IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15130

A Dual-Model Machine Learning System for Phishing Detection: URL Pattern Recognition and Email Content Analysis

Prof. K Thriveni[1], Praveen K[2], Manoj Kumar[3], Sharan S[4], Nishchal Gowda B R[5]

Assistant Professor, Dept. of AIML, The Oxford College of Engineering, Bangalore, India[1]

Final Year Student, Dept. of AIML, The Oxford College of Engineering, Bangalore, India[2,3,4,5]

Abstract: Phishing attacks persist as one of the most financially damaging cybersecurity threats, with recent global cybersecurity reports indicating a continuous year-over-year surge driven by large-scale automated phishing kits and AI-generated scam content. Standard reactive defenses like blacklists and static filters simply can't keep up with modern threats, failing specifically against zero-day attacks that leverage fresh domains or complex social engineering. A workable hybrid approach that uses both URL patterns and email content to identify phishing is needed to close this gap. In order to detect threats without any blacklist entries, the firstlay er employs a logistic regression model – character level TF-IDF vectorization to identify malicious sequence of n-grams 3 to 5 characters.The second layer is an email phishing detection laye r that uses a Random Forest Classifier trained on a UCI Spam base dataset with 57 markers, including word frequencies and capitalization patterns, to identify spam email contents. To avoid false flagging and promptly identify reliable websites, a whitelist is utilized. Both models are managed by the system, which is implemented as a Flask web application. By identifying both phishing URLs and spam patterns, the training results demonstrate the system's high detection rate and low false positives.

Keywords: phishing detection, TF-IDF, logistic regression, random forest, spambase, cybersecurity, URL analysis, and email security

I. INTRODUCTION

Phishing has become one of the most dangerous cyber threats we face today. Essentially, it's a con game where criminals pose as trustworthy organizations to trick people into handing over sensitive data—things like passwords, credit card details, or Social Security numbers. The damage can be devastating, both for individuals who lose their personal information and companies that suffer major financial hits. These attacks usually come through email, text messages, or malicious links, and they've gotten scary good at fooling people. Traditional security measures just can't keep up anymore. Take spear- phishing, for example—these are highly targeted attacks aimed at specific individuals. Or whaling attacks, which go after company executives and are incredibly hard to spot because they're so carefully crafted. The problem with old-school defense methods like blacklists is that they're always playing catch-up. Hackers constantly change their tactics, and these systems can only block threat they already know about. Every single day, attackers register thousands of new domains, making it nearly impossible for blacklists to stay current. Even when suspicious sites are discovered, it takes forever to properly identify and flag them as dangerous. And heuristic-based protection? It doesn't offer much defense either, especially against newer, more creative phishing schemes.

The reality is that attackers have gotten really clever at bypassing traditional filters. Take something as simple as the "@" symbol—they've found countless ways around it. They'll manipulate URLs, use visual tricks with look-alike characters (called homoglyphs), or set up entirely separate domains specifically for their phishing schemes. What's even more troubling is that some of these fake sites actually have legitimate SSL/TLS certificates, which makes them look completely trustworthy to the average person. When people fall for these scams, it doesn't just hurt them—it damages the reputation of the real companies being impersonated. Thank fully, we're seeing some promising solutions emerge from the AI and machine learning world. These newer systems can process massive amounts of data in real-time and, more importantly, they can adapt as phishing tactics evolve. That's a game-changer.

This research focuses on building a practical dual-layer detection system that tackles both email and URL phishing at the same time—since those are the two most common attack vectors. The URL detection component is particularly interesting. Instead of manually defining what makes a URL suspicious, it uses character-level TF-IDF vectorization combined with logistic regression to spot malicious patterns. Basically, it looks at character sequences and n-grams within

DOI: 10.17148/IJARCCE.2026.15130

the URL itself, learning to recognize things like obfuscation tricks and brand impersonation attempts without anyone having to explicitly tell it what to look for. That's a big improvement over older methods that relied heavily on hand-crafted features. On the email side, the system uses a Random Forest classifier trained on the UCI Spambase dataset to analyze the actual content of emails that users flag or submit for checking.

II. RELATED WORK

In the early days of phishing prevention, the go-to solution was static blacklists—basically just databases of websites that had already been flagged as malicious. This approach worked okay if you were dealing with known bad actors, but it completely fell apart when facing something new. The fundamental problem? Blacklists can't learn. They're reactive by nature, which means they're useless against zero-day attacks or freshly created phishing sites that haven't been reported yet. This limitation pushed researchers to think differently. Instead of maintaining endless lists of bad websites, why not look at the characteristics of the URLs themselves? That shift in thinking led to the development of machine learning models that focus on what's called feature- based or lexical analysis. These models examine the actual structure and patterns within a URL—things like unusual character sequences, domain length, or suspicious symbols. What makes this approach so appealing is its efficiency. Since these models only need to analyze the URL string itself, they're lightweight and fast. There's no need to actually load the webpage or render any content, which means they can make decisions almost instantly. It's a much smarter way to catch phishing attempts before they even have a chance to do damage.

A. URL – Based Phishing Detection

The earliest attempts at stopping phishing relied heavily on blacklists—essentially massive databases of known malicious domains. But here's the thing: they just don't work very well. Sure, they can block threats that have already been identified, but what happens when an attacker spins up a new domain, uses it for a quick phishing campaign, and then abandons it? The blacklist never even gets a chance to catch it. Newly registered domains, sophisticated fake websites, and novel phishing tactics slip right through these systems undetected. This reality pushed cybersecurity researchers to get more creative. For a long time, the solution involved painstaking feature engineering—experts would manually identify specific characteristics to look for in suspicious URLs. They'd extract details like URL length, the number of subdomains, whether an IP address was being used instead of a domain name, how old the domain was, SSL/TLS certificate information, and the presence of weird characters or suspicious keywords [1][2]. Machine learning algorithms, particularly ensemble methods like Random Forest, turned out to be pretty effective when fed these hand-crafted features, especially when working with large datasets.

But all that manual feature extraction had its limitations. It was time-consuming and required constant updating as attackers adapted their methods. That's when things started shifting toward more autonomous approaches. Modern systems now lean heavily on natural language processing techniques—specifically TF-IDF algorithms—that treat URLs like raw text rather than requiring someone to manually define what makes them suspicious. When you pair character- level TF-IDF with logistic regression, you get a system that can learn patterns on its own and achieve impressive accuracy. Some research has even taken this further by layering in unsupervised learning techniques, creating what's essentially a multi-level defense system that doesn't need humans constantly telling it what to look for.

B. Content-Based Email Detection

Email phishing is its own special beast. What makes these attacks so effective isn't just the technical sophistication—it's the psychological manipulation. Scammers create a sense of urgency, throw in some threats, and impersonate people or organizations you trust. They're banking on triggering an emotional response that makes you act before you think. The scary part? Most people don't even realize they're being manipulated. These threats hide in plain sight, disguised as legitimate correspondence that's nearly impossible for the average person to spot. Now, here's where machine learning comes in. Computers can't read text the way we do—they need numbers. That's where TF-IDF (Term Frequency-Inverse Document Frequency) becomes incredibly useful. Instead of just counting how many times words appear, TF-IDF measures how significant each term is across an entire collection of messages. This helps the system identify language patterns that are actually meaningful for spotting phishing attempts. When you combine TF-IDF with a solid classifier, the results are genuinely impressive. It's especially good at catching those carefully worded phishing emails because it can recognize important character sequences and word combinations (what researchers call n-grams) that might slip past simpler detection methods.

For testing these kinds of systems, researchers often turn to the UCI Spambase dataset—it's become something of a gold standard in the field. This dataset contains 57 different numerical features pulled from real emails: things like how often certain words appear, patterns in punctuation and symbols, and even bursts of capital letters (because apparently spammers love their ALL CAPS). What's great about this dataset is that it plays nicely with algorithms like Random Forest and

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.15130

Decision Trees without needing a ton of data preprocessing or normalization. Even older, more straightforward machine learning algorithms can achieve solid results when trained on it.

III. PROPOSED WORK

The proposed system implements a "defence-in-depth" strategy by establishing a hybrid, dual-model architecture. This framework is designed to detect phishing attempts across two orthogonal attack vectors: the delivery mechanism (URL) and the semantic payload (email content). The rationale for this hybridity is rooted in the multifaceted nature of modern attacks; some campaigns rely on obfuscated URLs to bypass network filters [1], [2], while others utilize sophisticated social engineering (e.g., Business Email Compromise) to elicit user compliance without malicious links [6]. To handle this, the system treats the problem as a two-part task. Let X be the input (the thing we want to check). The system uses a Routing Function R(X) to decide what kind of input it is. It then sends the input to the right expert: the URL Model ($M_{\{url\}}$)or the Email Model ($M_{\{nlp\}}$). The Final Answer Y is decided as follows:

$$\{Y\} = Murl(X) \text{ if } R(X) = \text{url}$$
$$Mnlp(X) \text{ if } R(X) = \text{email}$$

Here, Y is simply the final label the system gives: either "Phishing" or "Legitimate." Both models use the Random Forest algorithm, which is chosen because it is very accurate and good at handling complex data [3], [4].

A. Model 1: URL Phishing Detection via Character-Level TF-IDF

The entire Unlike traditional feature-engineering approaches requiring domain expertise to manually define URL characteristics, the URL model ($Murl$) employs automated pattern learning through character-level TF-IDF vectorization. This approach captures lexical patterns and character sequences discriminative of phishing URLs without explicit feature definition.

For a given URL string u, the system applies TF-IDF vectorization with the following parameters:

- Analyzer: Character-level (captures sub-token patterns)
- N-gram range: (3, 5) (extracts trigrams, 4-grams, and 5-grams)
- Maximum features: High-dimensional feature space for comprehensive pattern capture

This configuration transforms each URL into a high- dimensional sparse vector, where each dimension represents the TF-IDF weight of a specific character n-gram. For example, character sequences common in words like "login", "verify", or "secure" each receive weights based on their frequency in malicious versus legitimate URLs. This approach automatically identifies character sequences strongly associated with phishing URLs (e.g., sequences mimicking legitimate brands, hexadecimal encoding patterns, or obfuscation techniques) without requiring manual feature specification.

Classification Algorithm: Logistic Regression

The classification is performed using Logistic Regression on the TF-IDF vectors, with the regularization parameter adjusted for a tighter fit. We selected this method over heavier ensemble techniques because of its efficiency with sparse, high-dimensional data. It allows for linear-time predictions and provides easy-to-understand probability scores based on the sigmoid function. Additionally, the reduced regularization helps the model capture specific discriminative patterns more effectively. Technically, the model operates by learning a weight vector w and a bias b using maximum likelihood estimation.

$$P(y = 1|x) = \sigma(w^T x + b) = 1/(1 + e^{-(w^T x+b)})$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15130

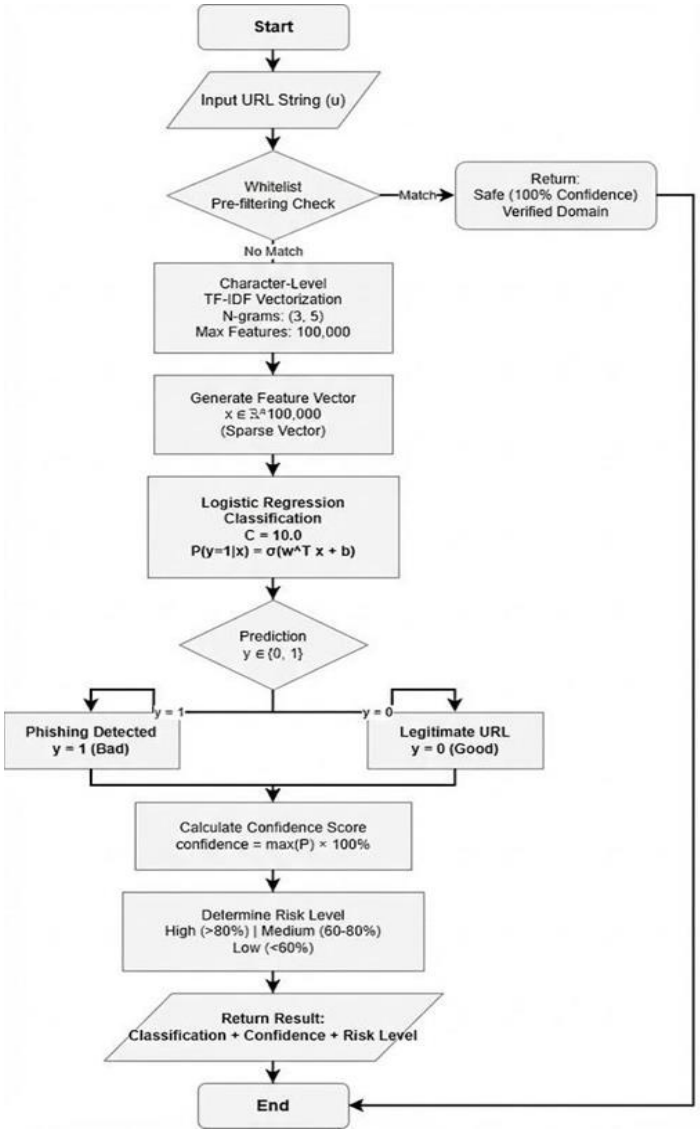

Fig. 1. Architecture of URL phishing detection

Whitelist Pre-Filtering:

To strictly limit false positives on established domains, we integrated a whitelist pre-filtering mechanism. Prior to any machine learning analysis, the URL's hostname is cross-referenced against a curated index of verified, popular websites. If a match occurs, the system bypasses the heavier computational steps and immediately classifies the URL as safe with absolute confidence. This architectural choice serves a dual purpose: it guarantees that known legitimate sites are never misclassified while significantly cutting down the processing load for trusted traffic.

B. Model 2: Email Content Detection via Spambase Features

For email analysis ($Memail$), we leverage the UCI Spambase dataset [8]. The feature space consists of three categories.

1. Word Frequency Features: The percentage of words matching specific keywords (e.g., financial terms, urgency indicators).

$$\text{word\_freq}WORD = 100 \times \frac{\text{count}(WORD \text{ in email})}{\text{total words in email}}$$

2. Character Frequency Features: The density of specific symbols (e.g., currency signs, exclamation marks).

$$\text{char\_freq}CHAR = 100 \times \frac{\text{count}(CHAR \text{ in email})}{\text{total characters in email}}$$

3. Capital Letter Run-Length Features: Metrics capturing the aggressive capitalization patterns characteristic of spam (average length, longest run, total count).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15130

Feature Extraction from Raw Text

To apply the trained model to real-world input, we implemented a feature extraction function that tokenizes raw email text and computes the Spambase statistics dynamically. This ensures the deployment data vector matches the training schema.

Classification Algorithm: Random Forest

Utilizes Random Forest for this classification due to its implicit feature selection capabilities and robustness against overfitting. The ensemble approach captures non-linear relationships between features effectively.

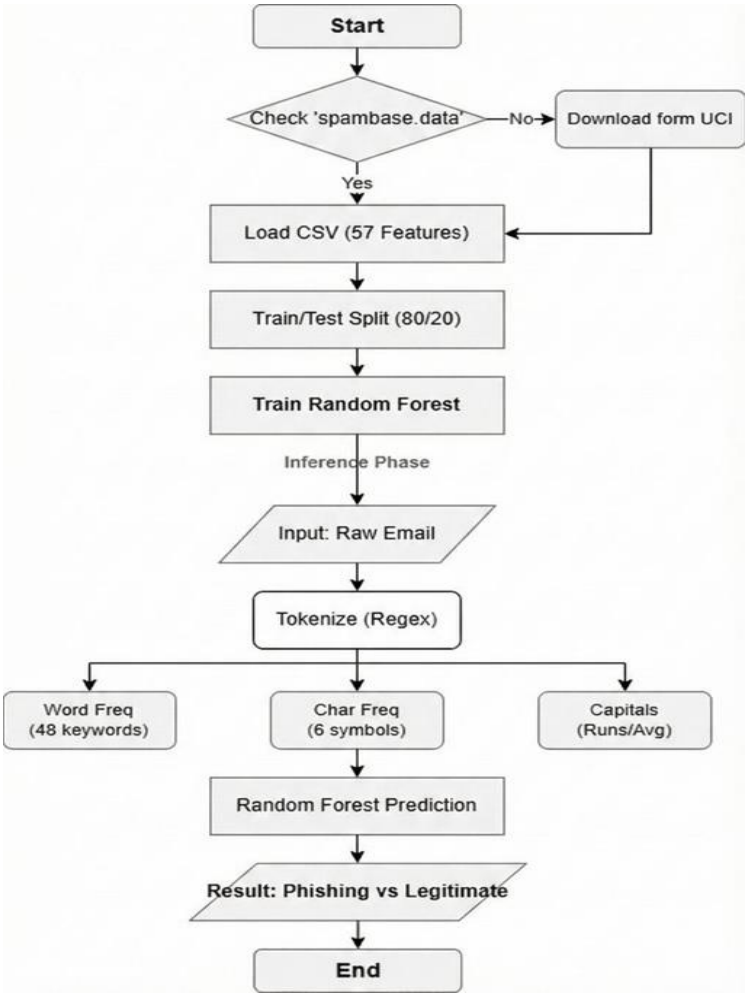

Fig 2. Architecture of Email Content Detections

IV. RESULTS AND DISCUSSIONS

Based on the provided paper, the results and discussions section highlights that the dual-model system was evaluated using standard machine learning practices to ensure effectiveness, utilizing stratified sampling to maintain class balance during training. The URL detection model, utilizing Logistic Regression on character-level TF-IDF features, demonstrated strong performance with an accuracy of 0.97, prioritizing high recall to minimize false negatives which is critical in cybersecurity. Simultaneously, the email content detection model, based on a Random Forest classifier and the UCI Spambase dataset, achieved exceptional performance with an accuracy of 0.94 and balanced precision and recall across all metrics. Ultimately, the quantitative data confirms that both models successfully identify phishing threats with high detection rates, validating the hybrid architecture's ability to handle both URL and email-based attack vectors.

V. CONCLUSION

This research presented a practical hybrid machine learning system for phishing detection, combining character-level TF-IDF URL analysis with Spambase-based email classification. The dual-model architecture effectively addresses the

DOI: 10.17148/IJARCCE.2026.15130

limitations of single-vector approaches. The URL model achieved high accuracy through automated pattern learning, while the email model leveraged robust statistical features [8]. The integration of whitelist pre-filtering further refined accuracy on trusted domains.

Future Enhancements Future work will focus on integrating advanced URL embeddings using deep learning (CNNs or Transformers) to capture semantic relationships [16]. We also aim to incorporate pre-trained language models for contextual email analysis and implement a continuous learning pipeline to address concept drift. Finally, improving adversarial robustness and integrating explainable AI techniques will be crucial for the next generation of phishing defence systems.

REFERENCES

[1]. M. A. Adebowale, K. T. Lwin, and E. A. Williams, "Machine Learning-Based Phishing Detection Using URL Features: A Comprehensive Review," *Journal of Cybersecurity Advances*, 2024.
[2]. A. R. R. S, S. S, and M. V, "Detection of Phishing Websites Using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 11, no. 4, 2024.
[3]. A. S. L, "Detecting Phishing URL Using Random Forest Classifier," *World Journal of Advanced Research and Reviews*, vol. 25, no. 2, pp. 762–769, 2025.
[4]. F. J. P. E, V. P, and C. A. C. B, "Web URLs Phishing Detection Model with Random Forest Algorithm," ResearchGate, 2024.
[5]. S. S. S. J, D. M, A. G, and R. A, "Phishing Detection System Through Hybrid Machine Learning Based on URL," *Dadi Institute of Engineering & Technology*, 2023.
[6]. A. A. Alhogail and A. Alsabih, "Applying Machine Learning and Natural Language Processing to Detect Phishing Emails," *Computers & Security*, vol. 110, 2021.
[7]. A. P. L and S. K, "Text Phishing Detection System Using Random Forest Algorithm," *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2024.
[8]. Hopkins, M., Reeber, E., Forman, G., & Suermondt, J., "Spambase Dataset," *UCI Machine Learning Repository*.
[9]. H. T, A. A, and E. E, "A Hybrid Phishing Detection System Using Deep Learning-Based URL and Content Analysis," *Elektronika ir Elektrotechnika*, vol. 28, no. 4, pp. 63-70, 2022.
[10]. S. Jain and S. S. H. M, "A Hybrid Approach for Alluring Ads Phishing Attack Detection Using Machine Learning," *Sensors*, vol. 23, no. 19, 2023.
[11]. R. J. van Geest, G. Cascavilla, J. Hulstijn, and N. Zannone, "The Applicability of a Hybrid Framework for Automated Phishing Detection," *Computers & Security*, vol. 139, p. 103736, April 2024.
[12]. A. A. Tawil, L. Almazaydeh, D. Qawasmeh, B. Qawasmeh, M. Alshinwan, and K. Elleithy, "Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT," *Computers, Materials & Continua*, vol. 81, no. 2, pp. 3395–3412, 2024.
[13]. S. Vajrobol, P. Pattanasethanon, and C. Tantisripreecha, "Mutual Information Based Logistic Regression for Phishing URL Detection," *Future Internet*, vol. 16, no. 3, 2024.
[14]. S. Das Guppta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques," *Annals of Data Science*, vol. 11, pp. 217–242, 2024.
[15]. A. I. Champa, M. F. Zibran, and W. Rahayu, "Curated Datasets and Feature Analysis for Phishing Email Detection with Machine Learning," *2024 3rd IEEE International Conference on Computing and Machine Intelligence (ICMI)*, 2024.
[16]. P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "A Systematic Review of Deep Learning Techniques for Phishing Email Detection," Electronics, vol. 13, no. 3823, September 2024.
[17]. N. Altwaijry, I. Al-Turaiki, R. Alotaibi, and F. Alakeel, "Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models," Sensors, vol. 24, no. 7, p. 2077, March 2024.
[18]. H. Takci and S. M. Rahman, "Highly Accurate Spam Detection with the Help of Feature Selection and Data Transformation Methods," The International Arab Journal of Information Technology, vol. 20, no. 1, pp. 29-40, January 2023.
[19]. R. Scavo and M. Bagić Babac, "UCI Spambase Dataset Analysis Using Classification Algorithms," University of Catania, Academic Year 2022/2023.
[20]. G. Bharath, "Hybrid Machine Learning Approaches for Phishing Email Detection," Insights2Techinfo, 2025.

