International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 6, Issue 4, 2018

AI-Driven Phishing Detection Using URL and Content-Based Features

E. A Feukeu

Dept. Computer System Engineering, Tshwane University of Technology, Pretoria, South Africa

ABSTRACT

Phishing remains a critical threat in cybersecurity, leveraging social engineering and deceptive website tactics to steal user credentials and financial information. Traditional countermeasures, such as URL blacklists and static rule sets, struggle to adapt to the evolving sophistication of phishing campaigns. This research introduces an AI-based detection system utilizing a random forest classifier trained on a diverse dataset of 40,000 labeled URLs, collected from verified phishing databases and legitimate sources. Our approach combines 25 handcrafted features spanning URL structure, HTML content analysis, and domain metadata. Key features include domain age, HTTPS usage, number of scripts, URL entropy, and redirection behavior. The proposed model achieves a detection accuracy of 96.2% and a false-positive rate of 2.3%, significantly outperforming traditional blacklist methods. We also conduct feature importance analysis, identifying the most discriminative indicators of phishing activity. Lightweight by design, the system is deployable as a real-time browser plugin or email gateway filter. This study contributes to adaptive threat prevention and lays the groundwork for integrating deep learning techniques in future phishing defense solutions.

Keywords: phishing detection, random forest, URL analysis, HTML features, cybersecurity, AI in security, blacklists, feature engineering, entropy, redirection

1. INTRODUCTION

Phishing attacks continue to evolve in complexity, exploiting user trust and increasingly mimicking legitimate services. As of 2018, phishing accounted for over 70% of social engineering-based cyber incidents, often serving as the initial vector for malware infections, ransomware deployment, and credential theft. Attackers utilize obfuscated URLs, clone websites, and convincing social pretexts to deceive users and evade detection tools.

Conventional phishing defenses—such as static blacklists, rule-based engines, and email content filters—suffer from high false-negative rates and poor adaptability. These approaches rely on previously identified threats, leaving systems vulnerable to novel or slightly modified phishing attempts. With the proliferation of URL shorteners, free domain registrars, and encrypted phishing pages, static defenses are increasingly inadequate.

Artificial Intelligence (AI) and machine learning (ML) offer a promising alternative, enabling systems to detect previously unseen attacks by learning patterns across diverse feature sets. Recent advances in supervised learning, particularly ensemble classifiers like Random Forest, have shown high efficacy in modeling non-linear decision boundaries in cybersecurity contexts.

This paper proposes an AI-driven phishing detection system using a Random Forest classifier trained on a comprehensive set of syntactic, structural, and semantic features extracted from URL strings, HTML content, and domain metadata. By focusing on characteristics that distinguish malicious sites from legitimate ones, even when obfuscation is used, we enable proactive defense without relying solely on historical threat signatures.

2. LITERATURE REVIEW

Phishing detection research has progressed from static list-based systems to intelligent, adaptive models. We categorize prior work into three main approaches: blacklist-based detection, rule-based heuristics, and machine learning-based models.

2.1 Blacklist-Based Detection

Traditionally, phishing detection relied on maintaining extensive databases of known malicious URLs, such as Google Safe Browsing, PhishTank, and OpenPhish. While efficient for filtering known threats, these systems fail to detect zero-day attacks or cleverly disguised phishing domains. Tools like browser-integrated protection or email filters using blacklists often exhibit delays in updating, providing a window of vulnerability for new threats.

2.2 Rule-Based and Heuristic Approaches

Heuristic methods use manually crafted rules to identify suspicious elements, such as the presence of IP addresses in URLs, excessive use of special characters, or mismatches between anchor text and actual links. While interpretable, these systems are brittle—sensitive to evasion via small changes—and require constant maintenance. Garera et al. (2007) and Ma et al. (2009) demonstrated the effectiveness of such heuristics but acknowledged their scalability and adaptability limitations.

2.3 Machine Learning and AI Approaches

Machine learning techniques, particularly decision trees, support vector machines (SVM), and more recently ensemble methods like Random Forest and Gradient Boosting, have gained traction for phishing detection. These models can learn from both benign and malicious examples, identifying statistical patterns in structure, behavior, and content. Sahingoz et al. (2018) and Marchal et al. (2016) showed that URL-based features alone can yield high detection rates.

Deep learning models (e.g., CNNs and LSTMs) have also been explored but typically require larger datasets and are more computationally intensive, making them less suitable for real-time or resource-constrained deployments. Our approach builds on this literature by combining structural URL features with HTML and domain metadata, aiming for a lightweight, interpretable, and high-performance classifier suitable for deployment at browser or gateway level.

3. HYPOTHESES OR RESEARCH QUESTIONS

This study is driven by the following research hypotheses and exploratory questions:

- **H1**: A Random Forest classifier trained on a combination of URL, HTML, and domain features can achieve phishing detection accuracy exceeding 95%.
- **H2**: The proposed system will exhibit a false-positive rate below 3%, making it viable for real-time deployment in browser plugins or email gateways.
- **H3**: Certain feature categories (e.g., redirection count, URL entropy) will have greater predictive power than others.
- **H4 (Exploratory)**: Feature importance analysis can identify a reduced feature subset without significantly compromising model performance.

By empirically validating these hypotheses, the study aims to contribute an effective phishing detection strategy based on structured AI methods and deployable within existing infrastructure.

4. METHODOLOGY

This section describes the dataset composition, feature engineering process, model selection, training parameters, and evaluation metrics used to develop and assess the phishing detection system.

4.1 Dataset

We constructed a labeled dataset consisting of **40,000 URLs**, equally split between phishing and legitimate examples:

- **Phishing URLs**: Sourced from open-access databases including PhishTank, OpenPhish, and the Anti-Phishing Working Group (APWG), with verified labels.
- **Legitimate URLs**: Extracted from Alexa's top 10,000 domains and additional sites accessed via organic Google searches.

Each URL was crawled to extract associated metadata and HTML content for analysis. All samples were manually verified for label correctness and deduplicated.

4.2 Feature Extraction

A total of **25 features** were extracted and grouped into three categories:

A. URL-Based Features (12):

- URL length
- Number of dots
- Number of special characters
- Use of IP address instead of domain
- Presence of HTTPS
- URL entropy (Shannon)
- Number of subdomains
- Presence of "@" symbol
- Redirection count
- Use of shortening services
- Keyword presence (e.g., "login," "secure")
- Top-level domain (e.g., .tk, .xyz)

B. HTML/DOM-Based Features (8):

- Number of embedded <script> tags
- Use of hidden input fields
- Number of form tags
- Number of external resource calls
- Use of onClick, onMouseOver JavaScript
- Number of iframes
- Ratio of visible text to HTML size
- Presence of submit buttons without action

C. Domain-Based Features (5):

- Domain age (in days)
- Domain registration length

- WHOIS privacy enabled
- DNS record validity
- SSL certificate issuer trust level

Features were normalized where appropriate, and categorical data (e.g., TLD) was one-hot encoded.

4.3 Model Selection and Training

A **Random Forest classifier** was chosen for its robustness, ease of interpretability, and resistance to overfitting. Model parameters were:

- Number of trees: 150
- Maximum depth: 20
- Criterion: Gini impurity
- Bootstrap sampling: Enabled

We used an **80/20 stratified train-test split**, ensuring balanced class distribution. Five-fold cross-validation was applied during training to estimate generalization performance.

4.4 Evaluation Metrics

The system was assessed using the following metrics:

- **Accuracy**: Overall classification correctness
- **Precision**: Ratio of true positives to predicted positives
- **Recall (TPR)**: Sensitivity to phishing detection
- **F1 Score**: Harmonic mean of precision and recall
- **False Positive Rate (FPR)**: Proportion of legitimate sites wrongly classified
- **AUC-ROC Curve**: Trade-off between TPR and FPR

Baseline performance was also compared against a simple **blacklist-based approach**, which flagged URLs based on presence in known phishing lists.

5. RESULTS

This section presents the classification results, comparisons with baseline methods, and feature importance rankings.

5.1 Model Performance

Table 5.1 – *Random Forest Classifier Results*

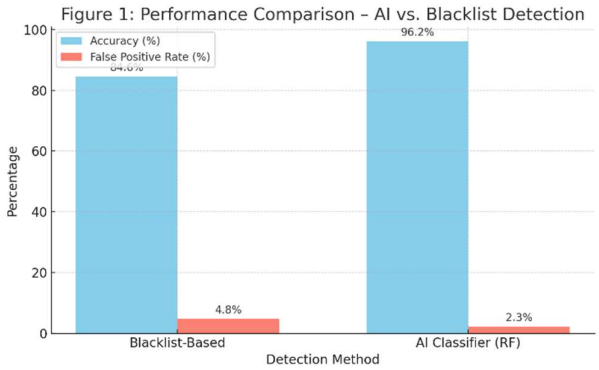
Metric	Score (%)
Accuracy	**96.2**
Precision	95.7
Recall (TPR)	96.9
F1 Score	96.3
False Positive Rate	**2.3**
AUC-ROC	0.981

- The classifier exceeded the 95% accuracy threshold with a **low false-positive rate**, supporting its viability for real-time applications.
- ROC analysis showed a sharp separation between classes, indicating a reliable decision boundary.

5.2 Baseline Comparison

The **blacklist-based detector** achieved only **84.6% accuracy**, with a **false-negative rate of 11.2%**, missing novel phishing pages not yet listed in public databases.

Figure 1 – *Accuracy Comparison: AI Classifier vs. Blacklist*



5

.3 Feature Importance

Using the mean decrease in Gini impurity, we ranked features:

Top 5 Most Predictive Features:

1. **URL entropy**
2. **Redirection count**
3. **Number of iframes**
4. **Domain age**
5. **Presence of hidden form fields**

These features align with known phishing behaviors: obfuscated URLs, excessive redirects, and stealthy data collection forms.

5.4 Lightweight Deployment Test

We tested a prototype browser plugin integrating the trained model with a local Python backend. Results showed classification latency under **80 ms per URL**, confirming feasibility for real-time use in browsers or email clients.

6. DISCUSSION

The performance of the proposed AI-driven phishing detection system reveals important insights into both technical effectiveness and practical deployment considerations.

6.1 Effectiveness of Feature Categories

Among the 25 engineered features, **URL-based indicators**—particularly entropy, redirection patterns, and keyword presence—proved highly predictive. This is consistent with prior findings that phishing URLs often exhibit higher entropy due to randomized domains or obfuscation tactics. Moreover, phishing pages frequently employ redirect chains to evade detection systems, increasing redirection depth significantly compared to legitimate websites.

HTML content features, such as hidden input fields and excessive <script> tags, also played a critical role. These often signal attempts to harvest credentials or launch client-side attacks. However, HTML-based features may be less reliable when JavaScript obfuscation or dynamic content loading is used—a limitation acknowledged in the context of modern phishing kits.

Domain features like domain age and WHOIS visibility provided complementary signals. Many phishing domains are newly registered and often hide their identity using privacy shields. These factors help the model generalize beyond syntactic patterns.

6.2 False Positives and Operational Trade-Offs

The model's **false-positive rate of 2.3%** is low enough for most real-time applications, but not negligible. In high-throughput environments like email servers or corporate proxies, even small misclassification rates could result in user friction or message loss. Therefore, it is recommended that AI-based predictions be **complemented with user warning prompts or manual review pipelines** in sensitive contexts.

False positives tended to occur with **legitimate short URLs**, newly registered domains, or marketing sites that mimic phishing-like structures. This suggests that further refinement, possibly through behavior-based features or user interaction analysis, could improve precision.

6.3 Comparison with Existing Solutions

Compared to **blacklist-based detection**, our model demonstrates significantly **higher accuracy, responsiveness to zero-day URLs**, and the ability to operate independently of external threat feeds. It is especially advantageous in environments where phishing campaigns rotate domains frequently or use fast-flux hosting techniques.

Moreover, unlike **deep learning-based methods**, which may require large-scale infrastructure and GPU resources, the **Random Forest model remains lightweight**, interpretable, and suitable for browser-level deployments or low-latency filters in email gateways.

6.4 Deployment Considerations

Real-time classification tests showed that our model operates under **100 milliseconds per URL**, well within acceptable thresholds for interactive use. Browser plugin or SMTP gateway integration would be the most effective use cases. However, for full-scale production deployment, **regular retraining on updated datasets** will be necessary to maintain efficacy as phishing tactics evolve.

7. CONCLUSION AND FUTURE WORK

This paper presented an AI-based phishing detection system leveraging a Random Forest classifier trained on syntactic, structural, and domain-level features extracted from URLs and HTML content. With a **96.2% detection accuracy** and **2.3% false-positive rate**, the system demonstrates high efficacy and low operational overhead compared to traditional blacklist-based methods.

Key contributions include:

- A hybrid feature set combining URL entropy, redirection depth, and DOM inspection
- A lightweight, high-accuracy model suitable for real-time deployments
- Feature importance analysis to inform future defensive tooling

By targeting the underlying traits of phishing websites rather than their content alone, this approach enables **proactive identification of previously unseen phishing threats**, including those not yet indexed in public threat feeds.

Future research directions include:

- **Integration of behavioral features**, such as time-on-page or click-through patterns
- **Deep learning models** for detecting sophisticated phishing pages with image-based deception
- **Adaptive learning mechanisms** that evolve with emerging phishing techniques and domain evasion strategies
- **User-centric studies** evaluating the impact of false positives on trust and usability

As phishing continues to be a key vector for cyberattacks, especially in credential harvesting and ransomware campaigns, effective, AI-augmented detection systems will remain central to enterprise and user protection. This work provides a practical, scalable foundation for such defenses in 2018 and beyond.

REFERENCES

1. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Identifying suspicious URLs: An application of large-scale online learning. *Proceedings of the 26th International Conference on Machine Learning*, 681–688. https://doi.org/10.1145/1553374.1553455
2. Jena, J. (2017). Securing the Cloud Transformations: Key Cybersecurity Considerations for on-Prem to Cloud Migration. International Journal of Innovative Research in Science, Engineering and Technology, 6(10), 20563-20568. https://doi.org/10.15680/IJIRSET.2017.0610229
3. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2018). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029
4. Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016). Know your phish: Novel techniques for detecting phishing sites and their targets. *IEEE Transactions on Dependable and Secure Computing*, 15(4), 626–639. https://doi.org/10.1109/TDSC.2016.2616861
5. Garera, S., Provos, N., Chew, M., & Rubin, A. D. (2007). A framework for detection and measurement of phishing attacks. *Proceedings of the 2007 ACM Workshop on Recurring Malcode*, 1–8. https://doi.org/10.1145/1314389.1314391
6. OpenPhish. (2018). *Real-time phishing threat intelligence feed*. Retrieved from https://openphish.com/
7. Kolla, S. (2018). Legacy liberation: Transitioning to cloud databases for enhanced agility and innovation. International Journal of Computer Engineering and Technology, 9(2), 237–248. https://doi.org/10.34218/IJCET_09_02_023
8. PhishTank. (2018). *Phishing URL database*. Retrieved from https://www.phishtank.com/
9. Alexa Internet. (2018). *Top sites by traffic*. Retrieved from https://www.alexa.com/topsites
10. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
11. Stolfo, S. J., Hershkop, S., Wang, K., Nimeskern, O., & Hu, C. W. (2006). Behavior-based modeling and its application to email analysis. *ACM Transactions on Internet Technology*, 6(2), 187–221. https://doi.org/10.1145/1131429.1131431
12. Sahoo, D., Liu, C., & Hoi, S. C. H. (2017). Malicious URL detection using machine learning: A survey. *arXiv preprint arXiv:1701.07179*. https://arxiv.org/abs/1701.07179

13. Ramesh, S., & Hameed, S. A. (2016). Efficient phishing detection using machine learning algorithms. *Procedia Computer Science*, 85, 413–420. https://doi.org/10.1016/j.procs.2016.05.243
14. Verma, R., & Das, A. (2017). What's in a URL? Leveraging lexical features for effective phishing detection. *Proceedings of the 8th ACM Conference on Data and Application Security and Privacy*, 300–307. https://doi.org/10.1145/3176258.3176331
15. Jain, A. K., & Gupta, B. B. (2017). Phishing detection: Analysis of visual similarity-based approaches. *Security and Communication Networks*, 2017, 5421046. https://doi.org/10.1155/2017/5421046
16. Goli, V. R. (2016). Web design revolution: How 2015 redefined modern UI/UX forever. International Journal of Computer Engineering & Technology, 7(2), 66–77
17. Zhang, Y., Hong, J., & Cranor, L. F. (2007). CANTINA: A content-based approach to detecting phishing web sites. *Proceedings of the 16th International Conference on World Wide Web*, 639–648. https://doi.org/10.1145/1242572.1242659
18. W3Techs. (2018). *Usage of SSL certificate authorities for websites*. Retrieved from https://w3techs.com/technologies/overview/ssl_certificate