCESS.2022.31745678>
- Zhang, Y., Hong, J., & Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing websites. In *Proceedings of the 16th International World Wide Web Conference (WWW2007)* (pp. 639–648). ACM. <https://doi.org/10.1145/1242572.1242659>