

**Olha Haborets**

PhD, Associate Professor,

Department of Operational and Investigative Activities and Information Security

Donetsk State University of Internal Affairs, Kropyvnytskyi, Ukraine

ORCID: 0000-0001-7791-6795

*olga-gaborets@ukr.net***EXPLAINABLE MACHINE LEARNING FOR RELIABLE PHISHING WEBSITE DETECTION: A TWO-DATASET EXPERIMENTAL STUDY**

Abstract. Phishing website detection is frequently presented as a classification task in which high predictive performance is treated as sufficient evidence of operational usefulness. This approach is incomplete in law-enforcement and cybersecurity settings, where a model output must be interpretable, independently verifiable and separated from conclusions about authorship, intent or legal qualification. The study develops and evaluates a human-centred explainable artificial intelligence framework combining supervised classification with structured analyst-oriented explanations. Five classifiers were assessed on the UCI Phishing Websites benchmark, and a second experiment used the larger PhiUSIIL dataset. Phishing was consistently designated as the positive class. Extra Trees achieved test accuracy of 0.9774, precision of 0.9784, recall of 0.9704, F1-score of 0.9744 and ROC-AUC of 0.9948. Stratified bootstrap intervals quantified test uncertainty: the 95% interval was 0.9670–0.9812 for F1 and 0.9915–0.9976 for ROC-AUC. The Brier score was 0.0187 and ten-bin expected calibration error was 0.0089. TreeSHAP analysis of 400 hold-out observations identified certificate state and anchor behaviour as the dominant global contributors. Across five fitted folds, the mean pairwise Jaccard overlap of the ten leading features was 0.891 (SD 0.094). In a mode-replacement fidelity test, replacing the three leading SHAP features by their training modes reduced F1 from 0.9744 to 0.3543, whereas random three-feature replacement yielded mean F1 of 0.9478. Near-perfect PhiUSIIL performance was not interpreted as deployment readiness: grouped validation, an explicitly defined 18-feature lexical ablation and duplicate analysis exposed collection-specific separability; the lexical-only grouped Extra Trees model retained mean F1 of 0.9970. The proposed Phishing Threat Explanation Card converts a prediction into an auditable record of the score, threshold, leading indicators, verification actions, uncertainty and analyst disposition. The contribution is empirical and procedural: it reports a documented and auditable comparison, quantifies predictive uncertainty, calibration and explanation fidelity, diagnoses inflated benchmark performance and defines safeguards for human oversight.

Keywords: phishing; explainable artificial intelligence; machine learning; phishing website detection; SHAP; decision support; law enforcement; cyber threat intelligence.

INTRODUCTION

Phishing remains one of the most adaptive forms of social engineering because it exploits both technical weaknesses and predictable human responses. A malicious website can imitate a trusted service, create artificial urgency, request credentials or payment data, and disappear before conventional blocklists have been updated. The defensive problem is therefore time-sensitive: analysts must prioritize suspicious objects quickly while preserving enough context for verification and subsequent action. Automated detection can reduce the volume of manual screening, but an unexplained label is a weak basis for a professional decision.

Machine-learning systems are well suited to identifying combinations of lexical, structural and infrastructure indicators that are difficult to express as static rules. Research has reported strong performance for tree ensembles, boosting algorithms and hybrid models [1-7]. Yet a recurring methodological weakness is the use of random train–test partitions from a single curated dataset as the main evidence of generalization. When a dataset is assembled from different sources or periods, classifiers may learn collection artefacts rather than stable characteristics of phishing. High scores can consequently coexist with poor resilience to new campaigns, new top-level domains, cloud hosting or deliberate feature manipulation.



Explainable artificial intelligence (XAI) has been proposed as a response to the opacity of high-performing models. LIME approximates a model locally around an observation [8], whereas SHAP assigns additive feature contributions based on Shapley-value principles [9]. These techniques can show which variables influenced a prediction, but they do not prove that a feature caused the malicious behaviour or that the explanation is stable under correlated inputs. In cybersecurity, such limitations are especially consequential because an adversary can manipulate observable indicators and because false positives may interrupt legitimate activity.

The problem has an additional evidentiary dimension in law-enforcement practice. A phishing score can support triage, preservation and prioritization, but it cannot by itself establish the identity of an offender, criminal intent, victimization or the legal status of the examined resource. The model and its explanation should therefore be embedded in a controlled workflow that records the input, model version, threshold, explanation, independent checks and analyst decision. This distinction between predictive assistance and evidentiary conclusion guides the present study.

The purpose of the research is to develop and empirically evaluate an explainable machine-learning framework for phishing website detection and to translate its outputs into a human-centred decision-support instrument. The objectives are to compare five classifiers under a documented protocol; evaluate errors using phishing as the positive class; examine robustness on a second dataset and under feature ablation; quantify the stability of global SHAP rankings; define an analyst-readable explanation format; and formulate safeguards for deployment, monitoring and audit. The study asks which evaluated model provides the most balanced benchmark performance, how stable its leading explanations are across training folds, whether very high results persist under restricted feature domains, and how feature attribution can be converted into verifiable analytical actions.

The original contribution lies in combining three elements that are often separated in the literature: an explicit and consistently class-specific evaluation, a robustness-oriented interpretation of a near-perfect contemporary benchmark, and a decision-support card designed for analysts. The study does not claim that a benchmark classifier is ready for autonomous blocking or evidentiary use. Instead, it proposes a transparent architecture in which automated risk assessment remains subordinate to human verification and ongoing validation.

Analysis of recent research and publications. Phishing detection research encompasses URL analysis, webpage-content analysis, domain and certificate intelligence, email-language modelling and multimodal approaches. Sahingoz et al. demonstrated the value of URL-derived features and machine-learning classifiers for real-time detection [1]. Zhang et al. surveyed explainable-AI applications across cybersecurity and systematized their methods, datasets and research challenges [2]. Safi and Singh reviewed contemporary phishing-detection methods and highlighted continuing problems of dataset comparability, concept drift and reproducibility [3]. These studies establish a mature predictive field, but also show that reported results are strongly conditioned by the selected corpus and experimental design.

Explainability has become a distinct research direction. Calzarossa, Giudici and Zieni studied explainable machine learning for identifying phishing-relevant features [4]. Zieni, Massari and Calzarossa surveyed phishing-website detection and organized the main data sources, feature families and classification approaches [5]. Capuano et al. surveyed XAI in cybersecurity and underlined the importance of matching an explanation technique to both the model and the stakeholder [6]. Rudin's critique of post-hoc explanation further warns that an apparently plausible explanation of a black-box system should not automatically be treated as a faithful account of its decision process [10].

Recent phishing-specific studies combine ensemble classifiers and SHAP. Uddin et al. reported an efficient boosting-based phishing-site detector with feature attribution [7]. Kehkashan et al. proposed an explainable phishing website detection approach for cyber infrastructure [11]. Prasad and Chandra introduced the PhiUSIIL dataset and a similarity-index-based detection approach [12]. These contributions expand both scale and interpretability, but operational claims still require careful examination of the dataset construction, feature availability at decision time and the possibility that a few variables encode source-specific distinctions.

Natural-language explanation is an emerging extension. EXPLICATE combines phishing detection with large-language-model-assisted interpretation, but it is presently a preprint and should be understood as preliminary evidence rather than a peer-reviewed operational standard [13]. The use of generative explanations may improve readability, yet it introduces additional risks: unsupported statements, non-determinism, disclosure of sensitive input and a mismatch between fluent language and actual model evidence. A safer design first creates a structured, machine-readable explanation record and only then permits controlled natural-language rendering.

Foundational XAI methods remain central. LIME provides local surrogate explanations by perturbing inputs and fitting a simple model near a selected case [8]. SHAP offers a unified additive framework and, for tree ensembles, computationally efficient implementations [9]. Both are useful when their scope is explicit. Global importance addresses recurring patterns across a dataset; local attribution addresses one prediction. Neither



replaces independent technical inspection. Correlated variables, background-sample selection and representation choices may substantially change attribution rankings.

The literature therefore reveals a practical research gap. Many studies report accuracy and an explanation plot, but fewer define what an analyst should verify, how uncertainty should be recorded, or how benchmark artefacts should constrain deployment claims. The present work addresses this gap through a documented two-dataset assessment and an explanation card that links each influential indicator to a verification task and records the final human disposition.

THEORETICAL FOUNDATIONS OF EXPLAINABLE PHISHING DETECTION

A phishing website is not defined by one universal technical property. Attackers may use newly registered domains, compromised legitimate hosts, URL shorteners, valid TLS certificates, cloud platforms, copied page elements and redirection chains. Some features that were once suspicious – such as HTTPS availability – are now common in both legitimate and malicious resources. A detector must consequently integrate several weak indicators and be reassessed as attacker behaviour and platform infrastructure evolve.

For analytical purposes, observable variables can be grouped into four domains. Lexical URL features describe length, symbols, tokens, subdomains and character composition. Host and domain features describe registration, DNS, certificate and reputation context. Webpage features characterize forms, anchors, scripts, external resources and visual or semantic similarity. Contextual features connect the object to campaign timing, referrers, messages, victim reports and related infrastructure. The first three groups support automated classification; the fourth is often essential for professional interpretation but absent from public benchmark datasets.

Adversarial conditions change the meaning of model performance. A conventional classifier assumes that future observations resemble training data, whereas a phishing operator intentionally modifies features to evade detection. Robustness must therefore be considered at three levels: statistical robustness to alternative samples, temporal robustness to new campaigns, and adversarial robustness to feature manipulation. The experiments in this article address the first level and partially probe feature-domain dependence; temporal and adversarial validation remain future requirements.

An explanation is useful only in relation to a defined user and task. A data scientist may need global feature dependence and error slices; a security operator needs concise reasons and verification actions; a supervisor needs threshold rationale and performance monitoring; an investigator needs provenance and preservation information. The same SHAP vector cannot be presented identically to all stakeholders. Human-centred XAI therefore begins with the decision process rather than the visualization technique.

Feature attribution answers a limited counterfactual-style question: how the selected model's output differs from a chosen baseline when feature information is introduced according to a formal rule. It does not establish legal relevance, causation or the truth of a feature value. For instance, a high contribution from certificate state indicates reliance by the model, not proof that the certificate is fraudulent. The analyst must examine certificate details, collection time, redirects and hosting context through independent tools.

Explanation quality has at least five dimensions: fidelity to the model, stability under small non-material changes, comprehensibility for the intended user, actionability for verification, and completeness of provenance [20]. A visually attractive explanation may fail if it is unstable or if its indicators cannot be checked. Conversely, a technically faithful vector may fail if it is too complex for rapid triage. The proposed framework uses a short ranked list plus explicit verification prompts to balance these requirements.

Table 1

Feature domains, analytical meaning and principal limitations

Feature domain	Examples	Analytical use	Principal limitation
URL lexical	Length, digits, symbols, subdomains, tokens	Fast pre-navigation screening	Easily manipulated; may penalize legitimate complex URLs
Domain/host	Age, DNS, registrar, certificate, reputation	Infrastructure verification and linkage	Availability and timing vary; compromised hosts appear legitimate
Webpage structure	Forms, anchors, scripts, external resources	Detect imitation and credential collection	Requires safe content acquisition; dynamic pages change



Similarity/content	Brand text, visual similarity, copyright or social links	Identify impersonation patterns	May encode collection-source artefacts
Context/campaign	Message source, redirects, victims, related domains	Connect technical object to incident or campaign	Usually absent from benchmark datasets

MATERIALS AND METHODS

The primary experiment used the UCI Phishing Websites dataset [14]. It contains 11,055 observations and 30 predictors, with 6,157 legitimate and 4,898 phishing cases. The variables are pre-engineered indicators typically encoded as -1 , 0 and 1 . Although the dataset is widely used and supports reproducibility, its age limits direct claims about current phishing infrastructure. It was therefore treated as a transparent baseline rather than a representation of all contemporary attacks.

The complementary experiment used the PhiUSIIL Phishing URL dataset distributed through the UCI Machine Learning Repository [16] and described by Prasad and Chandra [12]. It contains 235,795 observations, 54 features and a label, with 134,850 legitimate and 100,945 phishing instances. The dataset integrates URL, source-code and similarity-related variables. According to its documentation, label 1 denotes legitimate and label 0 denotes phishing. The larger sample and different feature construction make it useful for a complementary robustness assessment, but not for direct cross-dataset transfer because the two feature spaces are not equivalent.

A data audit preceded modelling. Feature types, missingness, class counts, duplicate observations and feature-label associations were inspected. In PhiUSIIL, several variables showed very strong univariate association with the label, including URLSimilarityIndex, HasSocialNet and HasCopyrightInfo. Duplicate URLs and duplicate numeric rows were also present. These facts do not invalidate the dataset, but they make random-split performance vulnerable to collection-specific separability. Consequently, near-perfect scores were interpreted cautiously and supplemented with grouped folds and lexical-only ablation.

Duplicate URLs and duplicate numeric feature vectors were quantified separately as part of the data audit. Under domain-grouped validation, observations assigned the same domain-group identifier were kept within one fold, which prevents exact URL duplicates sharing that identifier from crossing between training and validation partitions. Duplicate numeric profiles were not treated as independent evidence of generalization because they may represent either repeated collection records or genuinely recurring webpage templates. Their presence is therefore reported as a residual source of dependence and a limitation of the collection.

The positive class was defined consistently as phishing throughout preprocessing, model fitting and evaluation. Precision therefore measures the proportion of flagged websites that are phishing, whereas recall measures the proportion of phishing websites detected. The recoding rules and confusion-matrix orientation were verified before calculating all class-specific metrics.

For the UCI benchmark, the data were divided into 80% training and 20% testing subsets by stratified random sampling with seed 42. Five-fold stratified cross-validation with shuffling and the same seed was performed on the training data. Pre-processing was encapsulated in model pipelines to prevent leakage: StandardScaler was fitted within the logistic-regression pipeline, while tree models used the numeric indicators directly. The test subset was accessed only after the model specifications in Table 3 had been fixed; no test-set optimization was performed.

Five algorithms were evaluated: logistic regression, a regularized linear baseline; a constrained decision tree; a balanced random forest; an Extra Trees ensemble; and histogram gradient boosting. Logistic regression provides a useful low-complexity reference. The decision tree offers directly inspectable rules but may have higher variance. Random forest reduces variance through bootstrapped tree aggregation. Extra Trees additionally randomizes split thresholds and can perform strongly on heterogeneous indicators. Histogram gradient boosting models sequential residual structure and offers an efficient nonlinear comparator [17,18].

The PhiUSIIL experiment applied an analogous stratified protocol after excluding direct identifiers and the target label. Because several models reached effectively perfect random-split performance, two robustness checks were added. First, observations were allocated to five folds using a deterministic domain-based grouping so that related hostnames were less likely to appear in both training and validation subsets. Second, an Extra Trees model was trained on 18 URL-lexical variables only, excluding content, reputation and similarity variables that may encode the acquisition process. These checks do not constitute temporal external validation, but they are more informative than presenting a single random split without qualification.

For the grouped robustness check, each observation was assigned a deterministic domain-group identifier derived from the URL and domain information supplied in PhiUSIIL. All records with the same identifier were



allocated to the same fold, so no group identifier appeared simultaneously in the corresponding training and validation partitions. This procedure reduces direct hostname-related overlap but does not establish temporal or source independence: similar lexical patterns, duplicated numeric profiles and collection-specific artefacts may remain across groups. The grouped results are therefore interpreted as a stronger within-dataset diagnostic than a random split, not as external validation.

The 18-variable lexical subset comprised URLLength, DomainLength, IsDomainIP, TLDLength, NoOfSubDomain, HasObfuscation, NoOfObfuscatedChar, ObfuscationRatio, NoOfLettersInURL, LetterRatioInURL, NoOfDigitsInURL, DigitRatioInURL, NoOfEqualsInURL, NoOfQMarkInURL, NoOfAmpersandInURL, NoOfOtherSpecialCharsInURL, SpacialCharRatioInURL and IsHTTPS. The original dataset spellings, including Digits and Spacial, are retained here to keep the analysis traceable to the source data.

No synthetic oversampling was used because both datasets contain substantial representation of each class. Random seeds were fixed at 42. The UCI reproduction used Python 3.12.13, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, SHAP 0.52.0 and Matplotlib 3.10.8. The official UCI ARFF file was downloaded on 6 September 2026; its SHA-256 checksum was 3bcf93b1612306537ee2737751f1bac185b2a77886ab376644c3bf0c7f7cb3e6. The manuscript documents the data sources, class recoding, split strategy, model specifications, metric definitions and XAI sampling rules.

Accuracy was reported for comparability, but phishing-specific precision, recall and F1-score were treated as primary. For phishing as the positive class, precision = $TP/(TP+FP)$, recall = $TP/(TP+FN)$, and F1 is their harmonic mean. ROC-AUC evaluates ranking across thresholds but can remain high when a selected operating threshold produces an undesirable false-positive burden [19]; the confusion matrix was therefore retained as an operational complement. For the UCI experiment, five-fold means and standard deviations describe resampling variability. PhiUSIIL grouped results are reported as fold means and are interpreted descriptively rather than as estimates of future temporal performance.

Threshold choice is a policy decision. Lowering the threshold generally increases phishing recall while increasing legitimate-site alerts. In a high-risk triage queue, this trade-off may be acceptable; for automated blocking, the cost of false positives is more severe. The proposed model therefore separates the continuous risk score from the action threshold and requires threshold documentation, periodic calibration and differentiated actions such as monitor, request analyst review or temporarily contain.

No post-hoc probability recalibration was performed in the reported experiment. The Brier score, expected calibration error and reliability diagram were used to assess the empirical calibration of the original Extra Trees probability estimates on the UCI hold-out set. Expected calibration error was calculated using ten quantile-based bins as the weighted mean of the absolute differences between mean predicted probabilities and observed phishing frequencies. Accordingly, the reported probabilities should be interpreted as model-specific prediction scores rather than universally calibrated estimates of real-world phishing risk.

Table 2

Evaluation design and interpretive safeguards

Element	UCI Phishing Websites	PhiUSIIL robustness experiment
Positive class	Phishing (original label -1 / recoded 1)	Phishing (original label 0 / recoded 1)
Scale	11,055 cases; 30 predictors	235,795 cases; 54 supplied features
Primary split	Stratified 80/20; seed 42	Stratified split plus grouped five-fold checks
Cross-validation	Five-fold stratified on training set	Domain-grouped five-fold validation
Ablation	Not required for primary comparison	18 URL-lexical variables only
Main caution	Dataset age and concept drift	Strong source-specific separability and duplicates

Table 3

Fixed model specifications for the UCI experiment

Classifier	Fixed specification
Logistic regression	StandardScaler; C=1.0; lbfgs; maximum 2,000 iterations
Decision tree	Gini; maximum depth 10; minimum leaf size 2; balanced class weights
Random forest	250 trees; square-root feature sampling; balanced class weights
Extra Trees	250 trees; square-root feature sampling; balanced class weights
Histogram gradient boosting	200 iterations; learning rate 0.08; 31 leaf nodes; L2=0.1



TreeSHAP was applied to the fitted Extra Trees model to explain the phishing-class output [9]. A reproducible random sample of 400 hold-out observations (seed 42) was used for the global analysis; features were ranked by mean absolute SHAP value. A correctly classified phishing observation with the highest predicted score was selected for the local example. Explanation stability was assessed by fitting an Extra Trees model in each of five stratified training folds, calculating SHAP values for 100 validation observations per fold and comparing the top-ten feature sets using pairwise Jaccard overlap. This procedure evaluates ranking stability across resampled training sets; it does not establish causal validity.

The reported TreeSHAP analysis explains the phishing-class output of the fitted Extra Trees classifier. The 400-observation hold-out sample was used only to aggregate feature contributions and was not used for model fitting, threshold selection or probability assessment. The class mapping, feature order, software version and sampling seed were fixed for all reported explanations. Because SHAP values depend on the explainer configuration and reference distribution, the executable materials should preserve the exact explainer arguments and expected values; the attributions reported here are consequently interpreted as model- and configuration-specific rather than intrinsic properties of phishing.

Explanation fidelity was assessed through a mode-replacement perturbation test. The leading global SHAP features were successively replaced in the hold-out set by their training-set modes, and F1 and ROC-AUC were recalculated without retraining the classifier. The same intervention was repeated for 100 randomly selected feature subsets at each replacement size. A substantially larger performance decrease after SHAP-ranked replacement than after random replacement supports the faithfulness of the ranking to the fitted model. The procedure is diagnostic rather than causal: replacing features independently may generate uncommon combinations and does not measure the effect of a feasible real-world intervention.

The study adopts the governance principles of validity, reliability, transparency, accountability and human oversight reflected in the NIST AI Risk Management Framework [20]. These principles are operationalized through model-version logging, access control, input and output preservation, threshold approval, error monitoring, explanation review and escalation. The model may prioritize an object; an authorized analyst remains responsible for verification and disposition.

Ethical safeguards are also necessary. Public benchmark records were used, and no personal communications were processed. A deployment pipeline should minimize retention of credentials, tokens, personal data and page content; isolate active content; and preserve only information required by the lawful analytical purpose. Explanation interfaces must avoid exposing sensitive features to untrusted users in ways that make evasion easier.

RESEARCH RESULTS

The Extra Trees classifier produced the highest phishing F1-score, 0.9744, and test accuracy, 0.9774. Its phishing precision was 0.9784 and recall was 0.9704. The random forest was a close second by F1 and achieved the highest ROC-AUC, 0.9967. The difference between these two ensemble models is small enough that operational considerations – probability calibration, inference cost, explanation stability and maintenance – should influence selection. Logistic regression remained a credible baseline but had lower recall and F1, indicating that nonlinear interactions contribute materially to performance.

Table 4

Classifier performance on the UCI hold-out set (phishing is the positive class)

Classifier	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic regression	0.9281	0.9344	0.9010	0.9174	0.9785
Decision tree	0.9512	0.9308	0.9612	0.9458	0.9857
Random forest	0.9756	0.9783	0.9663	0.9723	0.9967
Extra Trees	0.9774	0.9784	0.9704	0.9744	0.9948
Histogram gradient boosting	0.9711	0.9731	0.9612	0.9671	0.9962

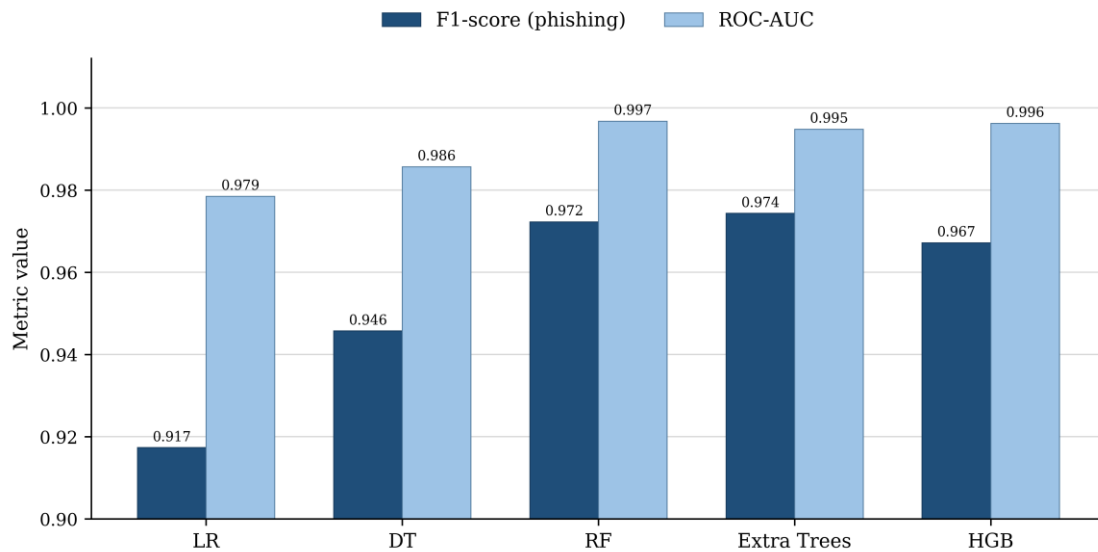


Figure 1. F1-score and ROC-AUC of the five classifiers on the UCI hold-out set

For Extra Trees, 951 phishing websites were correctly detected and 29 phishing websites were missed. Among legitimate websites, 1,210 were correctly classified and 21 were falsely flagged as phishing. Thus, the operationally relevant false-negative count is 29 and the false-positive count is 21. This interpretation follows directly from the class definition. The error distribution is favourable for a benchmark, but even a 2.96% phishing miss rate could be consequential at scale.

Case-level forensic interpretation of the 29 false negatives and 21 false positives is constrained by the UCI release format: the benchmark provides engineered indicators but not the original URLs, page captures or acquisition context. The study therefore reports the error counts and feature-space behaviour without attributing individual errors to verified real-world causes. A prospective evaluation should retain raw artefacts and analyse error patterns by domain type, certificate state, anchor behaviour, language and collection channel.

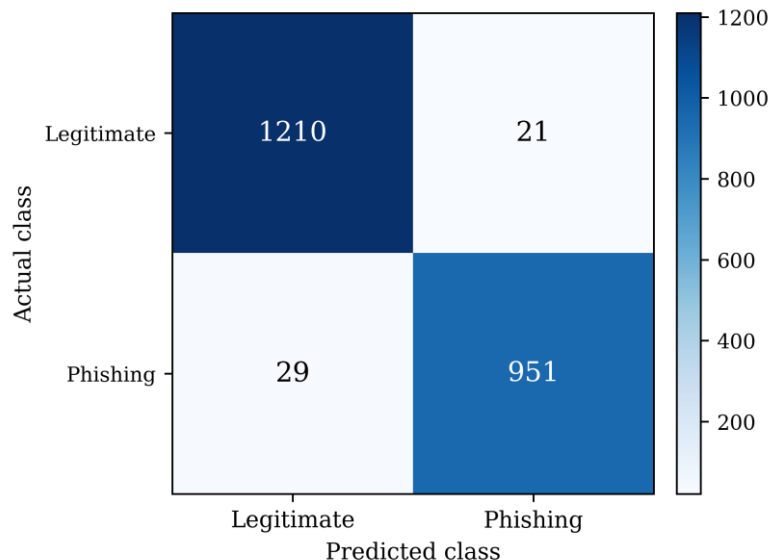


Figure 2. Confusion matrix for Extra Trees on the UCI hold-out set (rows: actual; columns: predicted)

For Extra Trees, five-fold cross-validation produced mean accuracy of 0.9686 (SD 0.0021), precision of 0.9643 (SD 0.0056), recall of 0.9648 (SD 0.0058), F1 of 0.9645 (SD 0.0024) and ROC-AUC of 0.9896 (SD 0.0026). The hold-out score is slightly higher than the training-fold mean but remains within a plausible range. Because all observations originate from the same historical corpus, this result demonstrates internal reproducibility rather than temporal generalization.

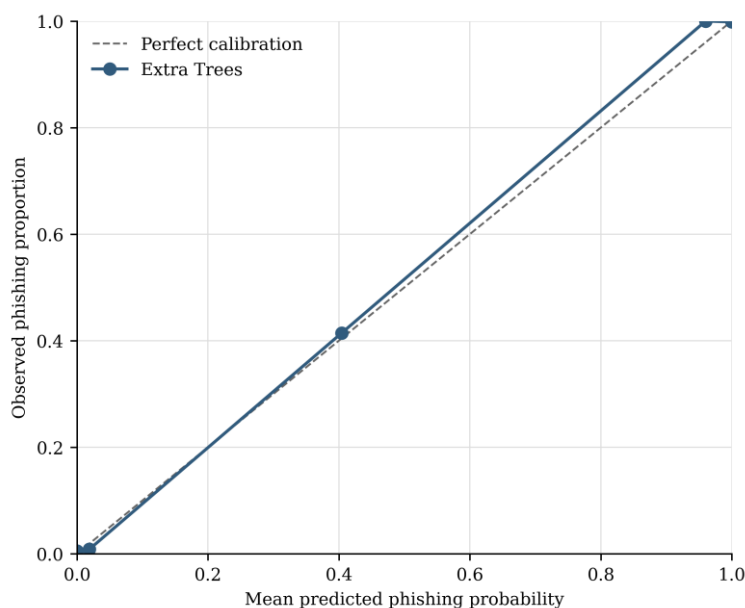


Figure 3. Reliability diagram for Extra Trees on the UCI hold-out set

The mode-replacement fidelity test provided a direct diagnostic of whether the SHAP ranking identifies features on which the fitted model materially depends. Replacing only `SSLfinal_State` by its training mode reduced F1 from 0.9744 to 0.8878. Replacing the three leading features reduced F1 to 0.3543 and ROC-AUC to 0.8711; replacing the five leading features reduced F1 to 0.1389 and ROC-AUC to 0.7682. By comparison, replacement of three randomly selected features yielded mean F1 of 0.9478 (SD 0.0376) across 100 repetitions. The pronounced differential supports ranking fidelity for this fitted model, while the artificial nature of independent mode replacement precludes causal interpretation.

A 2,000-replicate class-stratified bootstrap quantified uncertainty in the Extra Trees hold-out results. The 95% percentile intervals were 0.9711–0.9833 for accuracy, 0.9694–0.9866 for precision, 0.9592–0.9806 for recall, 0.9670–0.9812 for F1 and 0.9915–0.9976 for ROC-AUC. These intervals describe sampling uncertainty within the UCI test set; they do not account for temporal or cross-source distribution shift.

The original, non-recalibrated Extra Trees probabilities produced a Brier score of 0.0187 and a ten-quantile-bin expected calibration error of 0.0089. These are low aggregate errors within this hold-out sample, but neither statistic establishes calibration under a different prevalence or collection process. The reliability curve also contains local deviations where relatively few distinct probabilities are produced by the ensemble. The output should therefore be treated as a model-specific prediction score, not as an invariant estimate of real-world phishing risk; prospective recalibration would be required after deployment conditions are defined.

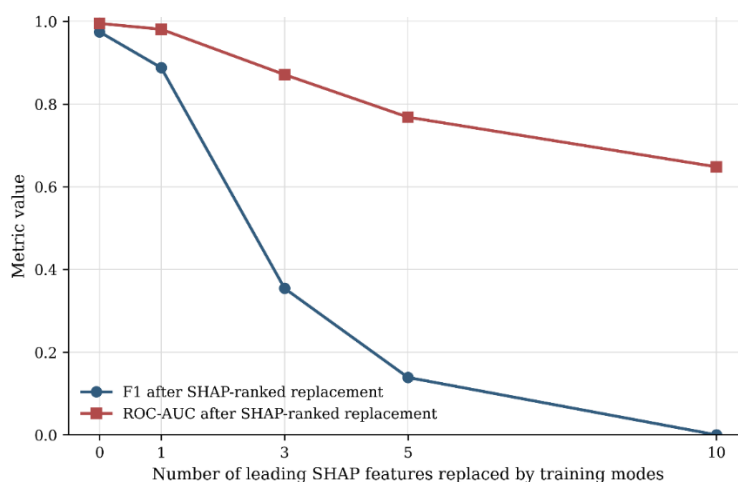


Figure 4. Mode-replacement fidelity test for the global SHAP ranking



On a conventional stratified random split of PhiUSIIL, the decision tree, random forest and histogram gradient boosting classifiers reached 1.0000 to the displayed precision; Extra Trees made one error and logistic regression made four. Reporting these results without qualification would suggest that the task is essentially solved. The data audit and robustness analysis support a more cautious conclusion: the dataset contains highly discriminative features and collection patterns that make its two classes unusually separable.

Domain-grouped five-fold validation using all numeric features remained extremely high, with mean accuracy of 0.999979 and mean phishing F1 of 0.999975. Grouping therefore removed little of the predictive separation. When the feature space was restricted to 18 URL-lexical variables, however, performance decreased measurably: mean accuracy was 0.997455, phishing precision 0.998879, recall 0.995188, F1 0.997030 and ROC-AUC 0.998332. The model retained substantial signal without content and similarity features, but the change confirms that feature-domain choice affects apparent difficulty.

The result should be interpreted as evidence about this collection, not as proof of universal performance. URLSimilarityIndex had an absolute association with the phishing label of approximately 0.860, HasSocialNet approximately 0.784 and HasCopyrightInfo approximately 0.743. Such variables may be substantively meaningful, but their strength can also reflect the way legitimate and phishing pages were sourced or processed. In addition, 425 duplicate URLs and 808 duplicate numeric feature rows were identified. Random splitting can distribute similar records across training and test subsets, thereby reducing the independence of evaluation.

The two datasets do not support direct model transfer because their engineered feature definitions differ. The PhiUSIIL experiment is therefore a complementary within-dataset robustness assessment, not external validation or a direct replication of the UCI classifier. A valid cross-source evaluation would require one versioned feature extractor applied prospectively to raw URLs and webpages from independently sourced, time-separated feeds.

Table 5

PhiUSIIL robustness results for Extra Trees

Evaluation setting	Accuracy	Precision	Recall	F1	ROC-AUC
All supplied numeric features; domain-grouped 5-fold	0.999979	0.999990	0.999961	0.999975	≈ 1.000000
URL-lexical features only; domain-grouped 5-fold	0.997455	0.998879	0.995188	0.997030	0.998332

In the 400-observation TreeSHAP sample, SSLfinal_State had the largest mean absolute contribution (0.1823), followed by URL_of_Anchor (0.1520), web_traffic (0.0512), Prefix_Suffix (0.0475) and having_Sub_Domain (0.0428). These variables are analytically plausible: they describe infrastructure trust, inconsistencies between displayed content and navigation targets, and URL structure. Nevertheless, the explanation must preserve the original encoded definition. A generic label such as ‘SSL risk’ is insufficient if the feature combines certificate issuer, validity period and HTTPS state.

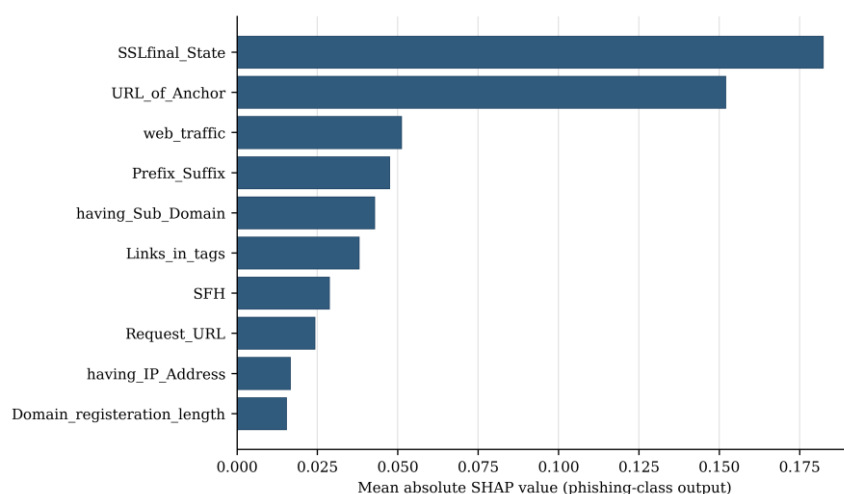


Figure 5. Global TreeSHAP importance for the Extra Trees model (400 hold-out observations)

A local explanation should be presented as a balance of risk-increasing and risk-reducing contributions relative to the explainer baseline. For a flagged website, the interface may report that abnormal anchor behaviour



and domain-related indicators increased the score, while a stable DNS record reduced it. The analyst must then verify the actual anchor destinations, certificate chain, registration timeline and redirect sequence. If a decisive feature is unavailable or changes after collection, the explanation card should state that limitation rather than silently presenting stale evidence.

Explanation stability must be monitored. If two semantically equivalent URLs receive materially different explanations because of minor formatting, the model may be relying on brittle artefacts. Stability tests should include removal of tracking parameters, normalization of case where appropriate, expansion of URL shorteners in a controlled environment and comparison across repeated acquisitions. A stable prediction with unstable reasons still warrants review because the attribution layer may not reliably support action.

Across the five independently fitted folds, the mean pairwise Jaccard overlap of the ten leading SHAP features was 0.891 (SD 0.094). The overlap is high enough to support a stable global narrative for this dataset, but it is not perfect; therefore, ranked indicators should be regenerated after retraining and versioned with the model rather than treated as permanent properties of phishing.

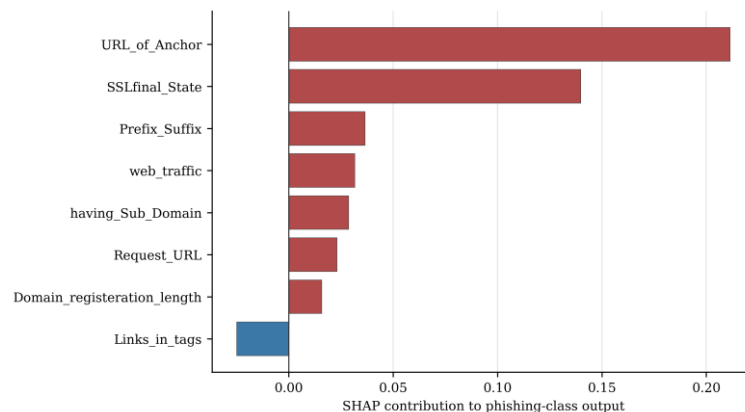


Figure 6. Local TreeSHAP explanation for a correctly classified phishing observation (test-row index 3588; predicted score 1.000)

For the local case in Figure 6, abnormal anchor behaviour (SHAP +0.2115) and certificate state (+0.1398) were the dominant risk-increasing contributions. Prefix/suffix use (+0.0366), low web-traffic state (+0.0316), subdomain structure (+0.0287) and external-request behaviour (+0.0232) also increased the phishing output, whereas Links_in_tags reduced it (−0.0249). The example demonstrates why an explanation must include both risk-increasing and protective evidence and why each encoded indicator must be verified against the original page, certificate and acquisition record.

The proposed Phishing Threat Explanation Card is an auditable interface between the classifier and the analyst. It is deliberately structured rather than generated as unconstrained prose. Each field has a defined provenance and purpose. The card records the object identifier and acquisition time; model and feature-extractor versions; predicted phishing score and decision threshold; predicted class; ranked risk-increasing and protective indicators; verification actions; uncertainty and data-quality notes; linked objects; analyst disposition; and the identity and time of review.

Table 6

Structure of the Phishing Threat Explanation Card

Card field	Required content	Analytical function
Object and provenance	Canonicalized URL/hash, source, collection time	Preserves identity and acquisition context
Prediction	Model score, threshold, class, model version	Separates score from policy decision
Leading indicators	Top risk-increasing and protective contributions	Shows what influenced this model output
Verification actions	DNS, certificate, redirect, content and reputation checks	Converts attribution into independent tasks
Uncertainty	Missing values, drift warning, unstable explanation, out-of-range features	Prevents false certainty
Analyst disposition	Escalate, monitor, close, contain; rationale and reviewer	Maintains human accountability and audit trail



Table 7

Completed explanation card for UCI test-row 3588

Card field	Recorded value
Object and provenance	UCI benchmark test-row 3588; raw URL unavailable in the public engineered-feature record; split seed 42
Prediction	Phishing; model score 1.000; decision threshold 0.500; Extra Trees, 250 trees
Leading risk indicators	URL_of_Anchor = -1 (SHAP +0.2115); SSLfinal_State = 0 (+0.1398); Prefix_Suffix = -1 (+0.0366); web_traffic = -1 (+0.0316)
Protective indicator	Links_in_tags = 1 (SHAP -0.0249)
Required verification	Inspect anchor destinations; validate the certificate chain and acquisition time; review domain syntax, redirects and traffic/reputation evidence
Uncertainty and limitation	Score is benchmark-specific; raw page and URL are absent; encoded features cannot establish causation, authorship, intent or legal qualification
Analyst disposition	Escalate for independent technical verification; do not autonomously block or characterize as evidence

The completed card illustrates an important boundary of benchmark-based XAI. Because the UCI record contains engineered indicators rather than the original URL and captured webpage, the contributions can be reproduced computationally but cannot be independently verified against the underlying object. The correct disposition is therefore escalation for technical verification, not an autonomous enforcement action. In a live system, the canonical URL, acquisition timestamp, page hash and protected evidence location would be mandatory fields.

OPERATIONAL WORKFLOW AND SAFEGUARDS

The proposed workflow comprises acquisition, explanation, verification, disposition and feedback. At acquisition, the system preserves the original URL or safely captured webpage, records provenance and isolates active content before extracting time-available indicators. It then returns a prediction score, which an approved policy maps to a preliminary action category. The explanation stage generates ranked risk-increasing and protective contributions together with data-quality warnings. During verification, the analyst checks the relevant certificate, domain, redirect, form, anchor, hosting and reputation evidence. The disposition record explicitly separates automated output, analyst inference and independently acquired facts. Confirmed errors and novel patterns are retained as monitoring data for controlled review and retraining.

Four safeguards govern this workflow. First, role separation distinguishes development, threshold approval, case verification and supervisory review. Second, provenance records the feature-extractor version, model version, explanation method, reference data and inference time. Third, abstention routes unsupported, incomplete, unstable or out-of-range inputs to manual review instead of forcing a binary conclusion. Fourth, human-factor and security evaluation tests comprehension, appropriate reliance, resistance to automation bias, adversarial probing, feedback poisoning and disclosure that could facilitate evasion.

A responsible transition to operational use begins with a narrowly specified intended purpose, such as prioritizing URLs from incident reports or supporting a secure review queue. The intended use determines the relevant population, permissible features, acceptable latency and consequences of each error type. Benchmark performance cannot be transferred to that population without prospective validation.

The initial pilot should run in shadow mode: predictions and explanation cards are recorded but do not alter access, blocking or case prioritization. Comparison with independently completed analyst decisions can then estimate actual prevalence, false-alert workload, phishing miss rate, feature availability and explanation usefulness across brands, languages, domain types and collection channels. Only a model that meets predefined performance, security and governance criteria should progress beyond decision support.

Before activation, a validation register should document the data period, inclusion criteria, labelling and duplicate-handling rules, split logic, extractor and model versions, hyperparameters, error slices and unsupported conditions. Thresholds should be selected from the pilot distribution and tied to differentiated actions rather than treated as universal constants. The explanation interface should be evaluated separately from classifier accuracy by testing whether analysts detect inconsistent recommendations, follow verification prompts and provide defensible rationales. Input validation, secure acquisition, authorized feedback and change control are mandatory whenever the model, extractor, threshold, collection method or source distribution changes.

Scientific reporting should distinguish development, internal testing, grouped validation, temporal validation and external validation. It should state the positive class, confusion-matrix orientation, class counts, duplicate handling, unavailable-at-inference variables and threshold-selection procedure. XAI reporting should



additionally identify the explained output, feature encoding, reference distribution and aggregation rule, while keeping global and local claims separate. Near-perfect benchmark results require stronger leakage, source and temporal diagnostics rather than stronger deployment claims.

DISCUSSION

The UCI results confirm that ensemble tree methods provide an effective balance of discrimination and compatibility with established attribution methods. Extra Trees led by phishing F1, while random forest led by ROC-AUC. This pattern illustrates why no single metric should determine deployment. If ranking is central and the threshold is adjustable, the random forest may be attractive; if the selected benchmark threshold and class balance resemble the operational setting, Extra Trees offers a marginally better F1 balance.

The PhiUSIIL experiment is arguably more instructive than a simple confirmation of high accuracy. Perfect or near-perfect results across diverse algorithms indicate that the supplied representation strongly separates the classes. This may reflect genuinely powerful signals, but it also lowers the evidential value of marginal algorithmic improvements. In such circumstances, the research priority should shift from another decimal place of accuracy to source analysis, leakage diagnosis, temporal testing and evaluation on features available in real time.

The lexical-only ablation demonstrates that strong separation persists even after excluding content and similarity variables. This supports the practicality of rapid pre-navigation screening because lexical features can be computed without rendering potentially malicious pages. However, it does not establish robustness to attacker adaptation. Lexical patterns are inexpensive to modify, and legitimate services often use long parameterized URLs. A production architecture should therefore combine early lexical scoring with later infrastructure and content verification.

The study's explanation card extends existing phishing XAI research by making verification explicit. Calzarossa et al. [4] and Uddin et al. [7] show how feature attribution can identify influential variables; the present framework asks what an analyst should do with that information and what must be logged. This procedural contribution is important because an explanation plot alone does not create accountability, reproduce evidence or define permissible actions.

The framework also responds to Rudin's warning about post-hoc explanations [10]. Where a transparent model achieves adequate performance, it may be preferable for high-impact decisions. Where an ensemble is used, the explanation should be treated as a bounded representation of model behaviour and should be cross-checked against permutation importance, local alternative explanations and external evidence. The system never presents the explanation as the true reasoning of an autonomous agent or as causal proof.

From an institutional perspective, implementation should begin with decision-support use in a monitored pilot. The model can rank incoming URLs and generate cards while analysts continue their existing verification process. Disagreement and missed cases should be reviewed weekly. Only after stable temporal performance, validated thresholds, security testing and documented governance should the system inform stronger containment actions. Autonomous legal characterization remains outside the model's role.

LIMITATIONS AND DIRECTIONS FOR FURTHER RESEARCH

The first limitation is the reliance on curated public datasets. The UCI corpus is historically important but does not reflect all contemporary infrastructure. PhiUSIIL is larger and newer, yet its strong feature-label associations and duplicate records make random-split results unusually optimistic. Neither dataset reproduces the full sequence by which an operational analyst encounters, acquires and verifies a suspicious resource.

The second limitation is the absence of harmonized raw-feature extraction across datasets. Because the two corpora define different variables, the experiment cannot train on one dataset and test on the other. Future work should implement a single versioned extractor for lexical, DNS, certificate and safely acquired webpage features, then apply it to multiple feeds collected in separated time windows.

The third limitation concerns explanation validation. The study evaluates TreeSHAP rankings, resampling stability and mode-replacement fidelity, but the mode-replacement intervention uses training-set modes and does not establish causal validity or robustness under every plausible perturbation. A formal user study with cybersecurity and law-enforcement analysts was also outside the present design. Future work should compare explanation-card variants, measure completion time and verification accuracy, and test whether explanations increase appropriate reliance without increasing acceptance of incorrect model outputs.

The fourth limitation is adversarial robustness. The current ablation estimates dependence on feature domains but does not simulate an adaptive attacker. Future experiments should evaluate feasible transformations such as token changes, subdomain variation, redirect modification, benign certificate acquisition and page-content manipulation. Robustness evaluation must preserve semantic attack functionality; arbitrary perturbations may create unrealistic samples.



The fifth limitation is temporal calibration and cost-sensitive thresholding. The reported reliability curve, Brier score and expected calibration error characterize the historical UCI hold-out set only. Future work should repeat calibration assessment prospectively, compare recalibration methods under temporal validation and optimize thresholds against documented false-positive and false-negative costs. Thresholds must be reviewed when case volume, prevalence or threat conditions change.

Finally, the study does not infer offender identity, intent or legal qualification. These conclusions require evidence beyond webpage features and are governed by applicable legal and procedural standards. The proposed system is a triage and analytical-support mechanism. Its value depends on integration with lawful data acquisition, competent human assessment and accountable institutional controls.

DATA AND CODE AVAILABILITY

The UCI Phishing Websites and PhiUSIIL datasets are publicly available from the UCI Machine Learning Repository [15,16]. The executable analysis script, fixed model specifications, random seeds, metric calculations and source data for the figures are retained by the corresponding author and are available for editorial or scholarly verification upon reasonable request. The materials contain no personal communications, credentials or operational case data.

CONCLUSIONS

The study developed a human-centred explainable AI framework for phishing website detection and evaluated five machine-learning models under a pre-specified, internally validated protocol. With phishing treated as the positive class, Extra Trees achieved accuracy of 0.9774, precision of 0.9784, recall of 0.9704, F1 of 0.9744 and ROC-AUC of 0.9948 on the UCI hold-out set. Random forest achieved the highest ROC-AUC of 0.9967. The confusion matrix contained 951 detected phishing cases, 29 missed phishing cases, 1,210 correctly recognized legitimate cases and 21 false alerts.

The complementary PhiUSIIL experiment produced near-perfect conventional results. Rather than treating them as evidence of deployment readiness, the study examined domain-grouped folds, feature-domain ablation, strong feature-label associations and duplicate records. The lexical-only grouped model retained F1 of 0.9970, demonstrating substantial within-collection signal while reinforcing the need to distinguish robustness diagnostics from temporal or external validation.

TreeSHAP analysis confirmed that certificate state and anchor behaviour dominated the UCI model's global explanations. The mean pairwise overlap of the ten leading features across five training folds was 0.891, showing substantial but non-absolute stability. The proposed Phishing Threat Explanation Card transforms these bounded model explanations into an auditable analytical record. It separates score from threshold, lists risk-increasing and protective indicators, connects them to independent verification tasks, records uncertainty and preserves the analyst's disposition.

Predictive uncertainty and probability reliability were assessed rather than inferred from point estimates alone. The Extra Trees 95% bootstrap interval was 0.9670-0.9812 for F1 and 0.9915-0.9976 for ROC-AUC; the Brier score was 0.0187. Replacing the three leading SHAP features with their training-set modes reduced F1 to 0.3543, compared with mean F1 of 0.9478 after random three-feature replacement. This supports the faithfulness of the global ranking to the fitted model while remaining a non-causal, configuration-specific diagnostic.

The principal practical conclusion is that explainability must be designed as a controlled socio-technical process. A strong classifier is only one component. Reliable use additionally requires provenance, temporal validation, validated thresholds, abstention, monitoring, secure feedback and trained human oversight. Under these conditions, explainable machine learning can improve phishing triage while preserving accountability and evidentiary caution.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE AND AI-ASSISTED TECHNOLOGIES

During the preparation of this manuscript, ChatGPT (OpenAI) was used for language editing, improving the clarity and consistency of the text, and refining figure labels. The author reviewed and approved all changes and takes full responsibility for the final content of the manuscript.

REFERENCES

1. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357. <https://doi.org/10.1016/j.eswa.2018.09.029>



2. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10, 93104–93139. <https://doi.org/10.1109/ACCESS.2022.3204051>
3. Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University - Computer and Information Sciences*, 35(2), 590–611. <https://doi.org/10.1016/j.jksuci.2023.01.004>
4. Calzarossa, M. C., Giudici, P., & Zieni, R. (2024). Explainable machine learning for phishing feature detection. *Quality and Reliability Engineering International*, 40(1), 362–373. <https://doi.org/10.1002/qre.3411>
5. Zieni, R., Massari, L., & Calzarossa, M. C. (2023). Phishing or not phishing? A survey on the detection of phishing websites. *IEEE Access*, 11, 18499–18519. <https://doi.org/10.1109/ACCESS.2023.3247135>
6. Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, 10, 93575–93600. <https://doi.org/10.1109/ACCESS.2022.3204171>
7. Uddin, K. M. M., Biswas, N., Rikta, S. T., Nur-A-Alam, M., & Mostafiz, R. (2025). Explainable machine learning for phishing site detection: A high-efficiency approach using boosting models and SHAP. *The Journal of Engineering*, 2025(1), Article e70110. <https://doi.org/10.1049/tje2.70110>
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
10. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
11. Kehkashan, T., Abdelhaq, M., Al-Shamayleh, A. S., Huda, N., Yaseen, I. A., Ahmed, A. I. A., & Akhunzada, A. (2025). Explainable phishing website detection for secure and sustainable cyber infrastructure. *Scientific Reports*, 15, Article 41751. <https://doi.org/10.1038/s41598-025-27984-w>
12. Prasad, A., & Chandra, S. (2024). PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. *Computers & Security*, 136, Article 103545. <https://doi.org/10.1016/j.cose.2023.103545>
13. Lim, B., Huerta, R., Sotelo, A., Quintela, A., & Kumar, P. (2025). EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2503.20796>
14. Mohammad, R., & McCluskey, L. (2012). *Phishing Websites* [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C51W2X>
15. Prasad, A., & Chandra, S. (2024). *PhiUSIIL Phishing URL (Website)* [Dataset]. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/967/phiusiil%2Bphishing%2Burl%2Bdataset>
16. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
17. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
18. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
19. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
20. Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>

**Габорець Ольга Андріївна**

доктор філософії, доцент,

доцент кафедри оперативно-розшукової діяльності та інформаційної безпеки

Донецький державний університет внутрішніх справ, м. Кропивницький, Україна.

ORCID: 0000-0001-7791-6795

olga-gaborets@ukr.net

**ПОЯСНЮВАНЕ МАШИННЕ НАВЧАННЯ ДЛЯ НАДІЙНОГО ВИЯВЛЕННЯ ФІШИНГОВИХ
ВЕБСАЙТІВ: ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ НА ДВОХ НАБОРАХ ДАНИХ**

Анотація. Виявлення фішингових вебсайтів часто розглядають як завдання класифікації, у якому високі прогностичні показники вважають достатнім підтвердженням практичної придатності моделі. Однак у правоохоронній діяльності та сфері кібербезпеки такий підхід є неповним, оскільки результат моделювання має бути зрозумілим, придатним до незалежної перевірки та відокремленим від висновків щодо авторства, умислу або юридичної кваліфікації. У дослідженні розроблено й експериментально оцінено орієнтовану на людину систему виявлення фішингових вебсайтів на основі пояснюваного штучного інтелекту, яка поєднує керовану класифікацію зі структурованими поясненнями для аналітика. П'ять моделей машинного навчання оцінено на наборі UCI Phishing Websites, а додатковий експеримент проведено на більшому наборі PhiUSIIL. В обох експериментах фішинг послідовно визначено позитивним класом. Модель Extra Trees забезпечила accuracy 0,9774, precision 0,9784, recall 0,9704, F1 0,9744 і ROC-AUC 0,9948. За результатами стратифікованого бутстреп-оцінювання 95% довірчий інтервал становив 0,9670–0,9812 для F1 та 0,9915–0,9976 для ROC-AUC. Показник Брієра дорівнював 0,0187, а очікувана похибка калібрування, розрахована за десятьма інтервалами, — 0,0089. TreeSHAP-аналіз 400 спостережень відкладеної тестової вибірки визначив стан сертифіката та поведінку посилань-якорів основними глобальними чинниками впливу. У п'яти фолдах середній попарний індекс Жаккара для множин десяти найважливіших ознак становив 0,891 за стандартного відхилення 0,094. У тесті відповідності пояснення моделі заміна трьох провідних SHAP-ознак їхніми модальними значеннями, визначеними на навчальній вибірці, знизила F1 з 0,9744 до 0,3543, тоді як випадкова заміна трьох ознак дала середній F1 0,9478. Майже ідеальні результати на наборі PhiUSIIL не розглянуто як свідчення готовності моделі до практичного впровадження: групована крос-валідація, абляція до чітко визначених 18 лексичних ознак URL та аналіз дублікатів виявили залежність роздільності класів від особливостей формування набору даних. Водночас модель Extra Trees, навчена лише на лексичних ознаках у межах групованої валідації, зберегла середній F1 0,9970. Запропоновано Картку пояснення фішингової загрози, яка перетворює прогноз на придатний до аудиту запис, що містить оцінку моделі, поріг ухвалення рішення, провідні індикатори, дії з незалежної перевірки, відомості про невизначеність і рішення аналітика. Науковий внесок дослідження має емпіричний і процедурний характер: здійснено документоване та придатне до аудиту порівняння моделей, кількісно оцінено прогностичну невизначеність, калібрування та відповідність пояснень моделі, виявлено чинники потенційного завищення результатів на бенчмаркових даних і визначено запобіжники для забезпечення належного людського контролю.

Ключові слова: фішинг; пояснюваний штучний інтелект; машинне навчання; виявлення фішингових вебсайтів; SHAP; підтримка рішень; правоохоронна діяльність; аналітика кіберзагроз.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357. <https://doi.org/10.1016/j.eswa.2018.09.029>
2. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10, 93104–93139. <https://doi.org/10.1109/ACCESS.2022.3204051>



3. Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University - Computer and Information Sciences*, 35(2), 590–611. <https://doi.org/10.1016/j.jksuci.2023.01.004>
4. Calzarossa, M. C., Giudici, P., & Zieni, R. (2024). Explainable machine learning for phishing feature detection. *Quality and Reliability Engineering International*, 40(1), 362–373. <https://doi.org/10.1002/qre.3411>
5. Zieni, R., Massari, L., & Calzarossa, M. C. (2023). Phishing or not phishing? A survey on the detection of phishing websites. *IEEE Access*, 11, 18499–18519. <https://doi.org/10.1109/ACCESS.2023.3247135>
6. Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, 10, 93575–93600. <https://doi.org/10.1109/ACCESS.2022.3204171>
7. Uddin, K. M. M., Biswas, N., Rikta, S. T., Nur-A-Alam, M., & Mostafiz, R. (2025). Explainable machine learning for phishing site detection: A high-efficiency approach using boosting models and SHAP. *The Journal of Engineering*, 2025(1), Article e70110. <https://doi.org/10.1049/tje2.70110>
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
10. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
11. Kehkashan, T., Abdelhaq, M., Al-Shamayleh, A. S., Huda, N., Yaseen, I. A., Ahmed, A. I. A., & Akhunzada, A. (2025). Explainable phishing website detection for secure and sustainable cyber infrastructure. *Scientific Reports*, 15, Article 41751. <https://doi.org/10.1038/s41598-025-27984-w>
12. Prasad, A., & Chandra, S. (2024). PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. *Computers & Security*, 136, Article 103545. <https://doi.org/10.1016/j.cose.2023.103545>
13. Lim, B., Huerta, R., Sotelo, A., Quintela, A., & Kumar, P. (2025). EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2503.20796>
14. Mohammad, R., & McCluskey, L. (2012). *Phishing Websites* [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C51W2X>
15. Prasad, A., & Chandra, S. (2024). *PhiUSIIL Phishing URL (Website)* [Dataset]. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/967/phiusiil%2Bphishing%2Burl%2Bdataset>
16. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
17. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
18. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
19. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
20. Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>

Отримано редакцією журналу / Received: 02.06.26

Прорецензовано / Revised: 16.06.26

Схвалено до друку / Accepted: 26.10.26



This work is licensed under Creative Commons Attribution-noncommercial-sharealike 4.0 International License.