



Beyond Random Splits: Class-Conditioned Site Exposure and Attribution Change in Phishing URL Evaluation

Sara Arif¹, Muhammad Ali Arif², Waheed Ur Rehman³ , Zanib Nawab⁴, Rimsha Minhaas⁵, Asma Ejaz⁶

^{1,4,5} Department of Computer Science, University of Chakwal, Chakwal 48800, Pakistan,

²Military College of Signals (MCS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan,

³Department of Mechatronics Engineering, University of Chakwal, Chakwal 48800, Pakistan,

⁶Department of Information Technology, University of Bahrain, Sakhir, Kingdom of Bahrain

✉ Corresponding Author: waheed.urrehman@uoc.edu.pk

Received: September 5,2026 **Accepted:** September 15,2026 **Published:** September 25,2026

ABSTRACT

Random URL-level evaluation can place test URLs from registrable sites already represented in training. Its generalization condition therefore differs from evaluation on previously unseen sites. Prior phishing research has examined site-disjoint, temporal, cross-dataset, robustness, and explanation settings. However, the controlled relationship among class-conditioned site exposure (the percentage of Test URLs in each class whose frozen private-PSL registrable-site key also occurs in Train), performance retention, and attribution change across random and site-disjoint evaluation has received limited direct study. We evaluated a provenance-controlled URL-only corpus of 124,154 URLs (52,589 benign and 71,565 phishing) under random URL-level Regime A and zero-overlap registrable-site-disjoint Regime C. The evaluation used three compact representation/model families, five frozen seeds per primary regime, site-cluster percentile bootstrap intervals, and within-model attribution comparisons. In A42, 99.7% of benign and 64.6% of phishing test URLs belonged to registrable sites represented in training; Regime C imposed 0.0% registrable-site overlap by construction. The A42–C42 MCC gaps were 0.095 for M1 (95% interval [-0.039, 0.235]), 0.200 for M2 [0.102, 0.315], and 0.151 for M3 [0.082, 0.247]; the M2 and M3 intervals excluded zero, whereas M1's included zero. M3's character n-gram TF-IDF logistic-regression representation retained the strongest absolute unseen-site MCC, although M2 and M3 had larger A-to-C decreases than M1. Attribution rankings remained broadly consistent within models while selected feature/token magnitudes and ranks changed across regimes; these summaries are descriptive and non-causal. The findings support reporting class-conditioned site exposure and using registrable-site-disjoint

evaluation alongside conventional random splits when the intended target is performance on previously unseen sites.

Keywords: Phishing URL detection; registrable-site-disjoint evaluation; unseen-site generalization; Public Suffix List; feature attribution; machine learning evaluation.

1. INTRODUCTION

A phishing attempt often presents a URL before any page content is rendered or inspected. The URL can be screened without executing page code, retrieving visual content, or waiting for reputation evidence, making it useful as an early signal. It also remains available when a page has disappeared or cannot safely be fetched, which permits offline analysis. URL-only classifiers cannot replace browser, content, or network controls, but they can provide a lightweight early layer for triage. Prior work has obtained useful discrimination from lexical and structural properties of the URL string and from representations learned directly from its characters and tokens [1]–[3].

The modelling literature includes engineered lexical features with linear or tree-based classifiers, sparse character patterns, convolutional URL encoders, and Transformer representations. These families differ in transparency, capacity, and computation [1], [2], [4]. A high-capacity sequence model may capture subtle character context. A linear sparse model can instead expose which substrings carry weight. Neither property guarantees transfer beyond the sites represented in the training sample. Headline benchmark scores are also difficult to compare when studies change the source data, preprocessing, and partitioning protocol together. PhishBench made this

comparability problem explicit [3]. Strong performance on an internal random partition does not by itself establish how the detector will behave when the test URLs come from sites absent during training.

URLs from the same registrable site are not necessarily independent observations. A registrable site may contribute several hostnames, paths, or query variants, so random URL-level partitioning can place URLs associated with the same site in both training and test data even after exact duplicates are removed. The model may then encounter related hostname tokens, path templates, or other site-specific structure at test time. Hostname separation alone does not necessarily remove this relation: two subdomains can belong to the same registrable site, while private-suffix rules may identify hosted tenants that an ICANN-only rule would aggregate. Such exposure could make the task easier than deployment on a previously unseen site, although its magnitude and effect are empirical questions. The two settings estimate performance on different targets: a new URL from infrastructure already represented in training, or a URL from a site excluded from training altogether. General guidance for structured data recommends aligning the grouping unit with the intended prediction task, and recent phishing work confirms that registrable-site-aware evaluation is feasible [5], [6]. We define a Test URL as site-exposed when its frozen private-PSL registrable-site key is also

present among Train URLs. Class-conditioned site exposure is the percentage of Test URLs satisfying this condition within each class.

Site-aware testing has clear precedents. Mamatov compares random, chronological, Public-Suffix-List-derived registrable-site-disjoint, and cross-dataset regimes, with duplicate and conflict controls [6]. Fed-urlBERT assigns eTLD+1 groups to a single partition and treats multi-tenant services separately, providing direct precedent for tenant-aware isolation [7]. Other studies address source transfer rather than DNS-site novelty. Rashid et al. formulate cross-source generalization as unsupervised domain adaptation, while Mia et al. and Yi et al. combine cross-dataset evaluation with feature attribution analyses [8]–[10]. Robustness and explanation stability are also established topics. Ahamed et al. evaluate structured URL perturbations, held-out attack families, external transfer, and top-K SHAP stability in one protocol [11]. These studies already cover domain-disjoint testing, cross-dataset analysis, robustness testing, and post-hoc explanation. We therefore do not treat those elements alone as contributions of this study.

The remaining question is more specific. The verified studies most directly related to this work do not report random-split exposure to registrable sites separately for phishing and benign examples. This class conditioning matters because the two labels may have different site multiplicity and source structure, which an aggregate exposure rate can conceal. Collection pipelines may also make label-associated properties such as root-path prevalence, scheme, hostname form, or URL depth easier to learn. These properties must be measured before a regime difference can be

interpreted as evidence about site novelty rather than source composition. Existing comparisons do not directly connect measured exposure and acquisition-linked URL structure to the performance change observed when the same corpus moves from random URL-level partitioning to a zero-overlap site partition. Attribution presents a related gap. Mia et al. observe qualitative changes in SHAP patterns across datasets, and Ahamed et al. quantify stability under input perturbation, but neither comparison measures attribution change between random and registrable-site-disjoint evaluation while holding the corpus and model family fixed [9], [11]. The verified literature did not provide a high-confidence direct treatment of that controlled regime comparison. This statement defines an evidence-qualified boundary, not a categorical priority claim.

This study compares class-conditioned registrable-site exposure, performance retention, and feature-attribution behaviour under two evaluation regimes applied to the same provenance-controlled corpus: conventional random URL-level partitioning and group-atomic partitioning with zero overlap under a frozen ICANN-plus-private Public Suffix List definition. The same protocols are used for compact URL-only representations, avoiding a simultaneous change in corpus or model family. The study focuses on how the evaluation target changes across regimes. It does not seek a new architecture benchmark. Source-associated structural properties are audited before modelling, and an ICANN-only grouping is retained as a sensitivity analysis. Attribution changes are interpreted as associations consistent with altered cue reliance, not as causal proof of memorization or shortcut learning.

The study is guided by the following research questions:

RQ1: How do phishing-detection performance and class-conditioned train–test site exposure differ between conventional random URL-level partitioning and zero-overlap, frozen ICANN+private-PSL registrable-site partitioning?

RQ2: Which compact URL-only representations retain performance most consistently when evaluation moves from random URL-level partitions to unseen registrable sites?

RQ3: How do feature attributions change between random URL-level and unseen registrable-site evaluation for the same model and feature families, and are any changes consistent with reduced reliance on repeated-site or acquisition-source-associated cues?

The contributions are:

1. We quantify train–test registrable-site exposure separately for phishing and benign classes under random URL-level partitioning and compare performance with a zero-overlap registrable-site evaluation.
2. We compare compact URL-only representation families under identical random and unseen-site protocols, focusing on performance retention rather than model novelty.
3. We quantify attribution change between those regimes and examine whether any changes are consistent with altered reliance on repeated-site or acquisition-source-associated cues.

4. We audit acquisition-linked URL structure and evaluate ICANN-plus-private versus ICANN-only PSL grouping as methodological sensitivities.

The analysis uses a fixed URL corpus to quantify class-conditioned Train–Test registrable-site exposure and compare three URL-only representation families under random URL-level and zero-overlap registrable-site-disjoint evaluation. It relates the evaluation change to performance retention, site-cluster uncertainty, and descriptive within-model attribution changes. Multiplicity controls and feature ablations provide supporting diagnostics

2. RELATED WORK

2.1 URL-BASED PHISHING DETECTION AND REPRESENTATION FAMILIES

URL-only phishing detection has followed two broad approaches. The first uses explicit lexical and structural variables, including URL length, delimiter counts, subdomain depth, path and query properties, suspicious tokens, and related host features. These variables support linear models and tree ensembles, are inexpensive to compute, and expose a feature space that can be inspected directly. Sahingoz et al. reported strong results from lexical and natural-language features with conventional classifiers, while Mamatov later compared Logistic Regression, Random Forest, XGBoost, a character CNN, and a stacking ensemble under several evaluation regimes [1], [6]. The broad PhishBench study had already shown why headline scores alone are difficult to compare: papers often change the data, feature set, and evaluation protocol at the same time [3].

A second approach learns directly from the URL string. URLNet combined character- and word-level convolutions so that rare or previously unseen tokens could still be represented [2]. Character CNNs removed much of the need for manual feature design, and Texception extended this line with a character/word architecture designed for low false-positive operation at production scale [12], [13]. URLTran brought Transformer representations into phishing URL detection [4]. More recently, Fed-urlBERT coupled a custom URL tokenizer and BERT with federated and split-learning settings [7]. These models differ in capacity and tokenization. Those choices alone do not define the generalization problem, because an unseen character sequence can still come from a registrable site observed during training.

Results outside conventional internal splits are mixed. Ghalechyan et al. evaluated neural URL models on later feeds and production data and found that useful performance could persist, although changes in the phishing feeds and the Alexa-derived benign source complicate a purely temporal interpretation [14]. Mamatov found that the relative value of feature groups changed under temporal, registrable-site-disjoint, and cross-dataset testing, and Fed-urlBERT reported performance close to centralized training under heterogeneous clients [6], [7]. The comparison is not simply classical versus neural. A compact set spanning transparent engineered features, sparse character patterns, and nonlinear trees is sufficient to test whether representation behavior changes with the evaluation unit.

2.2 DATASET CONSTRUCTION, PROVENANCE, AND SOURCE ARTEFACTS

Phishing URL corpora are commonly assembled rather than sampled from one population. PhishTank and OpenPhish supply phishing reports, but their reporting and verification processes are not the same. URLHaus is also used in malicious-URL research, although its labels concern malware distribution rather than phishing specifically. Benign examples have often come from popularity lists such as Alexa, Majestic, or Tranco [3], [14], [6], [7]. Tranco was designed as a persistent and manipulation-resistant top-sites ranking, so it is well suited to defining a reproducible site frame; it is not a collection of benign page URLs [15]. A ranked domain and an archived page from that domain are different observational units.

Public releases differ in what they preserve. PhiUSIIL provides a large engineered security profile and has become a frequent benchmark. Tamal and LegitPhish were released as dedicated phishing datasets with different mixtures of raw URLs and derived variables [16]–[18]. Kaggle-derived feature tables are also common in cross-dataset work [9], [10]. A file may be usable for its original classification task but unsuitable for a site-overlap audit when raw URLs, collection times, or source lineage are missing. Exact duplicates, normalized duplicates, and conflicting labels pose a separate problem. Some recent work reports their treatment carefully, whereas many benchmark papers state only that duplicate rows were removed.

Collection choices can also leave visible signatures in the URL string. A benign set drawn from popular homepages may contain more root paths, HTTPS URLs, or www hosts than a phishing feed rich in deep paths, query strings, IP literals, and disposable subdomains. Mia et al. found that feature ranks and directions

changed sharply between two derived datasets, while Yi et al. reported much stronger transfer on a different collection of benchmark sets [9], [10]. PhishVN is unusually explicit on this point: its documentation flags HTTPS as a collection artefact rather than a trustworthy signal [19]. Archived web indexes such as Common Crawl can supply page-level URLs, but crawl inclusion is not a benign label and must be tied to an independently defined site frame [20]. These observations motivate source-structure auditing; they do not establish that any new corpus is free of source bias.

2.3 RANDOM SPLITTING, LEAKAGE, AND UNSEEN-SITE EVALUATION

Random URL-level splitting treats rows as exchangeable, an assumption that is questionable when one registrable site contributes several URLs. Related hostnames, path templates, or tokens can then appear on both sides of the split. Exact duplicate URLs are only one part of the issue. A hostname-disjoint split can separate login.example.com from secure.example.com while leaving both under the same registrable site. General work on structured cross-validation states the underlying principle directly: blocking should follow the intended prediction task when observations share spatial, temporal, or hierarchical structure [5]. For phishing URL detection, the appropriate block depends on whether the target is a new URL from a known site or a URL from a site not represented during training.

Mamatov evaluates the stricter case. The study normalizes URLs, removes exact normalized duplicates and label conflicts, checks near-duplicates within registrable sites, and compares random, chronological, PSL-derived registrable-

site-disjoint, and cross-dataset regimes [6]. Its stacking model falls from an F1 of 0.973 under random evaluation to 0.918 in the registrable-site-disjoint setting and 0.884 across datasets. The paper does not report the PSL version or private-section semantics, and the bootstrap unit is unclear. Its central result is nevertheless unambiguous: domain-aware evaluation is established prior work. The study leaves unmeasured the amount of random-split site exposure and whether that exposure is distributed differently between phishing and benign examples.

Fed-urlBERT provides another important precedent. It assigns organization-owned eTLD+1 groups to one client and one train/test partition, while treating subdomains of multi-tenant services such as github.io as separate tenants [7]. The authors also report an automated leakage audit, though not numerical exposure counts. PhishVN uses registrable-domain groups in a temporal split and releases manifests and deterministic grouping artefacts, but its exact PSL and private-suffix implementation was not verified from the dataset documentation [19]. These studies establish precedents for eTLD+1 isolation and tenant-aware handling. They do not quantify random-split registrable-site exposure separately by class, and neither tests how an ICANN-only grouping differs from an explicitly frozen ICANN-plus-private definition.

2.4 TEMPORAL AND CROSS-DATASET GENERALIZATION

Temporal holdout, cross-dataset testing, and unseen-site testing answer different questions. A chronological split asks whether a model transfers to later observations; a cross-dataset test changes the source pipeline as well; a site-

disjoint split removes the specified site group regardless of time. Mamatov includes all three regimes and reports a different loss under each [6]. Ghalechyan et al. use later feeds and production traffic, while continual-learning work studies adaptation as phishing changes over time [14], [21].

PhishVN orders dated registrable-domain groups chronologically, providing useful recent dataset precedent, but its timestamp and grouping design should not be treated as interchangeable with a source-independent temporal benchmark [19].

Rashid et al. frame generalization as unsupervised domain adaptation between source datasets. In that study, a domain is a data distribution, not a DNS hostname or registrable site [8]. The reported results indicate that adaptation can recover some of the performance lost when training and testing sources differ. The methods available for verification did not permit checks of exact URL, hostname, or registrable-site overlap across the source sets. The comparison remains relevant, but it estimates a broader mixture of source, time, labeling, and representation changes than a zero-overlap site split.

The results of Mia et al. and Yi et al. show why cross-dataset performance cannot be summarized by one expected degradation. Using twenty harmonized features, Mia et al. obtained only 51% and 59% accuracy in the two transfer directions and observed large changes in SHAP rank and sign [9]. Yi et al. reported average accuracy above 94% across a primary and three secondary benchmark datasets [10]. The latter study describes manual duplicate review but does not report cross-set URL or site overlap; it also uses a different feature space and source

lineage. The disagreement may reflect real model behavior, overlap, class composition, preprocessing, benchmark difficulty, or a combination of these factors. Cross-dataset testing is established practice, but it is not a substitute for an explicit unseen-registrable-site estimand.

2.5 ROBUSTNESS OF PHISHING URL DETECTORS

Robustness to deliberate manipulation has an established literature. SpacePhish separates attacks in problem, preprocessing, and model spaces and evaluates twelve attacks against URL-, representation-, and combined-feature detectors. Several realistic attacks produce statistically significant degradation, although not every perturbation is effective [22]. Multi-Space Phish extends the same research line to combinations of attack spaces and shows that joint perturbations can be much more damaging than isolated changes [23]. The journal paper extends SpacePhish rather than providing an independent replication. Together, the two papers show that structured evasion analysis is already well developed.

Ahamed et al. combine clean evaluation, structured URL perturbations, held-out attack families, external PhishTank transfer, and explanation stability in one protocol [11]. Their accepted record reports sharp failures under subdomain manipulation for several model families and SHAP top-K Jaccard values around 0.53, indicating that the selected attribution set can change with the input. The final formatted methods were not available during verification, so details of preprocessing, label preservation, and statistical testing remain qualified. Even with this limitation, perturbation robustness and

quantitative explanation stability cannot be claimed as new here.

In this study, URL-string transformations have a narrower role. Once the unseen-site estimand and representation set are fixed, controlled changes to subdomain, path, query, encoding, or related lexical structure can test whether the observed representation ordering is brittle. Such an analysis is a sensitivity check. It should not be presented as a new adversarial framework or as evidence that a detector is operationally robust.

2.6 EXPLAINABILITY, STABILITY, AND ATTRIBUTION CHANGE

Post-hoc explanation is now common in phishing studies. Boosting and stacking papers use SHAP to rank engineered features, while other work combines global SHAP summaries with local explanations or LIME [24]–[27]. These analyses can show which variables a fitted model associates with its predictions. They do not automatically establish faithfulness, stability, or transfer of the explanation. Mayo et al., for example, study changing evidence across simulated email-reading windows; their temporal unit is a message sequence, not a later URL capture or an unseen registrable site. Cross-dataset work has begun to examine whether attributions travel with the model. Mia et al. compare TreeSHAP summaries across within-dataset, transfer, and merged-data settings and find that feature rank and effect direction can change with the dataset [9].

Their comparison is qualitative, with no reported scalar drift or faithfulness measure. Yi et al. provide global beeswarm and local waterfall explanations for tree models after cross-dataset performance tests, but they do not compare the

attribution distributions across datasets [10]. These papers establish cross-dataset XAI as prior art, while measurement of attribution change remains only partly developed. Ahamed et al. go further by quantifying top-K SHAP stability under perturbation, yet that is still a different comparison from changing the split regime while holding the corpus and model family fixed [11]. Mamatov supplies the relevant random and registrable-site-disjoint regimes but does not analyze explanations [6].

The specific open question is whether attribution rankings, distributions, or feature-group mass change between those two evaluation settings beyond normal seed and model variability. A shift toward or away from site- or source-associated feature groups would be consistent with changed cue reliance. It would not prove that the model learned a causal shortcut.

2.7 CLOSEST WORK AND REMAINING RESEARCH SPACE

Mamatov is the closest split study [6]. It combines normalized duplicate control, within-site near-duplicate checks, label-conflict removal, random evaluation, temporal evaluation, a PSL-derived registrable-site-disjoint split, and external testing. The split and the resulting degradation are therefore not new. The narrower focus here is the exposure mechanism left implicit in that comparison: how often random-split test URLs encounter a site represented during training, whether that rate differs by class, and how it relates to the measured regime difference. Fed-urlBERT is the closest grouping precedent because it enforces eTLD+1 isolation and makes an explicit exception for multi-tenant services [7].

This precedent also rules out a novelty claim for eTLD+1 or tenant-aware grouping. Related explanation and shift studies further narrow the novelty claim. Ahamed et al. already measure explanation stability under structured perturbations and test unseen attack families and external transfer. Mia et al. show that attribution rank and direction can change across datasets, Yi et al. combine external benchmark testing with ordinary SHAP, and Rashid et al. address cross-source transfer through domain adaptation [11], [9], [10], [8].

PhishVN adds non-peer-reviewed but relevant precedent for registrable-domain temporal grouping, manifests, and explicit treatment of an HTTPS collection artefact [19]. These elements do not support a general novelty claim for this study. Prior studies cover registrable-site and tenant-aware splitting, temporal and cross-dataset evaluation, structured robustness testing, and several forms of post-hoc explanation. The remaining issue is narrower.

The verified literature does not report random-split registrable-site exposure separately for phishing and benign classes, and source-structure diagnostics are rarely integrated with

that exposure. It also does not directly quantify how attributions change when the same model families move from random URL-level evaluation to a zero-overlap registrable-site-disjoint regime. Examining these relationships does not propose another site-disjoint split. It measures the exposure permitted by the conventional split, how that exposure is distributed, and whether performance and model evidence change under the corresponding unseen-site condition.

3. METHODOLOGY

3.1 CORPUS CONSTRUCTION AND PROVENANCE

The primary corpus combined a frozen PhishTank online-valid phishing snapshot dated 2026-08-17 with benign URLs acquired from the CC-MAIN-2026-30 index for a frozen 2,500-site Tranco frame. Both streams were processed under the frozen normalization, deduplication, and label-conflict controls. Registrable sites with mixed labels were quarantined as whole sites; the quarantine contains 175 rows from five registrable sites and was not relabeled or merged into the modeling corpus.

Table 1 Final post-quarantine primary corpus.

Quantity	Value
Total URLs	124,154
Benign URLs	52,589 (42.4%)
Phishing URLs	71,565 (57.6%)
Private-PSL registrable sites	29,091
Quarantined URLs	175
Quarantined mixed-label private-PSL sites	5

Table 2 Pre-quarantine combined-manifest site accounting.

Pre-quarantine grouping	Benign sites	Phishing sites	Union sites	Mixed-label intersection sites
private-PSL registrable-site accounting	1,428	27,673	29,096	5
ICANN-only registrable-site accounting	1,408	12,605	14,007	6

The final post-quarantine primary DNS-host corpus contains 124,154 URLs: 52,589 benign (42.4%) and 71,565 phishing (57.6%), across 29,091 registrable sites. This is a frozen acquisition corpus, not a primary temporal evaluation. Table 1 reports final post-quarantine corpus quantities. The pre-quarantine private-PSL and ICANN-only site accounting that documents the five quarantined mixed-label site keys is reported in Table 2.

3.2 URL PARSING AND REGISTRABLE-SITE DEFINITION

URLs were parsed offline with the frozen official Public Suffix List snapshot dated 2026-08-17 and a conservative normalization policy: scheme and host lowercasing plus IDNA UTS #46 handling with `std3_rules=True`. In this study, a frozen private-PSL registrable site is the PSL-derived top domain under the public suffix computed from the frozen PSL snapshot with both ICANN and PRIVATE rules; this operational unit defines primary Regime-C grouping and retains distinct tenant sites under private suffixes. The ICANN-only site key is a separate, broader grouping used only for a diagnostic regime. IP literals were identified before IDNA processing and retained as a

separate supplementary phishing stratum; the primary DNS-host corpus consequently has `is_ip=0`, although that feature remains in the frozen engineered-feature contract. Parsing and feature extraction made no DNS, webpage, or URL-shortener-resolution requests.

3.3 CLASS-CONDITIONED SITE EXPOSURE

The primary quantity of interest is class-conditioned Train/Test registrable-site exposure. Let T_c denote the set of Test URLs belonging to class $c \in \{\text{benign}, \text{phishing}\}$, let $g(u)$ denote the frozen registrable-site key of URL u , and let G_{train} denote the set of registrable-site keys represented in Train. The class-conditioned exposure for class c is:

$$E_c = \frac{|\sum_{u \in T_c} 1\{g(u) \in G_{\text{train}}\}|}{|T_c|} \times 100$$

where $1[\cdot]$ is the indicator function. This quantity measures the percentage of Test URLs in class c that share a registrable-site key with at least one Train URL. Under the primary uncapped A42 split, $E_{\text{benign}} = 99.7\%$ and $E_{\text{phishing}} = 64.6\%$. Under Regime C, $E_{\text{benign}} = 0.0\%$ and

$E_{\text{phishing}} = 0.0\%$ by construction, because each registrable-site-key group is assigned atomically to a single partition.

3.4 EVALUATION REGIMES AND SPLIT CONSTRUCTION

All assignments used frozen Train/Validation/Test targets of 70/15/15 and were verified by manifest hashes. Regime A is stratified random URL-level partitioning over seeds 13, 42, 73, 101, and 2026. Regime B is a hostname-disjoint diagnostic at seed 42; it enforces hostname separation but permits, and measures, residual registrable-site overlap. Regime C is the primary registrable-site-disjoint evaluation over seeds 42, 123, 456, 789, and 1234. Each registrable-site-key group is assigned

atomically to one partition, yielding zero private-PSL registrable-site, hostname, and normalized-URL overlap across Train, Validation, and Test. Regime D is an ICANN-only site-disjoint grouping diagnostic at seed 42. Seed 42 was used for the direct A42–C42 comparison because it had been frozen as the A-side matched-analysis seed and as the primary Regime-C split before main-model Test execution, providing a common pre-specified comparison context. All five A seeds remained formally equal for aggregate reporting; five-seed summaries are descriptive, and seeds are not inferential units.

Figure 1 depicts the frozen framework and split protocol; Table 3 summarizes the primary evaluation regimes.

Table 3 Frozen split and evaluation definitions.

Regime	Definition	Seed	Train/Test site condition
A	Random URL-level split	42 (matched-analysis); five frozen A seeds	Site overlap permitted by design
C	Registrable-site-disjoint split	42 (primary); five frozen C seeds	Zero registrable-site overlap

3.5 PRIMARY SPLIT COMPOSITION

Table 4 summarizes the Train, Validation, and Test composition of the direct Seed-42 random (A42) and registrable-site-disjoint (C42) evaluations. Under Regime A42, URLs are partitioned at the individual URL level across Train (86,907 URLs), Validation (18,623 URLs), and Test (18,624 URLs), permitting

registrable sites to appear across partitions. Under Regime C42, registrable-site keys are partitioned atomically, yielding zero site overlap across Train (86,931 URLs; 20,125 sites), Validation (18,665 URLs; 2,170 sites), and Test (18,558 URLs; 6,796 sites). The 29,091 total registrable sites in the corpus are thus completely partitioned across the three C42 splits.

Beyond Random Splits: Class-Conditioned Site Exposure and Attribution Change in Phishing URL Evaluation

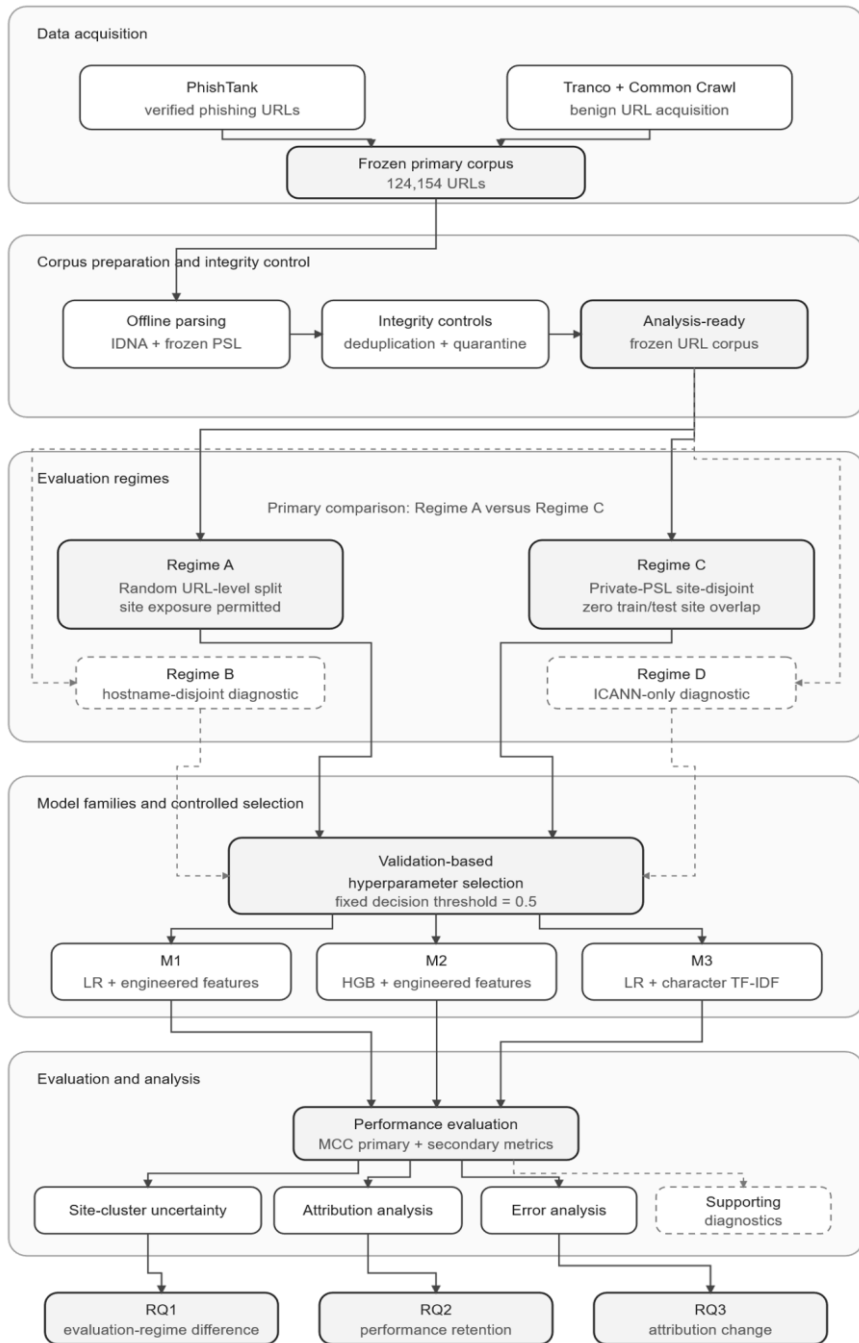


Figure. 1. Framework and frozen split/evaluation protocol.

Beyond Random Splits: Class-Conditioned Site Exposure and Attribution Change in Phishing URL Evaluation

Table 4 Primary split composition for A42 and C42 evaluations.

Regime	Split	Total URLs	Benign URLs	Phishing URLs	Total Sites	Site condition
A42	Train	86,907	36,812	50,095	21,869	Site overlap permitted
A42	Validation	18,623	7,888	10,735	6,245	Site overlap permitted
A42	Test	18,624	7,889	10,735	6,341	Site overlap permitted
C42	Train	86,931	36,834	50,097	20,125	Zero registrable-site overlap
C42	Validation	18,665	7,930	10,735	2,170	Zero registrable-site overlap
C42	Test	18,558	7,825	10,733	6,796	Zero registrable-site overlap

Overall corpus class prevalence: 42.4% benign (52,589/124,154), 57.6% phishing (71,565/124,154). A42 Train: 42.4% benign (36,812/86,907), 57.6% phishing (50,095/86,907). A42 Validation: 42.4% benign (7,888/18,623), 57.6% phishing (10,735/18,623). A42 Test: 42.4% benign (7,889/18,624), 57.6% phishing (10,735/18,624). C42 Train: 42.2% benign (36,834/86,931), 57.8% phishing (50,097/86,931). C42 Validation: 42.2% benign (7,930/18,665), 57.8% phishing (10,735/18,665). C42 Test: 42.2% benign (7,825/18,558), 57.8% phishing (10,733/18,558). Class prevalence affects interpretation of precision, PR-AUC, and recall at the fixed 0.5 threshold.

3.6 URL REPRESENTATIONS AND MODEL FAMILIES

The evaluated model families are M1–M3 (Table V). The model set spans three representation strategies so that evaluation behavior can be compared across representation families under the same protocol, with corpus and partitioning held fixed. The study makes no architecture-novelty claim.

M1 is logistic regression over 22 engineered URL-only features after Train-fitted standardization. It provides a transparent, engineered linear representation: coefficients are directly interpretable as signed linear

Beyond Random Splits: Class-Conditioned Site Exposure and Attribution Change in Phishing URL Evaluation

associations with the decision boundary after standardization.

M2 is `sklearn.ensemble.HistGradientBoostingClassifier` over the same 22 engineered features, with `early_stopping=False`; it is not an XGBoost model. M2 provides a nonlinear, tree-based engineered-feature representation capable of capturing feature interactions that the linear M1 cannot express.

M3 is logistic regression over Train-fitted raw-URL character TF-IDF n-grams, with `analyzer='char'`, an n-gram range of 2–6, `sublinear_tf=True`, `min_df=5`, and L2 normalization. M3 provides a sparse character-level representation that learns directly from the URL string without manual feature engineering.

M1–M3 cover three representation settings: (i) a transparent engineered linear representation; (ii) a nonlinear engineered-feature tree model; and (iii) a sparse character-level representation. This comparison tests whether evaluation behavior changes across representation families while the experimental protocol remains fixed.

No source, acquisition, partition, hostname, registrable-site, or row identifier was supplied as a predictive feature.

M1 and M2 use the same 22-feature numeric representation; M1 applies Train-fitted `StandardScaler`, whereas M2 consumes the numeric features directly. For presentation, the features are grouped as length/depth variables (`url_length`, `hostname_length`, `path_length`, `query_length`, `path_depth`, `subdomain_depth`); binary structural/URL flags (`has_query`, `has_port`, `is_root_path`, `is_ip`, `is_private_suffix`, `has_www`, `is_https`, `is_missing_scheme`, `has_punycode`); and character/count/ratio/entropy features (`digit_count`, `digit_ratio`, `hyphen_count`, `dot_count`, `special_char_count`, `percent_encoded_count`, `char_entropy`). These are descriptive groups within one representation, not separate model pipelines. These features are deterministic functions of the parsed URL fields. Metadata columns, including the frozen record identifier, label, the registrable-site key, the ICANN-only site key, and hostname, remained outside the predictive matrix.

Table 5 Frozen model and feature representations.

Model	Representation	Estimator	Attribution
M1	22 engineered URL features	Standardized logistic regression	Standardized coefficients
M2	22 engineered URL features	HistGradientBoosting	TreeSHAP + permutation importance
M3	Character TF-IDF n-grams	Logistic regression	Top n-gram coefficients

3.7 HYPERPARAMETER SELECTION AND PREDICTION PROTOCOL

Preprocessing and model selection were split- and seed-specific. For every executed identity, the scaler for M1, vectorizer for M3, and model parameters were fit only on that identity's Train partition; validation and Test data did not participate in fitting. M2 used the engineered numeric features without a scaler. Candidate selection used Validation MCC only: M1 searched C in $\{0.01, 0.1, 1, 10\}$; M2 searched the full product of max_iter in $\{100, 200\}$, learning_rate in $\{0.05, 0.1\}$, and max_leaf_nodes in $\{31, 63\}$; and M3 searched C in $\{0.1, 1, 10\}$ and max_features in $\{25,000, 50,000\}$. The selected configuration was frozen before Test access. Binary predictions used a fixed probability threshold of 0.5 throughout the study. Threshold optimization was intentionally excluded so that the regime comparison used a common decision rule; hyperparameters were selected using Validation MCC while the threshold remained fixed. Test outcomes were not used to choose a threshold.

For exact validation-MCC ties, the recorded lower-capacity ordering selected lower C for logistic regression; for M2 it selected lower max_iter , then lower max_leaf_nodes , then lower learning_rate ; and for M3 it selected lower C , then lower max_features . One supporting-analysis exception is documented explicitly: in the AB5b M2 feature-ablation setting, this M2 ordering was documented after all candidates tied on validation but before Test access. The resulting selection freeze and Test gate were preserved; the event is a nonblocking provenance limitation, not a Test-driven choice.

3.8 EVALUATION METRICS AND PRIMARY CONTRAST

The primary performance metric was Matthews correlation coefficient (MCC). For a binary classifier with true positives TP, true negatives TN, false positives FP, and false negatives FN (phishing as the positive class), MCC is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC ranges from -1 to $+1$ and accounts for class imbalance by incorporating all four entries of the confusion matrix, making it appropriate as a primary metric when class prevalence is unequal.

Secondary metrics were PR-AUC, macro F1, phishing recall, precision, false-positive rate (FPR), specificity, balanced accuracy, and ROC-AUC, with phishing as the positive class. ROC-AUC was not the primary metric, and per-site MCC was not used.

The primary regime contrast is defined as:

$$\Delta MCC_{A-C} = MCC_A - MCC_C$$

where MCC_A is the Test MCC under Regime A and MCC_C is the Test MCC under Regime C for the same model family and seed. The direct A42/C42 comparison at Seed 42 is the pre-specified inferential context; the five-seed summaries are descriptive only. The frozen primary A42–C42 ΔMCC values are: M1: 0.095, M2: 0.200, M3: 0.151.

3.9 SENSITIVITY AND DIAGNOSTIC ANALYSES

Supporting analyses were predeclared and retain their restricted roles. The cap sensitivity analysis evaluated Cap-10 and Cap-5 multiplicity variants for A42/C42 and M1–M3 by intersecting frozen cap membership with frozen uncapped assignments; it did not repartition data or transfer fitted state. Additional diagnostic evaluations considered the B42 hostname-disjoint and D42 ICANN-only regimes as diagnostics only. The feature-ablation analysis evaluated M1 and M2 on C42 under six feature configurations: removal of the site-associated group (AB1), `is_private_suffix` (AB2), scheme features (AB3), length features (AB4), and character-only (AB5a) or structure-only (AB5b) subsets. The visible-shortener sensitivity analysis excluded the frozen visible-shortener membership from A42 and C42 without replaying or repartitioning the splits, then performed fresh selection within the filtered partitions. The optional lexical-stress analysis was not executed and is not represented as a Methods experiment.

3.10 STATISTICAL UNCERTAINTY ANALYSIS

The statistical uncertainty analysis operated read-only on the canonical uncapped A/C Test predictions for M1–M3. It used a non-parametric, class-stratified private-PSL registrable-site cluster percentile bootstrap with $B=2,000$ replicates, master seed 99, and two-sided 95% limits. Each replicate sampled benign and phishing registrable-site clusters independently with replacement and included every URL in a selected cluster, preserving repeated cluster multiplicity; URLs were not treated as independent resampling units. The

reported uncertainty quantity is a 95% site-cluster percentile bootstrap interval. We resampled whole registrable-site clusters rather than individual URLs because multiple Test URLs can belong to the same site. Whole-site resampling preserves this within-site grouping and aligns uncertainty analysis with the study's registrable-site generalization unit; treating URLs as independent bootstrap units would ignore that clustering.

The principal contrast independently bootstrapped the native A42 and C42 Test sets and computed $MCC_A - MCC_C$ at each replicate index. It is an unpaired clustered comparison, not a row- or site-paired analysis. No p-values, null-hypothesis tests, standardized effect measures, or seed-level inference were used. Five-seed regime summaries report descriptive mean and sample standard deviation only; seeds were not bootstrap units.

Uncertainty estimation followed the pre-specified metric hierarchy. Site-cluster 95% percentile intervals were generated for MCC and the eight secondary performance metrics for all 30 canonical regime/seed/model identities. MCC was the primary metric and the only metric for which we formed the direct native A42–C42 difference interval. Secondary-metric intervals were retained in the frozen statistical outputs as supporting descriptive uncertainty, whereas the main-text interval display focuses on MCC and five-seed summaries are descriptive rather than seed-level inferential samples.

3.11 ATTRIBUTION ANALYSIS

For attribution comparison, we used the exact intersection of the authoritative A42 and C42 Test memberships, aligned by the frozen record

identifier. This produced 2,801 unique URLs (1,182 benign and 1,619 phishing), without additional balancing, random subsampling, correctness-based selection, or model-specific eligibility filtering. Restricting attribution analysis to URLs held out under both regimes allowed the A42 and C42 fitted pipelines to be compared on identical input instances; this restriction applies only to attribution and does not make the native-Test performance comparison paired. M1 used signed and absolute coefficients from Train-standardized logistic regression, including descriptive rank comparison across the 22 features. M2 used TreeSHAP with `feature_perturbation='tree_path_dependent'`, raw model output, and no background dataset; MCC permutation importance used 30 repetitions and seed 42 as corroborating evidence. M3 was limited to regime-specific top character-n-gram coefficient lists, top-k overlap, and exact shared-vocabulary comparisons; full-vocabulary magnitude comparisons, SHAP, and permutation importance were not used for M3 because its vectorizers have different vocabularies and IDF values. Attribution magnitudes were not compared across model families, and this analysis did not use bootstrap intervals or p-values.

3.12 ERROR ANALYSIS

The error analysis was a read-only analysis of M1–M3 false positives and false negatives from the native A42 and C42 Test predictions. Errors were assigned the prospectively operationalized, frozen multi-label taxonomy of co-occurring URL properties, with unmatched errors marked UNCATEGORIZED. The native A/C comparison is descriptive and unpaired; it uses no inferential statistics, confidence intervals,

source or template analysis, or attribution linkage. Ranked contributor-site aggregation was performed for error counts, but manuscript outputs use anonymized site identifiers and do not publish raw URLs or registrable-site names.

3.13 REPRODUCIBILITY AND EXECUTION CONTROLS

The corpus, quarantine, PSL, parsing contract, split assignments, seed registry, feature contract, and output artifacts were frozen and hash-verified. The execution process recorded split-specific selection freezes and enforced read-only Test governance: Test data could be transformed and evaluated only after selection was durable, and it was never used for preprocessing, hyperparameter selection, feature selection, or threshold tuning. The core execution environment was Python 3.10.11 with the version-locked scientific stack recorded in the reproducibility manifest; the attribution analysis additionally used its locked SHAP environment. Exhaustive file hashes, package versions, and artifact provenance are retained in the frozen reproducibility records rather than repeated in the main text. Manifests, frozen split assignments, SHA-256 hashes, version information, and execution records were retained throughout the study.

4. RESULTS

4.1 RANDOM VERSUS UNSEEN REGISTRABLE-SITE EVALUATION

Under the primary uncapped random URL-level split (A42), 99.7% of benign Test URLs and 64.6% of phishing Test URLs belonged to registrable sites represented in Train. Regime C imposed 0.0% Train/Test registrable-site overlap for both classes by construction (Fig. 2; Table 6).

Beyond Random Splits: Class-Conditioned Site Exposure and Attribution Change in Phishing URL Evaluation

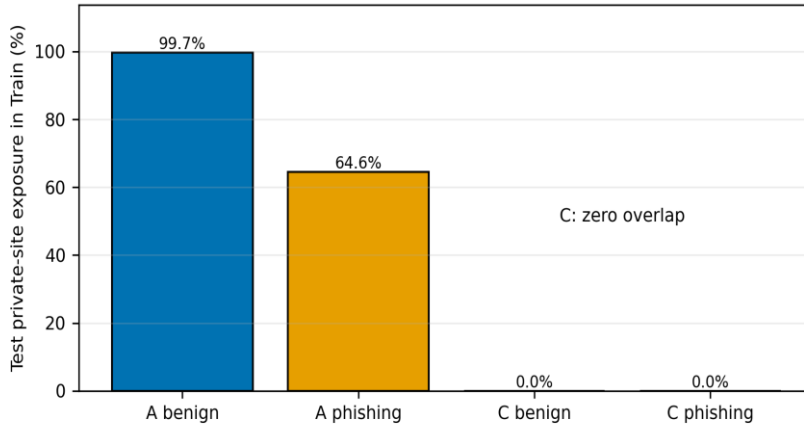


Figure 2. Class-conditioned registrable-site exposure under the frozen evaluation regimes.

Table 6 Frozen exposure and evaluation context.

Variant	A benign exposure (%)	C benign exposure (%)	A phishing exposure (%)	C phishing exposure (%)
uncapped	99.7	0.0	64.6	0.0
cap10	98.9	0.0	29.2	0.0
cap5	97.6	0.0	28.2	0.0

These values define the class-conditioned exposure contrast between the two evaluation regimes. The Seed-42 MCC comparison differed by model (Fig. 3; Table VII). The A42–C42 MCC gap was 0.095 for M1, with a 95% site-cluster percentile bootstrap interval of $[-0.039, 0.235]$, which included zero. The corresponding gaps were 0.200 for M2 $[0.102, 0.315]$ and 0.151 for M3 $[0.082, 0.247]$; both intervals excluded zero (Fig. 4). The five-seed results in Table 7

provide descriptive context for these Seed-42 native-Test comparisons.

4.2 REPRESENTATION-LEVEL PERFORMANCE RETENTION

Across the five frozen seeds, mean MCC under A versus C was 0.613 ± 0.011 versus 0.585 ± 0.056 for M1, 0.905 ± 0.003 versus 0.770 ± 0.066 for M2, and 0.974 ± 0.002 versus 0.863 ± 0.031 for M3 (Fig. 3; Tables 5 and 7). M3 had

Beyond Random Splits: Class-Conditioned Site Exposure and Attribution Change in Phishing URL Evaluation

the highest MCC in both regimes. M2 and M3 showed larger absolute A-to-C mean differences than M1, whereas M1 began from a substantially lower A-regime performance level. We therefore

report both the performance level and retention across the frozen evaluation regimes, rather than interpreting the absolute gap alone.

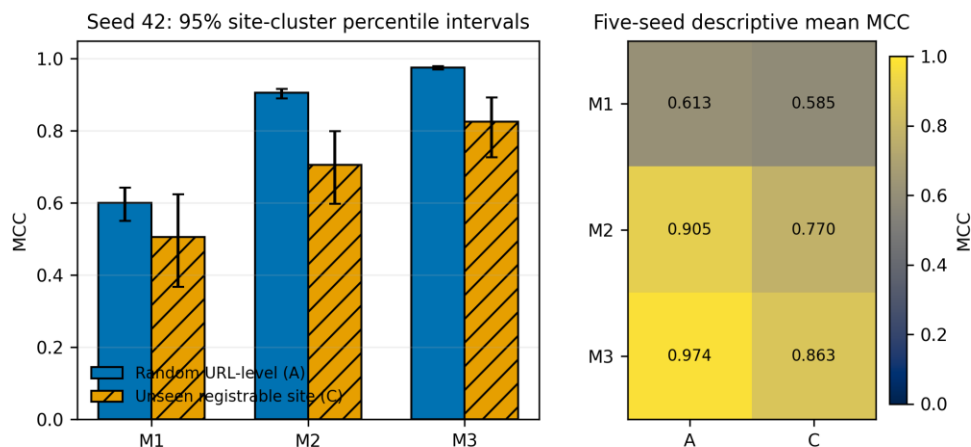


Figure 3. Performance under random URL-level and unseen registrable-site evaluation.

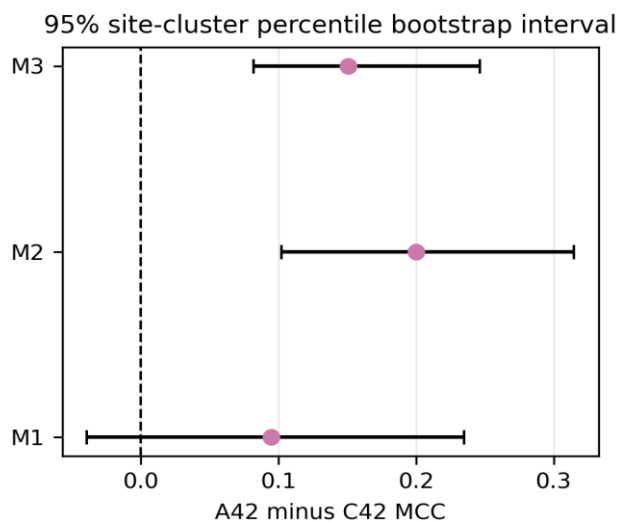


Figure 4. A-C MCC contrast with 95% site-cluster percentile bootstrap interval estimates.

4.3 SECONDARY METRICS

Table 8 reports the complete set of executed secondary metrics for the primary A42 and C42 evaluations at Seed 42. All values were computed on the corresponding frozen Test partitions with phishing as the positive class;

threshold-dependent metrics used the fixed 0.5 operating point, whereas PR-AUC and ROC-AUC used probability scores. Supporting 95% site-cluster percentile intervals for these metrics were generated and retained in the frozen statistical outputs; Table 7 displays the primary MCC uncertainty.

Table 7 Frozen performance summary across the primary evaluation regimes.

Model	A five-seed MCC (mean $\pm$ SD)	C five-seed MCC (mean $\pm$ SD)	A42–C42 MCC	95% site-cluster percentile bootstrap interval
M1	0.613 $\pm$ 0.011	0.585 $\pm$ 0.056	0.095	[−0.039, 0.235]
M2	0.905 $\pm$ 0.003	0.770 $\pm$ 0.066	0.200	[0.102, 0.315]
M3	0.974 $\pm$ 0.002	0.863 $\pm$ 0.031	0.151	[0.082, 0.247]

Table 8 Executed secondary metrics for primary A42 and C42 evaluations (Seed 42, uncapped, fixed 0.5 threshold, phishing positive).

Metric	M1 A42	M1 C42	M2 A42	M2 C42	M3 A42	M3 C42
MCC	0.600	0.506	0.906	0.706	0.976	0.825
PR-AUC	0.919	0.902	0.992	0.939	0.999	0.984
ROC-AUC	0.893	0.841	0.989	0.902	0.999	0.977
Macro F1	0.792	0.753	0.953	0.848	0.988	0.912
Phishing recall	0.912	0.810	0.954	0.803	0.988	0.906
Precision	0.785	0.783	0.965	0.925	0.992	0.942
Specificity	0.660	0.692	0.953	0.911	0.989	0.924
FPR	0.340	0.308	0.047	0.089	0.011	0.076
Balanced accuracy	0.786	0.751	0.954	0.857	0.988	0.915

4.4 SENSITIVITY AND ABLATION ANALYSES

The supporting cap analyses retained an A-versus-C MCC difference for all three models under both Cap-10 and Cap-5 corpus variants (Fig. 6(a)). These capped analyses are sensitivity

results and do not replace the uncapped primary comparison.

The predeclared C42 ablations showed model-specific sensitivity to the engineered-feature subsets (Table 9). Removing the site-associated engineered features (AB1) changed C42 MCC from 0.506 to 0.468 for M1 and from 0.706 to

0.681 for M2. Removing the length features (AB4) yielded 0.409 for M1 and 0.544 for M2. The tested structure-only subset (AB5b) yielded MCC of 0.000 for both M1 and M2 and produced all-phishing/majority-positive predictions in that tested setting. This result is confined to that specified subset and setting.

Table 9 Effect of engineered-feature ablations under the primary unseen-site evaluation (C42).

Ablation	Configuration	M1 MCC	Δ M1	M2 MCC	Δ M2
Full	All engineered features	0.506	—	0.706	—
AB1	Remove site-associated group	0.468	−0.038	0.681	−0.024
AB2	Remove is_private_suffix	0.492	−0.013	0.690	−0.015
AB3	Remove scheme features	0.507	+0.002	0.700	−0.005
AB4	Remove length features	0.409	−0.097	0.544	−0.161
AB5a	Character-oriented subset	0.382	−0.124	0.420	−0.285
AB5b	Structure-only subset	0.000	−0.506	0.000	−0.706

After exclusion of the frozen visible-shortener subset, the A/C difference remained evident for M2 (0.906 versus 0.702) and M3 (0.977 versus 0.824), while it narrowed for M1 (0.615 versus 0.542) (Fig. 6(b)). This is a sensitivity result, not a lexical-stress or robustness evaluation.

4.5 ATTRIBUTION CHANGES ACROSS EVALUATION REGIMES

All attribution comparisons are interpreted within model family on the frozen common cohort (Fig. 5). M1 coefficients are parameters

on Train-standardized engineered features, whereas M2 TreeSHAP values are contributions in raw model-output units; their magnitudes represent different quantities and are not directly comparable across M1 and M2. M3 coefficient summaries are likewise interpreted only within its own character representation. For M1, the Spearman correlation of absolute standardized-coefficient ranks was 0.920. The largest selected coefficient-magnitude changes included an increase of 1.125 for url_length, an increase of 0.603 for digit_count, and a decrease of 0.557 for digit_ratio.

Beyond Random Splits: Class-Conditioned Site Exposure and Attribution Change in Phishing URL Evaluation

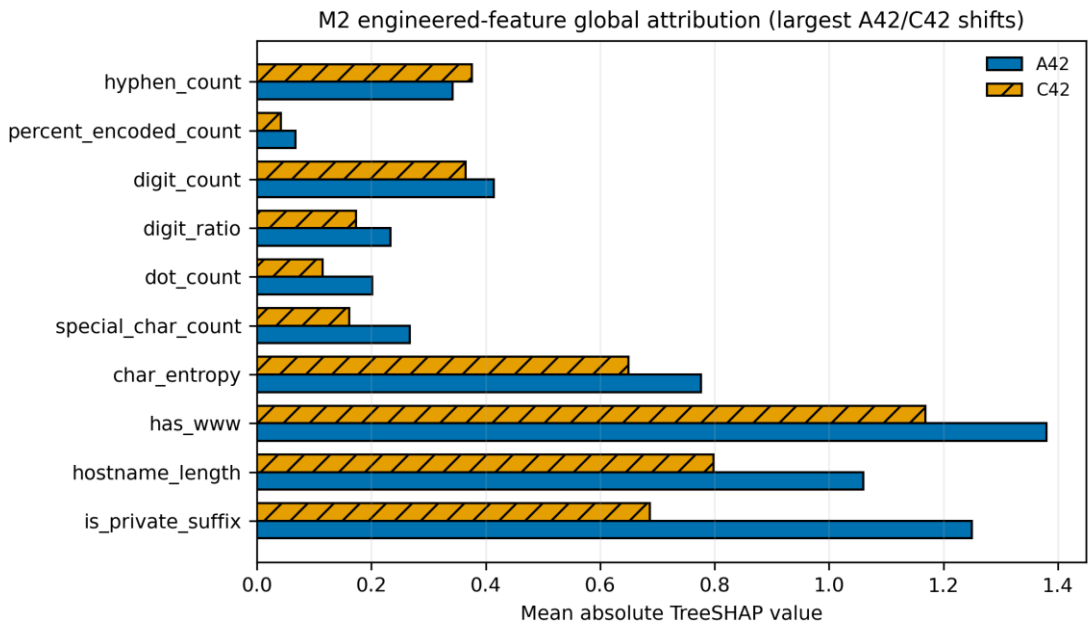


Figure 5. Within-model global attribution comparisons across evaluation regimes.

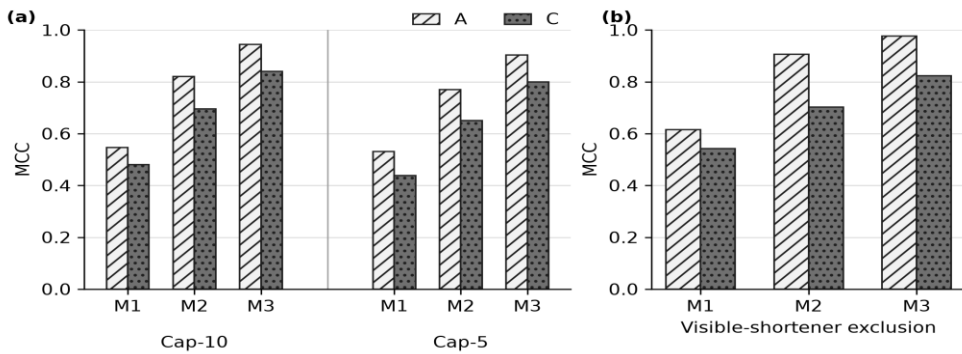


Figure 6. Sensitivity of the A-versus-C performance pattern to URL multiplicity caps and visible-shortener exclusion. (a) Cap-10 and Cap-5 multiplicity variants. (b) Visible-shortener exclusion under A42 and C42.

For M2, the global mean-absolute-SHAP rank correlation was 0.993. The largest pooled changes were decreases of 0.562 for `is_private_suffix`, 0.261 for `hostname_length`, and 0.212 for `has_www`; the largest selected

phishing-stratum change was a 0.752 decrease for `is_private_suffix`. The descriptive mean paired local change was 2.958 on the raw attribution scale and 0.392 after independent normalization. The frozen permutation-

importance analysis is retained as corroborating, model-specific supporting evidence in the supporting attribution records.

For M3, the separately fitted vocabularies had 24,992 exact shared tokens. Shared-token sign agreement was 87.6%, the absolute-coefficient rank correlation was 0.737, and the fixed top-50 absolute-coefficient sets had Jaccard overlap of 0.515. The within-model summaries show broad global rank consistency, with selected magnitude and rank changes across regimes.

4.6 ACQUISITION-LINKED STRUCTURAL DIFFERENCES

The acquisition-linked structural audit quantified URL-property differences between the benign and phishing streams in the pre-quarantine combined corpus. Table 10 reports the executed binary-property audit. The observed differences reflect acquisition-source composition and require bounded interpretation; they do not demonstrate that any model learned a causal shortcut.

Table 10 Acquisition-linked structural differences between phishing and benign URL streams in the pre-quarantine combined corpus (binary properties; phishing $n = 71,579$; benign $n = 52,750$ uncapped).

Property	Phishing (%)	Benign (%)	Absolute diff (pp)	Direction
is_root_path	38.2	7.0	31.2	Phishing higher
has_www	4.1	52.0	47.9	Benign higher
is_private_suffix	24.8	0.4	24.4	Phishing higher
is_https	91.9	94.6	2.7	Benign higher
has_query	15.2	18.6	3.4	Benign higher
has_percent_encoding	1.5	5.2	3.8	Benign higher

For continuous properties, phishing URLs had a lower median raw URL length (median 39 vs. 61 characters) and lower median path depth (median 1.0 vs. 2.0). The suffix-distribution Cramér's V association was 0.678, indicating strong label-associated suffix composition. These observations motivated source-structure controls in the model evaluation design and inform interpretation of attribution changes. They do not demonstrate that any model learned a causal shortcut.

4.7 ERROR CHARACTERISTICS UNDER UNSEEN-SITE EVALUATION

The read-only error analysis provides supporting descriptive evidence for the detailed taxonomy. Fig. 7 summarizes the false-positive and false-negative burden. Relative to A42, C42 false-negative counts and rates increased from 946 (8.8%) to 2,042 (19.0%) for M1, from 492 (4.6%) to 2,111 (19.7%) for M2, and from 134 (1.2%) to 1,004 (9.4%) for M3. The taxonomy

comparison used native A/C Test sets and was unpaired.

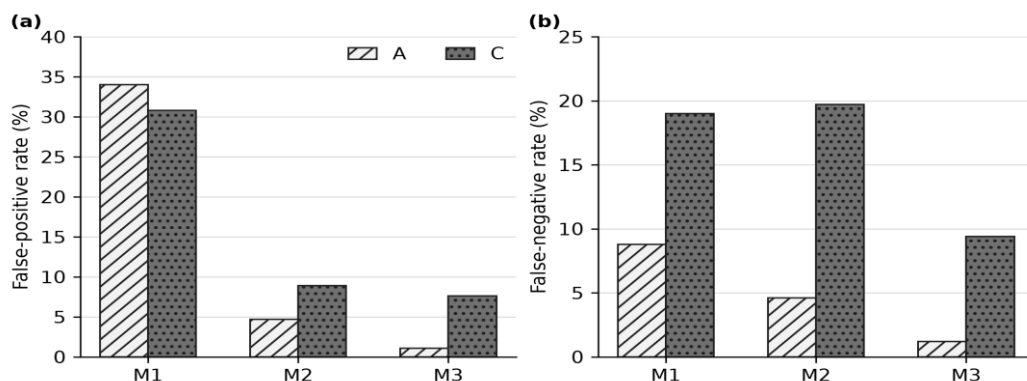


Figure 7. False-positive and false-negative burden under random URL-level (A42) and unseen registrable-site (C42) evaluation.

Categories are multi-label: a single error may satisfy more than one URL-property category. Consequently, category percentages are non-exclusive and are not expected to sum to 100%. Each category percentage uses all errors of the same type (false positive or false negative) within the corresponding regime and model as its denominator.

The largest category-prevalence changes were decreases in FN-D of 31.8 percentage points for M3, 26.4 percentage points for M1, and 20.9 percentage points for M2; M2 FN-E increased by 16.2 percentage points. For the C42 M1 false negatives, the leading anonymized contributor sites accounted for 626 (30.7%) and 345 (16.9%) errors, respectively. The taxonomy is multi-label, so category percentages need not sum to 100%; category membership is reported as a co-occurring URL-property description rather than an error mechanism.

5. DISCUSSION

5.1 EVALUATION-REGIME EXPOSURE AND PERFORMANCE CONTEXT

The primary comparison covers two generalization conditions: random URL-level and site-disjoint evaluation. Under random URL-level partitioning, Test URLs can belong to registrable sites represented in Train, and the observed exposure differed between benign and phishing URLs. Regime C removes registrable-site overlap by construction. Its Test URLs therefore measure performance when the grouped sites are unseen by both Train and validation. The lower MCC values under Regime C for the stronger M2 and M3 models, together with the Seed-42 site-cluster percentile intervals that exclude zero for their A42–C42 contrasts, indicate that measured performance depends on the partitioning regime for those models. The

corresponding M1 interval includes zero in this direct native-Test context. Random URL-level and site-disjoint evaluation measure different generalization conditions and should not be treated as interchangeable estimates of the same condition.

5.2 REPRESENTATION RETENTION UNDER UNSEEN REGISTRABLE SITES

The three model families represent engineered linear features (M1), engineered nonlinear tree-based features (M2), and character n-grams with a linear model (M3). M3 achieved the highest absolute MCC under both random URL-level and unseen-site evaluation. M2 and M3 showed larger absolute decreases from A to C than M1, which started from a much lower A-regime performance level. Retention should therefore be assessed using both absolute performance under Regime C and the change across regimes, not the decrease alone. By these criteria, the character n-gram representation retained the strongest absolute performance under the unseen-site regime within this frozen URL-only evaluation.

5.3 INTERPRETING ATTRIBUTION CHANGES

The attribution comparison is descriptive and restricted to within-model interpretation; its scales do not support magnitude comparisons across model families. M1 and M2 had high global ranking consistency between A42 and C42. M3 had lower, but still substantial, consistency among its shared tokens. Although the global rankings were broadly consistent, selected feature or token magnitudes and ranks changed across regimes. The same fitted model families therefore assigned different attribution to selected lexical and structural cues under the two evaluation regimes. This pattern is

consistent with different reliance on the available cues, but it does not identify a causal mechanism, prove shortcut learning, or establish site memorization or source dependence.

5.4 SUPPORTING DIAGNOSTICS AND ERROR CHARACTERISTICS

The cap, shortener, ablation, and error analyses (Fig. 6, Table 9, and Fig. 7) are supporting diagnostics and do not replace the uncapped A/C comparison. The primary A/C performance pattern persisted in the capped sensitivity variants. After the visible-shortener subset was removed, the M2 and M3 A/C differences remained, so visible shorteners alone do not account for the observed pattern. The predeclared ablations show that selected engineered feature groups materially affected C42 performance. The AB5b collapse is confined to the tested structure-only configuration. Error analysis found a substantially greater false-negative burden under unseen-site evaluation for M1 through M3, with category shifts and ranked contributor-site concentration providing descriptive context. The taxonomy was prospectively operationalized, is multi-label, and compares unpaired native Test sets. Its category patterns and site concentration therefore do not identify an error mechanism or demonstrate site memorization. Results from the Regime B and D diagnostic evaluations remain diagnostic-only context and do not change the primary A/C interpretation.

5.5 METHODOLOGICAL IMPLICATIONS

This study provides a controlled comparison of class-conditioned site exposure, performance, and attribution under random and frozen registrable-site-disjoint evaluation. Evaluation

reports can state the generalization condition more clearly by measuring class-conditioned Train/Test site exposure explicitly instead of assuming that a random URL-level split isolates registrable sites. Registrable-site-disjoint evaluation can complement conventional random evaluation when the question concerns sites unseen during development. The public-suffix semantics used to define registrable sites should be frozen and reported because the grouping definition sets the scope of the comparison. Repeated frozen seeds provide descriptive variation, while site-cluster intervals quantify uncertainty without turning the direct Seed-42 comparison into seed-level inference. Model comparisons should consider both absolute performance and retention under distributional restrictions. Attribution should also be compared across evaluation regimes rather than interpreted under one split alone. These implications apply to evaluation practice within the tested protocol and do not prescribe deployment policy.

6. THREATS TO VALIDITY

6.1 CONSTRUCT VALIDITY

The operational definition of registrable-site novelty rests on the frozen private-PSL grouping key. Under primary Regime C, an unseen Test site has a registrable-site key absent from both Train and Validation; the Train/Test exposure measure itself refers only to Train membership. This operationalization depends on the frozen PSL snapshot dated 2026-08-17 and the private-suffix semantics it encodes; a different PSL version or different private-suffix inclusion policy would produce different groupings and potentially different exposure estimates. The ICANN-only sensitivity analysis partially

bounds this dependence. The study is URL-only; no claim is made about page content, browser behavior, or non-URL features. The fixed 0.5 decision threshold is a modeling choice, not a deployment recommendation.

6.2 INTERNAL VALIDITY

The study audited acquisition-source structural differences between the phishing (PhishTank snapshot) and benign (Common Crawl / Tranco frame) streams. Differences in root-path prevalence, www-host prevalence, private-suffix prevalence, and path depth reflect acquisition-pipeline composition rather than the underlying phishing phenomenon. The audit motivated the chosen model features and the inclusion of source-associated diagnostics, but it cannot establish that any model learned a causal source-associated shortcut. Attribution changes between A42 and C42 are descriptive associations. They neither identify a causal mechanism nor explain why the performance difference occurred. These findings do not support a causal source effect, site-memorization mechanism, or shortcut-learning claim.

6.3 EXTERNAL VALIDITY

The phishing corpus is a frozen snapshot of online-valid PhishTank records dated 2026-08-17. Findings are specific to that snapshot and cannot be generalized to later phishing campaigns or different acquisition windows without primary temporal evaluation. No temporal holdout evaluation was conducted, so the study makes no time-robustness claim. No external deployment evaluation was performed, so zero-day detection capability and adversarial robustness cannot be claimed. The benign corpus is derived from a Tranco top-sites frame via

Common Crawl page-level URLs; it is not a probability sample of all benign web activity. The corpus is URL-only; findings do not cover content-based, network-based, or multimodal detectors.

6.4 STATISTICAL CONCLUSION VALIDITY

The direct inferential comparison is A42 versus C42 at Seed 42, using the frozen site-cluster percentile bootstrap intervals. This comparison is seed-42 specific; the confidence interpretation applies to that particular draw, not to a population of seeds. The bootstrap is unpaired: A42 and C42 native Test sets are bootstrapped independently and the Δ MCC distribution is computed at each replicate index. Five-seed summaries (mean $\pm$ SD) are descriptive and do not support seed-level inference. No p-values, null-hypothesis significance tests, or standardized effect measures were used or are reported. The supporting error taxonomy analysis is multi-label, descriptive, and unpaired, and cannot support exclusive-category or causal error inferences.

7. CONCLUSION

This study compared class-conditioned Train/Test registrable-site exposure, URL-only model performance, representation retention, and within-model attribution patterns under random URL-level and registrable-site-disjoint evaluation. Under random URL-level partitioning, Test URLs could belong to registrable sites represented in Train, and exposure differed by class. Regime C removed registrable-site overlap by construction. The stronger M2 and M3 pipelines had lower MCC under Regime C. Their direct Seed-42 site-

cluster percentile intervals for the A42–C42 contrasts excluded zero, while the corresponding M1 interval included zero. Within the frozen corpus and protocol, random URL-level and site-disjoint evaluation therefore measure different generalization conditions.

M3 retained the strongest absolute performance under unseen registrable-site evaluation among the tested URL-only representations. M2 and M3 had larger A-to-C decreases than M1, whose random-regime baseline was substantially lower. Retention should therefore be interpreted using both unseen-site performance and regime change. Within-model attribution rankings were broadly consistent across regimes, although selected feature and token magnitudes and ranks changed. This descriptive pattern does not establish a causal mechanism. Cap sensitivity, visible-shortener exclusion, predeclared ablations, and descriptive error analysis provided supporting context for the primary comparison.

The methodological contribution is a direct treatment of the generalization condition. Phishing-URL evaluations should report class-conditioned site exposure and can complement conventional random URL-level splits with registrable-site-disjoint evaluation under frozen PSL semantics. The conclusions are limited to the frozen acquisition snapshot, URL-only features, the operational private-PSL grouping definition, and analyses without primary temporal or external deployment evaluation. The attribution evidence is non-causal. Future work can include timestamped temporal evaluation, external and source-shift validation, additional collections, content or multimodal features, prospectively governed lexical perturbation studies, and broader deployment-oriented validation.

APPENDIX: SUPPORTING RESULTS

A. SIMPLE DIAGNOSTIC BASELINES

Two diagnostic baselines were evaluated before the main-model analyses. B1 used URL length alone in logistic regression, while B2 used a Train-fitted private-suffix one-hot representation. These baselines are supporting diagnostics and do not address RQ1–RQ3.

Across five Regime-A seeds, mean Test MCC was -0.039 ± 0.020 for B1 and 0.545 ± 0.004 for B2. Under Regime C, the corresponding mean Test MCC values were 0.191 ± 0.154 and 0.389 ± 0.103 , respectively. At Seed 42, B1 achieved MCC of 0.224 under C42 and B2 achieved 0.505. These results show that the private-suffix representation carried more predictive signal than URL length alone in the evaluated corpus, while its lower mean performance under site-disjoint evaluation is consistent with reduced reuse of label-associated suffix composition. The findings are descriptive and do not establish a causal mechanism.

B. DIAGNOSTIC REGIME RESULTS (B42 AND D42)

At Seed 42, the hostname-disjoint Regime B achieved Test MCC values of 0.536 (M1), 0.793 (M2), and 0.898 (M3). B42 enforced hostname separation while retaining measurable Train/Test registrable-site exposure of 21.9% for benign Test URLs and 27.7% for phishing Test URLs.

The ICANN-only Regime D achieved Test MCC values of 0.151 (M1), 0.824 (M2), and 0.899 (M3). Relative to the private-PSL grouping, the broader ICANN-only grouping collapses tenant sites: the 29,091 private-PSL registrable sites correspond to 14,002 ICANN-only groups, and the C42 and D42 assignments differ for 60,822 rows. Regimes B and D are supporting diagnostics and are not primary evaluation conditions.

DATA AVAILABILITY

The primary corpus was constructed from publicly accessible source streams: a frozen PhishTank online-valid phishing snapshot dated 17 August 2026 and page-level benign URLs obtained from Common Crawl CC-MAIN-2026-30 for a frozen 2,500-site Tranco frame. Registrable-site grouping used a frozen Public Suffix List snapshot dated 17 August 2026. The final post-quarantine corpus contains 124,154 URLs (52,589 benign and 71,565 phishing). Raw and derived URL data are not redistributed in the public repository because redistribution eligibility across the combined third-party sources has not been fully established. The public reproducibility repository at <https://github.com/SaraArif6198/phishing-unseen-site-evaluation> and its v1.1-revision-support release at <https://github.com/SaraArif6198/phishing-unseen-site-evaluation/releases/tag/v1.1-revision-support> provide acquisition and preprocessing code, provenance records, frozen split-regime definitions, SHA-256 verification records, canonical aggregate results, and statistical uncertainty outputs. Source datasets remain subject to their providers' access and redistribution conditions

CODE AVAILABILITY

Code and reproducibility artifacts supporting this study are publicly available at <https://github.com/SaraArif6198/phishing-unseen-site-evaluation> and in the stable v1.1-revision-support release at <https://github.com/SaraArif6198/phishing-unseen-site-evaluation/releases/tag/v1.1-revision-support>. The repository includes frozen execution scripts, configuration records, environment specifications, canonical aggregate outputs, figures, split documentation, SHA-256 verification artifacts, and statistical uncertainty outputs. The revision-support release preserves the submitted scientific results and adds documentation and previously retained frozen

reproducibility artifacts; no analyses were recomputed for the revision.

REFERENCES

- [1] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019, doi: 10.1016/j.eswa.2018.09.029.
- [2] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, "URLNet: Learning a URL representation with deep learning for malicious URL detection," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2018, doi: 10.1145/3219819.3219945.
- [3] A. El Aassal, S. Baki, A. Das, and R. M. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *IEEE Access*, vol. 8, pp. 22170–22192, 2020, doi: 10.1109/ACCESS.2020.2969780.
- [4] S. Maneriker *et al.*, "URLTran: Improving phishing URL detection using transformers," in *Proc. IEEE Military Communications Conf. (MILCOM)*, 2021, pp. 197–204, doi: 10.1109/MILCOM52596.2021.9653028.
- [5] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.
- [6] I. Mamatov, "GenAI-augmented ensemble learning framework for phishing URL detection using lexical host-based and sequential features," *International Journal of Theoretical and Applied Issues of Digital Technologies*, vol. 9, no. 3, pp. 45–52, 2026, doi: 10.62132/ijdt.v9i3.397.
- [7] Q. Zan, L. Song, H. He, Q. Wang, Y. Deng, Y. Wang, Y. Liu, and Y. Li, "Fed-urlBERT: Client-split federated BERT for malicious URL detection," *IEEE Access*, vol. 13, pp. 191164–191179, 2025, doi: 10.1109/ACCESS.2025.3629370.
- [8] F. Rashid, B. Doyle, S. C. Han, and S. Seneviratne, "Phishing URL detection generalisation using unsupervised domain adaptation," *Computer Networks*, vol. 245, Art. no. 110398, 2024, doi: 10.1016/j.comnet.2024.110398.
- [9] M. Mia, D. Derakhshan, and M. A. Pritom, "Can features for phishing URL detection be trusted across diverse datasets? A case study with explainable AI," in *Proc. ACM Int. Conf. on Network and System Security (NSysS)*, 2024, pp. 137–145, doi: 10.1145/3704522.3704532.
- [10] L. Yi, A. Omotosho, and H. Balogun, "Phishing URL detection and interpretability with machine learning: A cross-dataset approach," *Security and Privacy*, vol. 9, no. 1, Art. no. e70175, 2026, doi: 10.1002/spy2.70175.
- [11] T. Ahamed, S. C. Kakon, F. A. Farid, J. Uddin, and H. A. Karim, "An integrated evaluation protocol for adversarial robustness generalization and explanation stability in URL-based phishing detection," *Frontiers in Computer Science*, vol. 8, Art. no. 1834407, 2026, doi: 10.3389/fcomp.2026.1834407.
- [12] A. Aljofey *et al.*, "An effective phishing detection model based on character level convolutional neural network from URL," *Electronics*, vol. 9, no. 9, Art. no. 1514, 2020, doi: 10.3390/electronics9091514.
- [13] F. Tajaddodianfar, J. W. Stokes, and A. Gururajan, "Texception: A character/word-level deep learning model for phishing URL detection," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2857–2861, doi: 10.1109/ICASSP40776.2020.9053670.
- [14] H. Ghalechyan *et al.*, "Phishing URL detection with neural networks: An empirical

- study,” *Scientific Reports*, vol. 14, Art. no. 25134, 2024, doi: 10.1038/s41598-024-74725-6.
- [15] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, “Tranco: A research-oriented top sites ranking hardened against manipulation,” in *Proc. Network and Distributed System Security Symp. (NDSS)*, 2019, doi: 10.14722/ndss.2019.23386.
- [16] A. Prasad and S. Chandra, “PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning,” *Computers & Security*, vol. 136, Art. no. 103545, 2024, doi: 10.1016/j.cose.2023.103545.
- [17] C. M. Tamal *et al.*, “Dataset of suspicious phishing URL detection,” *Frontiers in Computer Science*, vol. 6, Art. no. 1308634, 2024, doi: 10.3389/fcomp.2024.1308634.
- [18] R. S. Potpelwar *et al.*, “LegitPhish: A large-scale annotated dataset for URL-based phishing detection,” *Data in Brief*, vol. 63, Art. no. 111972, 2025, doi: 10.1016/j.dib.2025.111972.
- [19] T. N. Vu, “PhishVN: A time-stamped Vietnamese URL phishing dataset with impersonation-scenario labels and confidence tiers,” *Data in Brief*, Art. no. 113195, 2026.
- [20] Common Crawl, “Get started,” Common Crawl, 2026. [Online]. Available: <https://commoncrawl.org/get-started>. Accessed: Sep. 3, 2026.
- [21] M. Ejaz, A. Mian, and A. Manzoor, “Life-long phishing attack detection using continual learning,” *Scientific Reports*, vol. 13, Art. no. 11488, 2023, doi: 10.1038/s41598-023-37552-9.
- [22] G. Apruzzese, M. Conti, and Y. Yuan, “SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning,” in *Proc. Annual Computer Security Applications Conf. (ACSAC)*, 2022, pp. 171–185, doi: 10.1145/3564625.3567980.
- [23] Y. Yuan, G. Apruzzese, and M. Conti, “Multi-SpacePhish: Extending the evasion-space of adversarial attacks against phishing website detectors using machine learning,” *Digital Threats: Research and Practice*, vol. 5, no. 2, Art. no. 16, 2024, doi: 10.1145/3638253.
- [24] J. Uddin *et al.*, “Explainable machine learning for phishing site detection: A high-efficiency approach using boosting models and SHAP,” *The Journal of Engineering*, vol. 2025, no. 1, Art. no. e70110, 2025, doi: 10.1049/tje.2.70110.
- [25] T. Kehkashan *et al.*, “Explainable phishing website detection for secure and sustainable cyber infrastructure,” *Scientific Reports*, vol. 15, Art. no. 41751, 2025, doi: 10.1038/s41598-025-27984-w.
- [26] H. A. Khan *et al.*, “Explainable AI for phishing URL detection: A Bayesian-optimized stacking ensemble framework with SHAP-guided feature learning,” *Frontiers in Artificial Intelligence*, vol. 9, Art. no. 1854934, 2026, doi: 10.3389/frai.2026.1854934.
- [27] C. Mayo, M. Tchuindjang, S. Brohi, and N. Ersotelos, “A temporally dynamic feature-extraction framework for phishing detection with LIME and SHAP explanations,” *Future Internet*, vol. 18, no. 2, Art. no. 101, 2026, doi: 10.3390/fi18020101.