



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning-Based Detection Of Phishing And Authentic Urls

Priyanshu Popat Narale

M. Sc. Computer Science

Mamasahab Mohol College, Pune, Maharashtra, India

Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract—

Phishing continues to pose a serious challenge in cybersecurity, where attackers manipulate deceptive URLs to obtain confidential user information [1], [2]. This paper proposes a machine learning framework for phishing versus authentic URL detection using lexical URL features [9], [10]. The study follows a mixed-method design: an Experimental Research Method is used to build and evaluate machine learning models, and a Survey Research Method is used to analyze user awareness and behaviors toward suspicious links [20], [21]. URLs are preprocessed and transformed into eight computationally inexpensive lexical features including URL length, digit count, dot count, hyphen count, slash count, parameter count, subdomain count, and HTTPS presence [5], [6]. Logistic Regression is selected for interpretability, while Random Forest is selected for modeling non-linearity and robustness [15], [16]. Experiments on a dataset of **22,783 URLs** achieved **0.9388 accuracy** using Logistic Regression and **0.9382 accuracy** using Random Forest, respectively, with ROC-AUC reaching **0.9411**. Survey findings indicate users often trust weak safety cues such as HTTPS and frequently ignore registered domain verification, increasing phishing susceptibility [11], [12]. The results demonstrate that lexical-feature-based models can support practical phishing URL detection and awareness-driven security recommendations [10], [26].

Keywords— phishing detection, URL classification, lexical features, Logistic Regression, Random Forest, PhishTank-style feeds, Tranco, user survey.

I. INTRODUCTION

Phishing is among the most prevalent cyber threats, where adversaries impersonate legitimate services to trick users into revealing confidential information such as credentials, financial details, or OTPs [1], [2]. These attacks remain successful because they exploit human trust and attention rather than relying only on technical vulnerabilities [3], [11]. The growth of online banking, e-commerce, and digital payment services has increased user exposure to malicious URLs distributed through emails, SMS (smishing), and social media platforms [4], [26].

A phishing attack commonly begins with a deceptive URL that mimics a legitimate domain using techniques such as typosquatting, misleading subdomains, special characters, and shortened links [5], [6]. Traditional blacklist and reputation-driven defenses are effective against known threats but struggle to detect newly generated or short-lived phishing URLs [7], [8]. Consequently, Machine learning-based URL classification has become increasingly relevant due to its ability to learn generalized URL patterns and identify previously unseen phishing attempts [9], [10].

However, phishing success is also influenced by user awareness; users often trust weak signals such as HTTPS indicators and brand keywords, while ignoring domain structure [11], [12]. Therefore, this research combines ML-based detection with survey-based awareness analysis.

The main contributions of this work are:

1. A phishing vs authentic URL detection framework using lexical features.
2. Comparative evaluation of Logistic Regression and Random Forest under domain-aware splitting.
3. A user-awareness survey identifying behavioral phishing risks and misconceptions.

II. RELATED WORK

Early phishing defenses relied heavily on blacklists and web reputation services, which block URLs already reported as malicious [7], [8]. Although effective for known threats, such approaches struggle with new phishing domains, short-lived attacks, and fast-flux phishing infrastructure [13]. Rule-based systems examine lexical URL patterns such as excessive length and abnormal token structure [5], [6]. However, heuristics can be bypassed because attackers continuously evolve URL construction techniques [14].

Machine learning approaches improve phishing detection by learning decision boundaries from labeled URL datasets, enabling detection of unseen phishing URLs [9], [10]. Lexical feature-based ML is widely adopted due to low computational overhead and suitability for real-time detection without downloading webpage content [10]. Logistic Regression remains a strong baseline because it provides interpretability and explainability [15]. Random Forest is frequently used because it captures non-linear relationships and provides robustness to feature noise [16]. Deep learning approaches such as CNNs and LSTMs classify URLs at character level but require larger datasets and higher compute cost [17], [18].

Unlike many lexical-only phishing studies, this work enhances rigor by adding domain-aware splitting to prevent leakage, introducing label-noise modeling for realism, and integrating survey-based human-factor evidence [22], [24].

TABLE I. Comparison with Prior Studies

Study	Features	Model	Dataset	Reported Accuracy
Ma et al. [9]	Lexical + host	Online learning	URL corpus	~0.90+
Abu-Nimeh et al. [19]	Lexical	Multiple ML	Phishing corpora	0.89–0.94
Marchal et al. [14]	URL patterns	ML	Phishing URLs	~0.92
This study	Lexical	LR + RF	Verified phishing + top domains	0.9388/0.9382

III. METHODOLOGY

This research uses a mixed-method design with two methodologies.

(A) Experimental Research Method: model development, training, and evaluation using measurable performance metrics such as accuracy, F1-score, and ROC-AUC [20], [25].

(B) Survey Research Method: questionnaire-driven primary data collection to evaluate user awareness and behaviors toward phishing URLs [21], [26].

Model evaluation was performed using an 80:20 train-test split while ensuring domain-level separation between training and testing data, ensuring that URLs from the same domain do not appear in both training and test sets [22]. Logistic Regression was selected for interpretability and baseline comparison, while Random Forest was selected due to its non-linear learning capability and robustness [15], [16]. The Random Forest classifier was trained with 100 decision trees using default scikit-learn parameters.

Figure 1. Overall Research Workflow

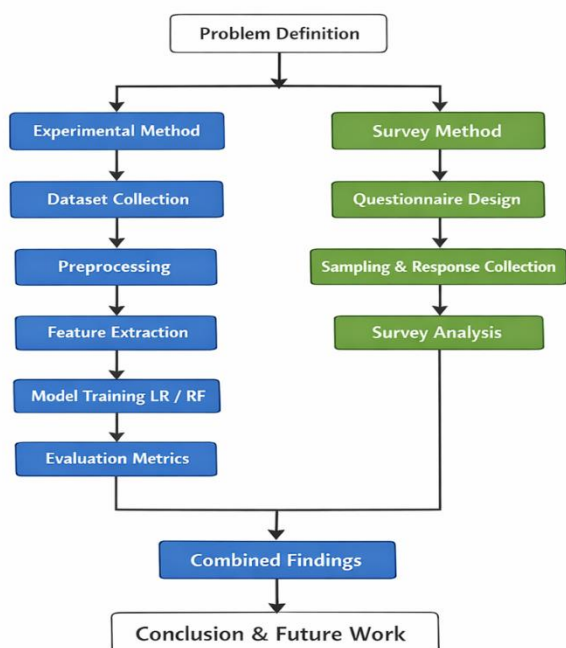


Figure 1. Overall Research Workflow

IV. DATASET

Phishing URLs were collected from verified phishing feeds aligned with threat intelligence repositories, and authentic URLs were derived using a top-domain list representing legitimate websites [7], [13], [23]. The final dataset contains **22,783 URLs**, including **10,843 phishing URLs** and **11,940 authentic URLs**. Phishing URLs were obtained from publicly available PhishTank-style feeds, while authentic URLs were sourced from the Tranco top-domain ranking.

A. Label Noise Modeling

A label noise ratio of **0.06** was modeled by randomly flipping 6% of training labels to simulate real-world dataset uncertainty caused by outdated phishing reports, inactive URLs, and feed inconsistency [24].

TABLE II. Dataset Summary

Category	Value
Total URLs	22,783
Phishing URLs	10,843
Authentic URLs	11,940
Total Features	8
Label Noise	0.06

V. SYSTEM ARCHITECTURE

The proposed system processes an input URL through preprocessing and feature extraction, then classifies it using trained ML models to output phishing or authentic decisions [10], [16].

Figure 2. System Architecture of the Proposed Framework

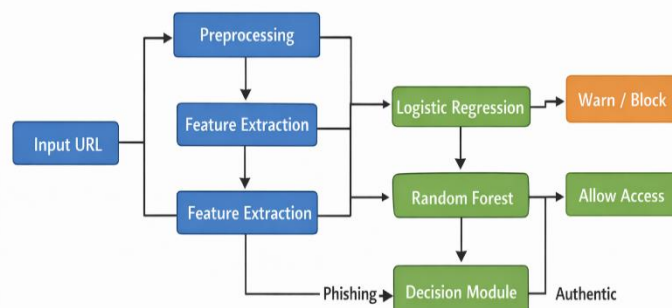
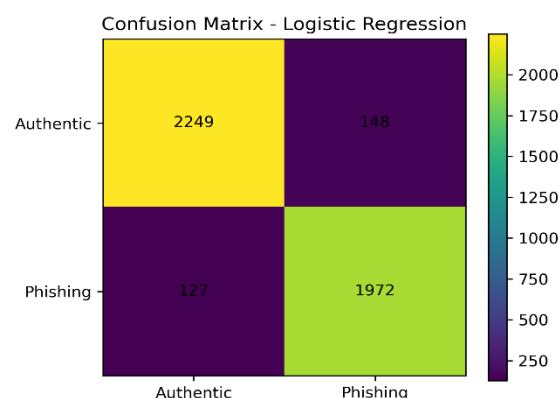


Figure 2. System Architecture of the Proposed Framework

VI. PREPROCESSING AND FEATURE ENGINEERING

URLs are



normalized by trimming spaces, ensuring consistent parsing, and removing duplicates [9], [10]. The system extracts eight lightweight lexical indicators, including URL length, character-based counts, structural delimiters, and HTTPS availability. [5], [6]. HTTPS presence is treated as a binary indicator; however, The presence of HTTPS alone does not ensure website legitimacy, as modern phishing sites frequently deploy SSL certificates [12].

Figure 3. Data Processing Pipeline

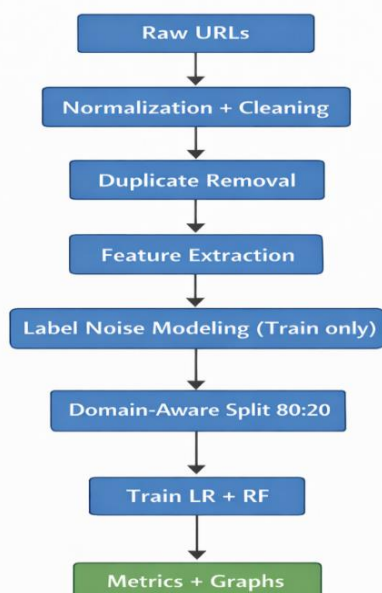


Figure 3. Data Processing Pipeline

Figure 4. Confusion Matrix Representation

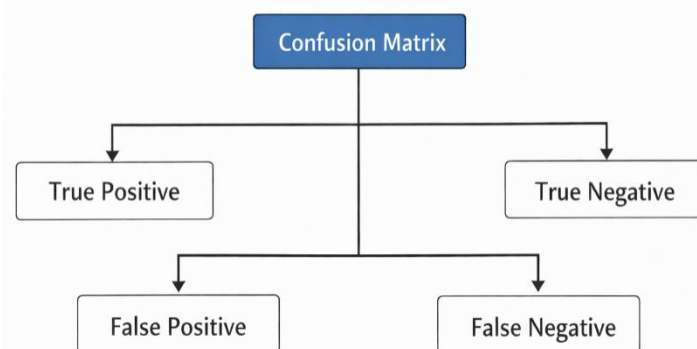


Figure 4. Confusion Matrix Representation

Figure 5. Confusion Matrix – Logistic Regression

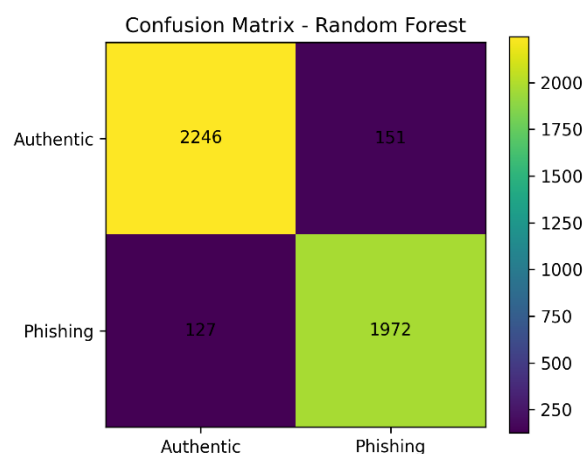


Figure 6. Confusion Matrix – Random Forest

VII. RESULTS AND EVALUATION

Performance evaluation employed multiple metrics, including accuracy, precision, recall, F1-score, ROC-AUC, along with 5-fold cross-validation. [25]. Domain-aware evaluation was used to avoid domain leakage and yield realistic performance estimates [22].

TABLE III. Experimental Results

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	5-Fold CV
Logistic Regression	0.9388	0.9302	0.9395	0.9348	0.9381	0.9400
Random Forest	0.9382	0.9289	0.9395	0.9342	0.9411	0.9384

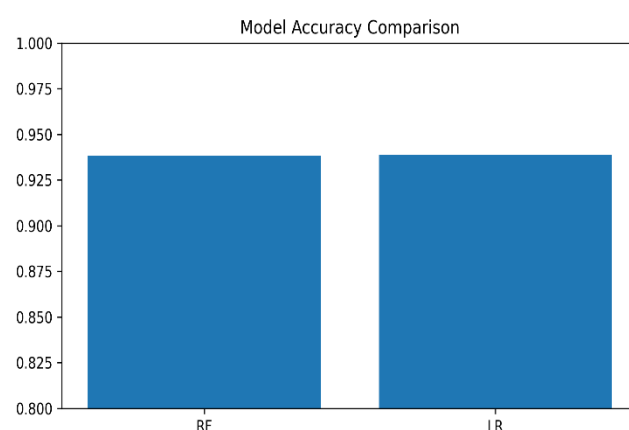


Figure 7. Accuracy Comparison Plot

VIII. USER SURVEY (Survey Research Method)

A structured questionnaire survey was conducted to measure user awareness and behavior regarding phishing URLs and link safety indicators [21], [26].

Ethics Statement

Participation was voluntary and anonymous, and the survey was conducted strictly for academic research purposes. No personally identifiable information was collected [21].

Figure 8. Survey Research Workflow

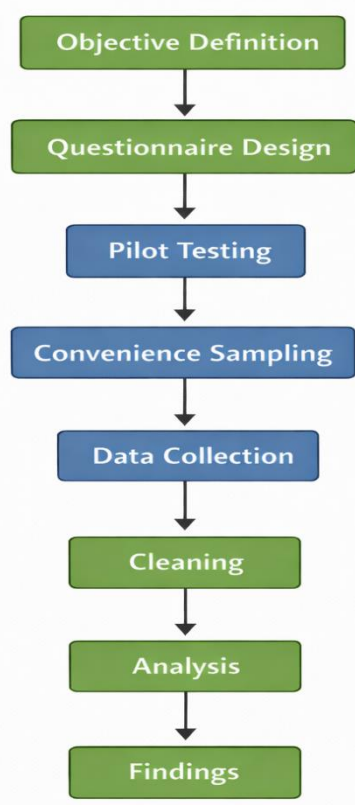


Figure 8. Survey Research Workflow

A. Questionnaire Design

The questionnaire included 12–15 items covering phishing awareness, URL-checking habit, HTTPS trust perception, shortened link behavior, and domain vs subdomain understanding [11], [12]. Likert-scale and Yes/No formats were used to standardize interpretation [21].

B. Sample Size and Demographics

A total of $N = 120$ responses were collected, mainly from participants aged 18–25 years. Convenience sampling was used due to time constraints [21].

TABLE IV. Demographic Summary

Category	Value
Sample size	120
Age group	18–25
Education	UG + PG
Sampling technique	Convenience sampling

C. Key Findings

Survey findings indicate human-factor weaknesses contribute significantly to phishing susceptibility [11]. Many respondents reported trusting HTTPS as a primary safety indicator when judging URL legitimacy [12]. Many participants did not verify registered domains and were vulnerable to misleading subdomain-based URL deception [11]. Clicking shortened URLs without destination verification was also observed [26].

D. Survey Limitations

Convenience sampling and self-reporting bias may affect generalization; larger probability-based surveys are recommended in future work [21].

IX. DISCUSSION

The results demonstrate lexical URL features provide strong predictive value for phishing classification [9], [10]. Logistic Regression performs strongly due to interpretability and near-linear separability of lexical indicators [15]. Random Forest achieves higher ROC-AUC, indicating stronger discrimination under non-linear feature interactions [16]. Label noise modeling strengthens experimental realism by simulating uncertainty in phishing feeds and reporting [24]. Survey findings confirm that user misconceptions and weak domain-checking behaviors increase phishing risk, supporting combined technical detection and user awareness strategies [11], [26].

X. CONCLUSION

This study presented a lightweight machine learning-based framework for distinguishing phishing URLs from authentic ones using lexical features and mixed-method validation. Logistic Regression achieved **0.9388 accuracy**, and Random Forest achieved **0.9382 accuracy**, with ROC-AUC reaching **0.9411**. The survey component confirmed that user misconceptions and weak domain validation behaviors contribute

to phishing risk. Overall, the proposed system is suitable for real-time detection and awareness-driven security improvement.

XI. FUTURE WORK

Future enhancements include integrating DNS and WHOIS signals, exploring deep learning models for URL classification, deploying browser extensions for real-time blocking, and conducting larger probability-sampled surveys for improved generalization [17], [18], [23].

REFERENCES

2024. [Online]. Available: <https://phishtank.org>. Accessed: Jan. 2026.
- [8] Google, "Safe Browsing," 2024. [Online]. Available: <https://safebrowsing.google.com>. Accessed: Jan. 2026.
- [9] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in *Proc. ACM SIGKDD*, 2009.
- [10] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Systems with Applications*, vol. 106, pp. 1–20, 2018.
- [11] S. Sheng et al., "Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proc. CHI*, 2010.
- [12] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: An empirical study of the effectiveness of web browser phishing warnings," in *Proc. SOUPS*, 2008.
- [13] A. Oest et al., "Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis," in *Proc. IEEE SPW*, 2018.
- [14] S. Marchal, J. Francois, R. State, and T. Engel, "Proactive discovery of phishing related domain names," in *Proc. RAID*, 2014.
- [15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.
- [16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [17] J. Huang et al., "Phishing URL detection via CNN and hybrid features," in *Proc. IEEE ICC*, 2020.
- [18] A. Bahnsen et al., "Classifying phishing URLs using recurrent neural networks," in *Proc. IEEE MLSP*, 2019.
- [19] M. Abu-Nimeh et al., "A comparison of machine learning techniques for phishing detection," in *Proc. eCrime Researchers Summit*, 2007.
- [20] C. R. Kothari, *Research Methodology: Methods and Techniques*, 2nd ed. New Age International, 2004.
- [21] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*. Pearson, 2019.
- [22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [23] Tranco, "Tranco Top Sites Ranking," 2024. [Online]. Available: <https://tranco-list.eu>. Accessed: Jan. 2026.
- [24] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
- [25] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC," *Journal of Machine Learning Technologies*, 2011.
- [26] R. Wash and E. Rader, "Too much knowledge? Security beliefs and protective behaviors among U.S. internet users," in *Proc. SOUPS*, 2015.
- [27] A. Y. Fu, L. Wenying, and X. Deng, "Detecting phishing web pages with visual similarity assessment," in *Proc. IEEE IWQoS*, 2006.
- [28] B. Gupta, N. A. Arachchilage, and K. Psannis, "Defending against phishing attacks: taxonomy and future directions," *Telecommunication Systems*, vol. 67, pp. 247–267, 2018.
- [29] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable AI," *IEEE*

Access, vol. 6, pp. 52138–52160, 2018.

[30] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *JMLR*, vol. 12, pp. 2825–2830, 2011.

[31] N. Japkowicz and S. Stephen, “The class imbalance problem,” *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[32] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE TKDE*, vol. 21, no. 9, pp. 1263–1284, 2009.

[33] ICANN, “WHOIS Lookup,” 2023. [Online]. Available: <https://www.icann.org>. Accessed: Jan. 2026.

[34] ISO/IEC, “ISO/IEC 27001: Information security management systems,” 2022.

[35] Y. Zhang et al., “Phinding phish: Evaluating anti-phishing tools,” in *Proc. NDSS*, 2007.

[36] M. Canali, D. Balzarotti, and A. Francillon, “The role of web hosting providers in detecting compromised websites,” in *Proc. WWW*, 2013.

[37] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[38] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[39] J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” in *Proc. ICML*, 2006.

[40] S. Van der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy array,” *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[41] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence,” *arXiv:1708.08296*, 2017.

[42] NIST, “Framework for Improving Critical Infrastructure Cybersecurity,” ver. 1.1, 2018. [Online]. Available: <https://www.nist.gov/cyberframework>. Accessed: Jan. 2026.

