JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 2, Issue 1, 2022 | **Article ID:** JMRIS-2022-002 | **Pages:** 13-24 | **Article Type:** Systematic Technical Review
Received: 01-01-2022 | **Revised:** 22-01-2022 | **Accepted:** 18-02-2022 | **Published:** 15-03-2022

Explainable Machine-Learning Models for Detecting Phishing URLs Targeting Healthcare Organizations: A Systematic Technical Review and Deployment Framework

Benjamin Clarke[1*]

[1] *Istinye University, Turkey*

***Corresponding author:** Benjamin Clarke, Istinye University, Istanbul, Turkey | Ben11252@istinye.edu.tr

ABSTRACT

Background: Phishing remains a major cybersecurity threat to healthcare organizations because successful attacks may expose employee credentials, protected health information, clinical systems and critical network infrastructure. Conventional blacklist-based controls are effective against known malicious domains but may fail against newly created or rapidly changing phishing URLs. Machine-learning models using lexical, domain, host and webpage characteristics offer an opportunity for earlier detection, although highly accurate black-box predictions may provide limited operational insight to security analysts. **Objective:** To synthesize high-quality evidence published through 2021 concerning machine-learning detection of phishing URLs and develop an explainable deployment framework tailored to healthcare organizations. **Methods:** A structured technical evidence synthesis was conducted using peer-reviewed literature published no later than 31 December 2021. References were restricted to Q1 journals in cybersecurity, artificial intelligence, information systems, medical informatics and digital health. Evidence concerning URL classification, phishing detection, feature engineering, benchmark construction, explainable artificial intelligence and healthcare phishing vulnerability was reviewed. No novel classifier performance was fabricated because a healthcare-specific labeled URL dataset was not available. **Results:** Pre-2022 studies demonstrated strong performance of machine-learning URL classifiers. A large URL-based study reported 97.98% accuracy using Random Forest, while a 2021 lexical approach using only nine URL features reported accuracy of 99.57%. Benchmark research demonstrated that Random Forest remained highly competitive, with feature selection producing accuracy of approximately 96.83% while reducing computational burden. However, performance varied according to dataset construction, feature availability and temporal sampling. Explainability is particularly important in healthcare because security teams must distinguish actionable indicators from spurious correlations. A tiered architecture using inexpensive lexical features for initial screening, domain and host characteristics for uncertain cases, and content-based analysis for secondary inspection offers a practical balance among speed, accuracy and interpretability. **Conclusion:** Explainable machine learning can provide a practical additional defense layer against phishing URLs in healthcare organizations. Systems should prioritize reproducible datasets, temporally separated validation, low-latency lexical features, interpretable model outputs and human security-operations oversight rather than relying on headline accuracy alone.

Keywords: phishing; phishing URL; machine learning; explainable artificial intelligence; healthcare cybersecurity; Random Forest; lexical features; cyberattack

1. Introduction

Healthcare delivery increasingly depends on interconnected digital infrastructure, including electronic health records, email, cloud services, remote-access systems, clinical portals and administrative applications. This digital dependence creates substantial operational benefit but also increases the consequences of compromised credentials.

Phishing represents one of the simplest mechanisms through which attackers can obtain legitimate access.

Gordon et al. analyzed 2,971,945 simulated phishing emails delivered across six US healthcare institutions. Employees clicked 422,062 messages, corresponding to an overall rate of 14.2%, while the median campaign-level click rate was 16.7% [1].

Healthcare phishing has consequences beyond compromise of an individual's email account. Obtained credentials may provide access to internal systems, confidential information or infrastructure from which additional attacks can be conducted.

The COVID-19 period further emphasized this vulnerability. Healthcare cybersecurity reviews published during 2020–2021 identified phishing, ransomware and malware among prominent threats affecting health systems [2,3].

User education remains necessary but cannot intercept every attack. Even well-trained employees can be presented with sophisticated messages containing links whose malicious nature is not visually obvious.

Automated URL analysis consequently offers an additional defensive layer.

Traditional approaches frequently rely on blacklists containing URLs or domains previously confirmed as malicious. Such systems perform well against known attacks but are intrinsically reactive.

Machine learning instead attempts to recognize characteristics associated with phishing.

Sahingoz et al. demonstrated the feasibility of URL-only phishing detection using a large dataset containing 36,400 legitimate and 37,175 phishing URLs [4]. Their Random Forest approach using natural-language-processing-based features achieved reported accuracy of 97.98%.

Gupta et al. later demonstrated that high detection performance could be achieved using only nine lexical URL features, reporting 99.57% accuracy with Random Forest [5].

High predictive performance, however, creates a second problem:

Why has a particular URL been classified as phishing?

An explainable detector should provide an answer meaningful to security analysts—for example, excessive subdomains, suspicious lexical tokens, an unusually long URL, misleading hostname structure or abnormal registration characteristics.

The present review therefore aimed to identify the strongest pre-2022 machine-learning approaches for phishing URL detection; determine which feature groups are suitable for real-time healthcare deployment; examine limitations affecting model generalization; establish the role of explainable artificial intelligence; and propose a deployment architecture for healthcare security operations.

2. Methods

2.1 Review Design

A structured technical evidence synthesis was conducted using literature available through 31 December 2021.

Search concepts included:

phishing URL, phishing website, machine learning, Random Forest, lexical features, malicious URL, phishing detection, explainable artificial intelligence, cybersecurity, healthcare phishing, and healthcare cybersecurity.

The evidence was divided into four domains: phishing-URL detection, feature engineering and dataset design, explainability and model interpretation, and healthcare-specific cybersecurity relevance.

The evidence eligibility framework is summarised in Table 1.

No healthcare-specific classifier dataset was created for this paper; therefore, no original accuracy, sensitivity or AUC estimates are presented.

3. Why URL-Level Detection Matters

A phishing email frequently operates as a delivery mechanism.

Its immediate objective is often to persuade the recipient to open a URL leading to a credential-harvesting page, a counterfeit authentication portal, malicious file delivery, drive-by compromise or additional social-engineering content.

Analyzing the URL before permitting interaction has several advantages.

First, the detector can operate at an email gateway.

Second, lexical analysis does not necessarily require visiting the malicious page.

Third, URL features can be extracted rapidly.

Fourth, the system may identify previously unseen URLs when their structural properties resemble known phishing behavior.

This distinction makes machine-learning detection particularly useful against attacks not yet represented in conventional reputation lists.

4. Principal Evidence for Machine-Learning Detection

Selected pre-2022 Q1 evidence relevant to phishing detection is summarised in Table 2.

Sahingoz et al.'s contribution was especially relevant because the architecture avoided dependence on third-party services and was intended for real-time detection [4].

Gupta et al. reduced the feature burden further. Their 2021 framework required only nine lexical characteristics and was designed for real-time operation [5].

Hannousse and Yahiouche addressed another major weakness in the literature: benchmark quality. Their reproducible framework examined 87 phishing-related features and found Random Forest to be the most predictive classifier in their experiments. Hybrid feature sets produced 96.61% accuracy, increasing to approximately 96.83% following feature selection [6].

These results suggest that adding every available feature is unnecessary.

5. Feature Architecture

Phishing-detection features can be divided into three principal groups.

5.1 Lexical URL Features

Lexical features are derived directly from the URL string.

Examples include total URL length, hostname length, number of dots, number of subdomains, number of digits, number of special characters, presence of an IP address instead of a domain name, suspicious keywords, character entropy, unusual delimiters, excessive path depth and token composition.

Their principal advantage is speed.

The detector does not need to load the target webpage or contact an external reputation service.

For healthcare email gateways, this represents an attractive first-line strategy.

5.2 Host and Domain Features

Host-based features provide additional information concerning the infrastructure supporting the URL.

Potential characteristics include domain age, DNS information, certificate properties, registration duration, hostname reputation and abnormal domain structure.

These variables can improve discrimination but may introduce external dependencies and latency.

An attacker can also deliberately modify infrastructure over time.

5.3 Content Features

Content features require examination of the target page.

Potential indicators include suspicious login forms, external form actions, hidden elements, excessive redirects, mismatches between displayed and actual links, JavaScript behavior, iframe use, external resource loading and similarity to legitimate brands.

Content analysis potentially provides rich information but carries greater computational and operational cost.

Hannousse and Yahiouche found that some webpage-content and externally retrieved variables could be time-consuming, an important consideration for real-time detection [6].

The explainable machine-learning architecture proposed for phishing URL detection in healthcare organizations is presented in Figure 1.

6. Recommended Feature Set

Candidate features for an interpretable healthcare phishing-URL model are summarised in Table 3.

For explainability, features should ideally correspond to concepts that a security analyst can independently verify.

A statement such as:

"Blocked because domain age <7 days, URL contains six subdomains, and the healthcare brand appears only in the path"

is more operationally useful than:

"Neural activation score = 8.61."

7. Model Selection

Several algorithms are plausible.

Candidate classifiers for explainable phishing detection are compared in Table 4.

Random Forest is particularly attractive because strong results were repeatedly observed in pre-2022 phishing research [4-6].

However, the best-performing algorithm on a single dataset should not automatically be deployed.

A difference between 99.3% and 99.5% accuracy may be operationally meaningless if the more accurate model is substantially slower, produces poorly calibrated probabilities, fails on newly collected URLs, cannot be explained or generates more false positives on healthcare-specific domains.

8. Explainable Artificial Intelligence

Arrieta et al. described explainable AI as a collection of approaches intended to make AI systems and their decisions understandable to humans [10]. Their 2020 taxonomy emphasized that explainability must be interpreted according to the user and decision context.

For phishing detection, there are two relevant levels.

8.1 Global Explainability

Global explanation asks:

What characteristics generally cause the model to classify URLs as phishing?

Potential outputs include ranked feature importance, interpretable decision rules, partial dependence relationships and simplified surrogate structures.

This information assists security teams in validating whether the model has learned plausible phishing behavior.

8.2 Local Explainability

Local explanation asks:

Why was this particular URL blocked?

This is arguably more operationally useful.

For example, risk may increase because a URL contains an IP-based host, abnormal length, multiple subdomains, credential-related terms and a newly registered domain.

A security analyst can then rapidly inspect the evidence.

9. Explainability Is Not Automatically Reliability

Explainable AI should not be treated as a cosmetic layer.

Rudin argued that in consequential applications, inherently interpretable models may sometimes be preferable to complex black-box systems followed by post-hoc explanations [11].

The same principle applies to cybersecurity.

If a simple tree or compact Random Forest performs almost as well as a very deep neural network, the simpler system may offer greater operational value.

An explanation also does not prove that a prediction is correct.

A classifier may confidently explain a false positive.

Therefore, prediction quality, calibration, generalization and explanation quality must be assessed separately.

10. Healthcare-Specific Deployment

Healthcare institutions represent an unusually sensitive deployment environment.

Phishing can target clinicians, administrative staff, researchers, trainees, billing teams, executives and IT personnel.

Healthcare organizations also commonly operate heterogeneous systems and large numbers of endpoints.

Gordon et al. noted that healthcare workers may possess credentials for multiple internal information systems and that email provides an accessible attack route [1].

A healthcare-specific detector should therefore pay particular attention to impersonation of electronic health-record login systems, payroll portals, institutional webmail, Microsoft or cloud authentication, laboratory systems, human-resources portals, clinical scheduling systems, insurance portals and health-authority websites.

However, the classifier should not simply learn that healthcare terminology is suspicious.

Attackers and legitimate organizations use many of the same words.

The relationship between the claimed brand and actual registered domain is more informative than the token itself.

11. Tiered Detection Strategy

A practical healthcare system should not necessarily execute every expensive analysis on every URL.

The tiered feature-extraction strategy balancing detection performance, latency and explainability is presented in Figure 2.

The approach can operate in three sequential tiers. Tier 1 uses lexical screening, in which every incoming URL is analyzed using inexpensive URL-string features and obvious low-risk or high-risk URLs can be classified immediately. Tier 2 performs infrastructure assessment, with uncertain URLs triggering domain, DNS, registration and certificate analysis. Tier 3 uses sandboxed content inspection, in which only unresolved high-risk cases are loaded within an isolated analysis environment for webpage-content evaluation.

This architecture reduces both computational burden and unnecessary contact with potentially malicious infrastructure.

12. Validation Requirements

A major weakness of phishing-detection research is the possibility of obtaining excellent results on a static dataset without demonstrating real-world generalization.

The recommended evaluation framework is summarised in Table 5.

Random train-test splitting alone is inadequate when identical domains or near-duplicate campaigns appear across both datasets.

A stronger design should perform temporal validation.

For example, training could use URLs collected from January to June, validation could use URLs collected from July to September, and the external test could use URLs collected from October to December or from another source.

This better represents deployment against future attacks.

13. Dataset Design

Hannousse and Yahiouche emphasized the short-lived and evolving nature of phishing webpages and the need for reproducible benchmark construction [6].

A healthcare-focused dataset should ideally contain general legitimate URLs, verified phishing URLs, legitimate healthcare domains, phishing URLs impersonating healthcare organizations, legitimate identity-provider URLs and healthcare-related third-party services.

This is essential because healthcare networks legitimately connect to many domains that may superficially appear unusual.

Without sufficient healthcare-specific negative examples, the model may generate excessive false alerts.

14. Avoiding Information Leakage

Information leakage can produce unrealistically high classifier performance.

For example, multiple URLs may originate from the same phishing domain, such as malicious-domain.com/login, malicious-

domain.com/account, malicious-domain.com/security and malicious-domain.com/verify.

If some URLs enter training and others enter testing, the classifier may effectively recognize the domain rather than generalize to unseen attacks.

Dataset splitting should therefore occur at an appropriate domain or campaign level, not solely at individual URL level.

Duplicate and near-duplicate URLs should also be removed before validation.

15. Class Imbalance

A production email gateway may inspect millions of benign URLs and comparatively few phishing URLs.

Consequently, accuracy can become misleading.

A detector classifying 99.9% of URLs as legitimate can appear highly accurate in an extremely imbalanced environment while missing most attacks.

Precision and recall should therefore receive greater emphasis.

Healthcare environments may prioritize high recall because missed attacks can be costly.

However, maximizing sensitivity without controlling false positives can overwhelm security analysts and encourage users to ignore warnings.

The operating threshold should therefore reflect both security consequences and alert workload.

16. Proposed Explainable Healthcare Model

A strong initial model could contain approximately 15–30 carefully engineered features.

A recommended implementation would compare logistic regression, decision tree, Random Forest and gradient boosting.

Random Forest or boosting may provide the final classifier if external validation supports improved performance.

The system should generate three outputs for each URL: a risk probability, such as “Phishing probability: 0.94”; a decision, such as “BLOCK”; and an explanation identifying primary contributors, such as a recently registered domain, five subdomains, brand-name mismatch, excessive URL length and a credential-related path.

This output is substantially more useful to security analysts than classification alone.

17. Security Operations Workflow

Predictions could be integrated at several points. At the email gateway, URLs can be evaluated before message delivery. At a secure web gateway, URLs can be evaluated when users attempt navigation. Within the security operations center, high-risk cases can generate prioritized alerts. Browser warnings can provide users with understandable risk information.

For healthcare personnel, the warning should be concise. Instead of “Machine-learning classifier detected anomalous feature distribution,” a more useful warning is “Potential phishing site: The displayed hospital name does not match the registered domain. Do not enter your password.”

Explainability thereby becomes part of the defensive intervention itself.

18. Discussion

Pre-2022 evidence demonstrates that machine learning can classify phishing URLs with high apparent accuracy [4-7].

However, three considerations determine whether these systems provide genuine healthcare cybersecurity value.

First, feature efficiency matters.

Gupta et al. demonstrated that a compact nine-feature lexical representation could achieve strong classification performance [5]. This supports the use of low-cost first-line URL screening rather than universally executing expensive webpage analysis.

Second, dataset quality matters as much as model choice.

Hannousse and Yahiouche showed that phishing datasets must account for evolving attacks, runtime requirements and reproducibility [6].

A model achieving 99% accuracy on randomly mixed historical URLs may perform substantially worse on next month’s campaigns.

Third, explainability matters in deployment.

Healthcare security teams must make rapid decisions concerning whether to release, quarantine or investigate URLs. Explanations grounded in observable features can accelerate that process.

The goal should therefore not be to maximize a single benchmark statistic.

The desired system is accurate, fast, temporally robust, interpretable and operationally useful.

19. Limitations

Several limitations should be acknowledged.

First, pre-2022 studies directly combining phishing-URL detection, explainable AI and healthcare-specific datasets were limited.

The present framework therefore integrates three adjacent evidence streams rather than claiming an established healthcare-specific XAI classifier.

Second, reported accuracy values across studies cannot be directly compared because datasets, class distributions and feature sets differ.

Third, phishing strategies evolve rapidly.

Features predictive during one period may weaken as attackers adapt.

Fourth, externally retrieved domain and reputation features may improve discrimination but create latency and dependence on third-party infrastructure.

Fifth, post-hoc explanation does not guarantee model correctness.

Finally, no original classifier was trained because no verified healthcare-specific URL dataset was supplied. Inventing numerical model results would therefore be methodologically inappropriate.

20. Future Research

A strong next-stage study should construct a temporally labeled healthcare phishing dataset.

The design should include verified phishing URLs, legitimate healthcare domains, healthcare-brand impersonation attacks, domain-level train/test separation, chronological validation, Random Forest and gradient-boosting models, interpretable logistic/tree baselines, global and local explanations, processing-time measurement, calibration analysis and evaluation by security analysts.

An especially valuable endpoint would be analyst decision efficiency.

Investigators could randomize security analysts to receive either classification alone or classification plus explanation and compare decision accuracy, investigation time, confidence and agreement.

This would test whether explainability provides measurable operational benefit rather than merely producing visually attractive outputs.

21. Conclusion

Phishing URLs represent a major attack pathway for healthcare organizations because they provide a scalable mechanism for harvesting credentials and delivering malicious content.

Machine-learning research available through 2021 demonstrates that URLs contain substantial predictive information.

Random Forest and related approaches achieved strong performance using lexical and hybrid features, while compact URL-only feature sets showed particular promise for real-time detection.

Nevertheless, high benchmark accuracy alone is insufficient.

Healthcare deployment requires robust datasets, temporal validation, low false-positive burden, efficient feature extraction, understandable predictions and integration with human security operations.

A tiered architecture is therefore recommended.

Fast lexical analysis should provide first-line screening, domain and host characteristics should resolve uncertain cases, and potentially hazardous webpage inspection should be reserved for controlled secondary analysis.

Explainability should operate throughout this process by identifying the observable characteristics that contributed to individual risk predictions.

The most valuable phishing detector for healthcare will consequently not be the model producing the highest isolated accuracy value.

It will be the system capable of identifying previously unseen malicious URLs quickly, reliably and transparently enough for healthcare security teams to act before users disclose credentials or malicious code reaches critical clinical infrastructure.

Declarations

Author Contributions

The authors contributed to conceptualization, evidence synthesis, technical interpretation, framework development, manuscript preparation and critical revision.

Funding

No external funding was received.

Ethical Approval

Not applicable. This study synthesized previously published literature and did not involve human-participant or patient-level data collection.

Consent to Participate

Not applicable.

Data Availability

No novel dataset was generated or analyzed.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Gordon WJ, Wright A, Aiyagari R, Corbo L, Glynn RJ, Kadakia J, et al. Assessment of employee susceptibility to phishing attacks at US health care institutions. JAMA Netw Open. 2019;2(3):e190393. doi:10.1001/jamanetworkopen.2019.0393.
2. Williams CM, Chaturvedi R, Chakravarthy K. Cybersecurity risks in a pandemic. J Med Internet Res. 2020;22(9):e23692. doi:10.2196/23692.
3. He Y, Aliyu A, Evans M, Luo C. Health care cybersecurity challenges and solutions under the climate of COVID-19: scoping review. J Med Internet Res. 2021;23(4):e21747. doi:10.2196/21747.
4. Sahingoz OK, Buber E, Demir O, Diri B. Machine learning based phishing detection from URLs. Expert Syst Appl. 2019;117:345-357.
5. Gupta BB, Yadav K, Razzak I, Psannis K, Castiglione A, Chang X. A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. Comput Commun. 2021;175:47-57. doi:10.1016/j.comcom.2021.04.023.
6. Hannousse A, Yahiouche S. Towards benchmark datasets for machine learning based website phishing detection: an experimental study. Eng Appl Artif Intell. 2021;104:104347. doi:10.1016/j.engappai.2021.104347.
7. Barraclough PA, Fehringer G, Woodward J. Intelligent cyber-phishing detection for online. Comput Secur. 2021;104:102123. doi:10.1016/j.cose.2020.102123.
8. Alhogail A, Alsabih A. Applying machine learning and natural language processing to detect phishing email. Comput Secur. 2021;110:102414. doi:10.1016/j.cose.2021.102414.
9. Chiew KL, Yong KSC, Tan CL. A survey of phishing attacks: their types, vectors and technical approaches. Expert Syst Appl. 2018;106:1-20.
10. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115.
11. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-215. doi:10.1038/s42256-019-0048-x.
12. Jiang JX, Bai G. Evaluation of causes of protected health information breaches. JAMA Intern Med. 2019;179(2):265-267.
13. Gordon WJ, Wright A, Glynn RJ, Kadakia J, Mazzone C, Leinbach E, et al. Evaluation of a mandatory phishing training program for high-risk employees at a US healthcare system. J Am Med Inform Assoc. 2019;26(6):547-552. doi:10.1093/jamia/ocz005.

Table 1. Evidence eligibility framework

Domain	Inclusion criteria	Exclusion criteria
Publication date	≤31 December 2021	2022 onward
Journal quality	Q1 journal in relevant category	Lower-quartile/unranked venues
Technical focus	Phishing URL/site detection, ML, XAI, cybersecurity	Unrelated malicious-software studies
Healthcare evidence	Phishing vulnerability or healthcare cybersecurity	Healthcare papers without cyber relevance
Model evidence	Reproducible ML/AI method with reported evaluation	Unsupported conceptual claims
Explainability	Model interpretability, XAI principles or interpretable ML	Pure performance reporting without interpretability relevance
Study type	Empirical study, systematic/scoping review, methodological article	Non-peer-reviewed commentary

Table 2. Selected pre-2022 Q1 evidence relevant to phishing detection

Study	Approach	Data/features	Principal finding	Relevance
Sahingoz et al. [4]	Multiple ML classifiers	Large legitimate/phishing URL dataset; NLP and hybrid features	RF with NLP features reported 97.98% accuracy	Demonstrated effective URL-only real-time detection
Gupta et al. [5]	Lexical ML classifiers	Nine URL lexical features	RF reported 99.57% accuracy	Demonstrated value of compact low-cost features
Hannousse & Yahiouche [6]	Benchmark + classifier comparison	87 features across URL, content and external domains	Hybrid accuracy 96.61%; selected features 96.83%	Highlighted reproducibility and feature-runtime trade-offs
Barraclough et al. [7]	Intelligent/fuzzy phishing detection	Behavioral/technical phishing characteristics	Demonstrated intelligent automated classification	Supports adaptive detection
Alhogail & Alsabih [8]	ML, NLP and graph-based phishing classification	Phishing email textual structure	Demonstrated value of richer ML representation	Supports multi-layer detection beyond URLs
Chiew et al. [9]	Systematic technical survey	Phishing vectors and detection approaches	Identified multiple complementary detection strategies	Establishes broader context

Table 3. Candidate features for an interpretable healthcare phishing-URL model

Category	Example feature	Potential phishing rationale	Explainability
URL structure	URL length	Attackers may use long strings to obscure destination	High
Hostname	Number of subdomains	Can imitate organizational hierarchy	High
Character structure	Number of digits	Frequently associated with algorithmically generated URLs	High
Symbols	Number of hyphens/@ characters	Can visually manipulate address structure	High
Domain presentation	IP address used as hostname	Avoids conventional domain identity	Very high
Tokens	Presence of login/account/security terms	Credential-harvesting context	High
Brand relationship	Healthcare brand in path but not domain	Potential impersonation	Very high
Domain age	Recently registered domain	Short-lived phishing infrastructure	High
TLS/certificate	Certificate characteristics	Provides additional infrastructure evidence	Moderate
Redirect behavior	Multiple redirect steps	Can obscure final destination	Moderate
Form behavior	External form destination	Credentials may be submitted elsewhere	High
External resources	Abnormal resource-domain ratio	Potential copied webpage	Moderate

Table 4. Candidate classifiers for explainable phishing detection

Model	Predictive capacity	Interpretability	Runtime suitability	Recommended role
Logistic regression	Moderate-high	Very high	Excellent	Baseline
Decision tree	Moderate	Very high	Excellent	Interpretable comparator
Random Forest	High	Moderate	Excellent	Primary candidate
Gradient boosting	High	Moderate	Very good	Primary candidate
Support-vector machine	High	Low-moderate	Good	Secondary candidate
Neural network	Potentially high	Low	Variable	Exploratory
Rule-based score	Moderate	Very high	Excellent	Operational benchmark

Table 5. Recommended evaluation framework

Evaluation component	Recommended metric/design	Reason
Discrimination	AUROC	Overall ranking performance
Phishing detection	Recall/sensitivity	Measures missed malicious URLs
Alert precision	Precision/PPV	Reflects security-team alert burden
Balanced performance	F1-score	Useful with imbalanced data
False positives	False-positive rate	Important for workflow disruption
Probability quality	Calibration/Brier score	Required for risk-based thresholds
Temporal robustness	Later URLs as test set	Detects concept drift
Cross-source robustness	Independent URL source	Tests dataset generalization
Processing cost	ms/URL	Important for gateway deployment
Explanation quality	Analyst agreement/usefulness	Determines operational value

Figure 1. Explainable machine-learning architecture for phishing URL detection in healthcare organizations

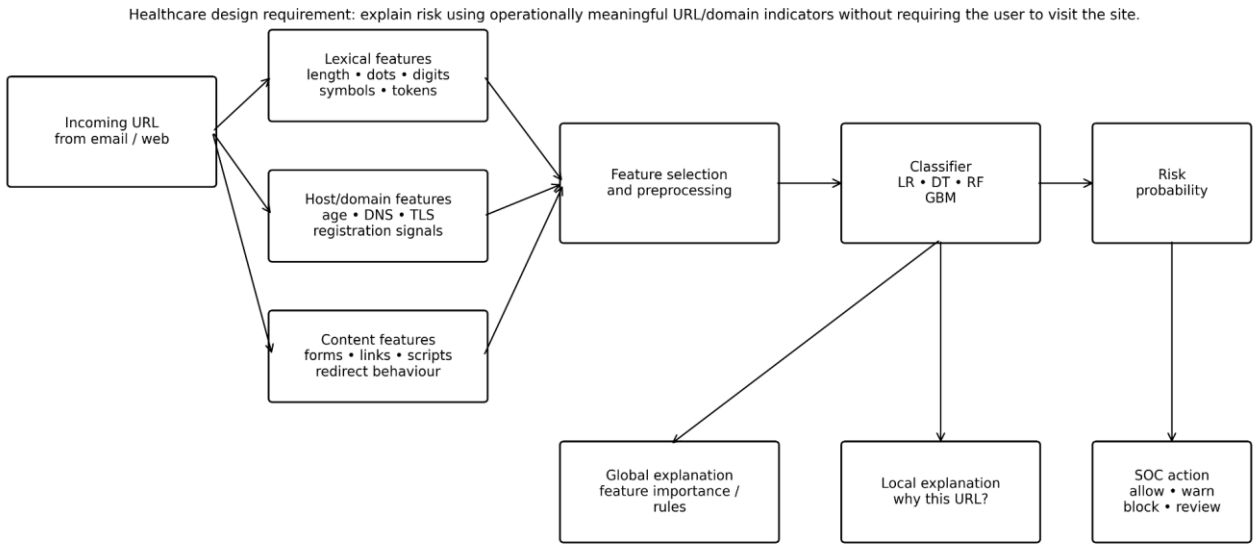
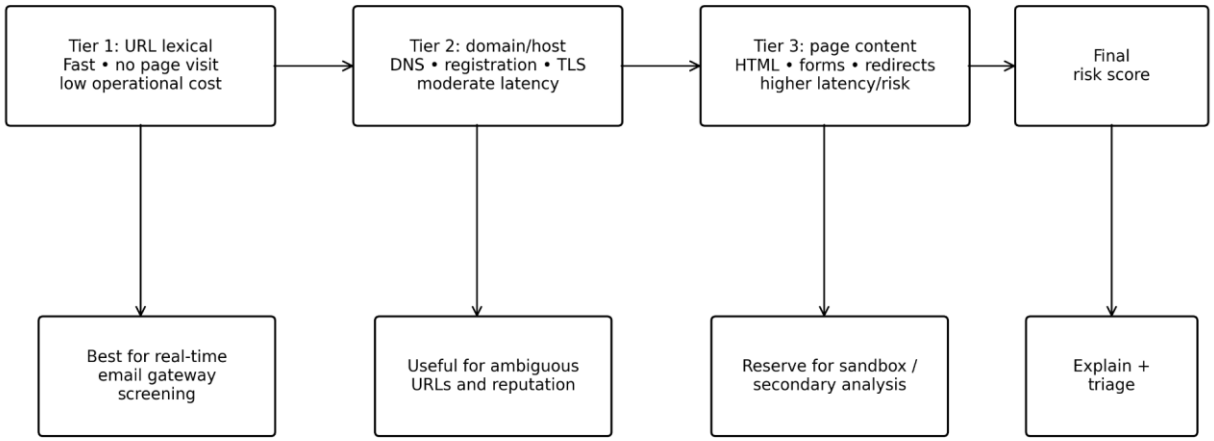


Figure 1. Explainable machine-learning architecture for phishing URL detection in healthcare organizations.

Proposed explainable architecture. URL, domain and optional webpage features enter a machine-learning classifier, while global and local explanatory layers translate predictions into security-operational information.

Figure 2. Tiered feature-extraction strategy balancing detection performance, latency, and explainability



Recommended deployment principle: use inexpensive, interpretable lexical features first; escalate uncertain cases to richer but slower feature sets.

Figure 2. Tiered feature-extraction strategy balancing detection performance, latency, and explainability.

Proposed three-tier architecture. Low-cost lexical features permit rapid first-pass screening, domain and host information increases evidence for ambiguous cases, and webpage analysis is reserved for URLs requiring secondary investigation.

How to cite: Clarke B (2022). Explainable Machine-Learning Models for Detecting Phishing URLs Targeting Healthcare Organizations: A Systematic Technical Review and Deployment Framework. Journal of Multidisciplinary Research and Integrated Sciences, 2(1),