International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 39 العدد
Part 1 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection

http://www.doi.org/10.62341/istj-vol39-1-ad48

Received	**2026/08/20**	**تم استلام الورقة العلمية في**
Accepted	**2026/09/08**	**تم قبول الورقة العلمية في**
Published	**2026/09/10**	**تم نشر الورقة العلمية في**

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection

Albahlool Abood
Faculty of Information Technology, University of Gharyan
Gharyan, Libya
Albahlool.abood@gu.edu.ly,
https://orcid.org/0009-0007-3264-3720

Abstract

Feature reduction is widely used in machine learning based phishing URL detection to lower model complexity while retaining detection performance. However, conclusions about how much reduction is acceptable may depend on how training and testing data are partitioned. This study examines whether conclusions about feature reduction remain stable when evaluation changes from random stratified to domain disjoint cross validation. Experiments were conducted on URL-Phish Version 2 using Mutual Information ranking within each training fold and four predefined feature budgets, from 22 to 5 features. Logistic Regression and Random Forest were evaluated under both protocols, with PR AUC as the primary metric. PR AUC decreased as the feature budget was reduced for both classifiers under both evaluation protocols, so the qualitative conclusion about feature reduction remained consistent within this experimental setting. However, the size of the protocol gap varied across classifiers and feature budgets: it was clearest for Random Forest under the smallest feature budget, whereas the Logistic Regression gaps were less clearly separated from fold-level variability. Domain disjoint evaluation also produced greater fold-level variability. The results show that claims about compact phishing URL representations should be interpreted together with the evaluation protocol used. Future work should examine whether

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

this pattern persists across additional datasets, classifiers, and domain grouping rules.
Keywords: Phishing URL Detection, Feature Reduction, Domain Disjoint Evaluation, Machine Learning, Mutual Information, Generalization

حساسية اختزال السمات لبروتوكول التقييم في اكتشاف روابط التصيد الاحتيالي

البهلول عبود
كلية تقنية المعلومات، جامعة غريان، ليبيا
Albahlool.abood@gu.edu.ly

الملخص

يُستخدم اختزال السمات على نطاق واسع في اكتشاف روابط التصيد الاحتيالي المعتمد على تعلم الآلة بهدف تقليل تعقيد النموذج مع الحفاظ على أداء الكشف. ومع ذلك، قد تعتمد الاستنتاجات المتعلقة بمدى إمكانية اختزال السمات على الطريقة المستخدمة في تقسيم بيانات التدريب والاختبار. تبحث هذه الدراسة فيما إذا كانت الاستنتاجات المتعلقة باختزال السمات تظل مستقرة عند الانتقال من التحقق المتقاطع العشوائي الطبقي إلى التحقق المتقاطع المنفصل حسب النطاق. أُجريت التجارب على مجموعة بيانات URL–Phish Version 2 باستخدام ترتيب السمات بطريقة Mutual Information داخل كل طية تدريب، مع أربع ميزانيات محددة مسبقًا للسمات تتراوح من 22 إلى 5 سمات. وتم تقييم كل من Logistic Regression و Random Forest تحت كلا البروتوكولين، مع استخدام PR AUC بوصفه المقياس الأساسي. انخفضت قيمة PR AUC مع تقليل ميزانية السمات لدى كلا المصنفين وتحت كلا بروتوكولي التقييم، ولذلك ظل الاستنتاج النوعي بشأن اختزال السمات متسقًا ضمن هذا الإعداد التجريبي. ومع ذلك، اختلف حجم الفجوة بين بروتوكولي التقييم باختلاف المصنف وميزانية السمات؛ إذ كانت أوضح بالنسبة إلى Random Forest عند أصغر ميزانية للسمات، في حين كانت فجوات Logistic Regression أقل وضوحًا مقارنة بالتباين على مستوى الطيات. كما أدى التقييم المنفصل حسب النطاق إلى تباين أكبر على مستوى الطيات. وتوضح

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

النتائج أن الادعاءات المتعلقة بفعالية التمثيلات المختصرة لروابط التصيد الاحتيالي ينبغي تفسيرها بالاقتران مع بروتوكول التقييم المستخدم. وينبغي أن تبحث الدراسات المستقبلية فيما إذا كان هذا النمط يستمر عبر مجموعات بيانات ومصنفات وقواعد تجميع نطاقات إضافية.

الكلمات المفتاحية :اكتشاف روابط التصيد الاحتيالي، اختزال السمات، التقييم المنفصل حسب النطاق، تعلم الآلة، المعلومات المتبادلة، التعميم

1. Introduction

Phishing remains a persistent cybersecurity threat because deceptive websites and links are used to obtain sensitive information from users. Machine learning has been widely investigated for detecting suspicious URLs, particularly through lexical and structural features extracted directly from the URL. Such features can support lightweight detection without requiring complete webpage analysis or continuous access to external services (Bustio-Martínez et al., 2022; Kapan and Sora Gunal, 2023).

Phishing detection performance depends not only on the classifier but also on the information used to represent each URL. Although larger feature sets may contain useful discriminatory information, they can also increase model complexity and computational requirements. Feature selection has therefore been extensively studied as a way to retain informative inputs while reducing dimensionality. Previous studies have reported strong performance using reduced feature subsets and different selection strategies (Wazirali et al., 2021; Kocyigit et al., 2024; Vajrobol et al., 2024; Kaur and Jain, 2025). However, conclusions about reduced representations may also depend on how training and testing data are constructed. Random partitioning is commonly used to estimate machine learning performance, but in URL based phishing datasets it can place URLs from the same domain in both training and testing data. This does not by itself demonstrate information leakage or invalidate random evaluation, but it raises a generalization question because domain related characteristics may be shared across the partition boundary. Several studies have also shown that phishing detection performance can change when evaluation involves

http://www.doi.org/10.62341/istj-vol39-1-ad48

different datasets, domains, or stricter generalization conditions (Rashid et al., 2024; Ahamed et al., 2026). This issue is particularly relevant to feature reduction. A reduced feature set that performs well under random evaluation may not lead to the same conclusion when domain overlap is prevented. The present study therefore examines two related questions: whether the overall conclusion about feature reduction remains consistent across evaluation protocols, and how the random minus domain disjoint PR AUC gap varies across classifiers and feature budgets. These questions are addressed through a controlled comparison of random stratified and domain disjoint evaluation while keeping the dataset, feature ranking procedure, feature budgets, and classification models fixed. Figure 1 presents the conceptual design of the study and summarizes the comparison between random stratified and domain disjoint evaluation across the four feature budgets and two classifiers.

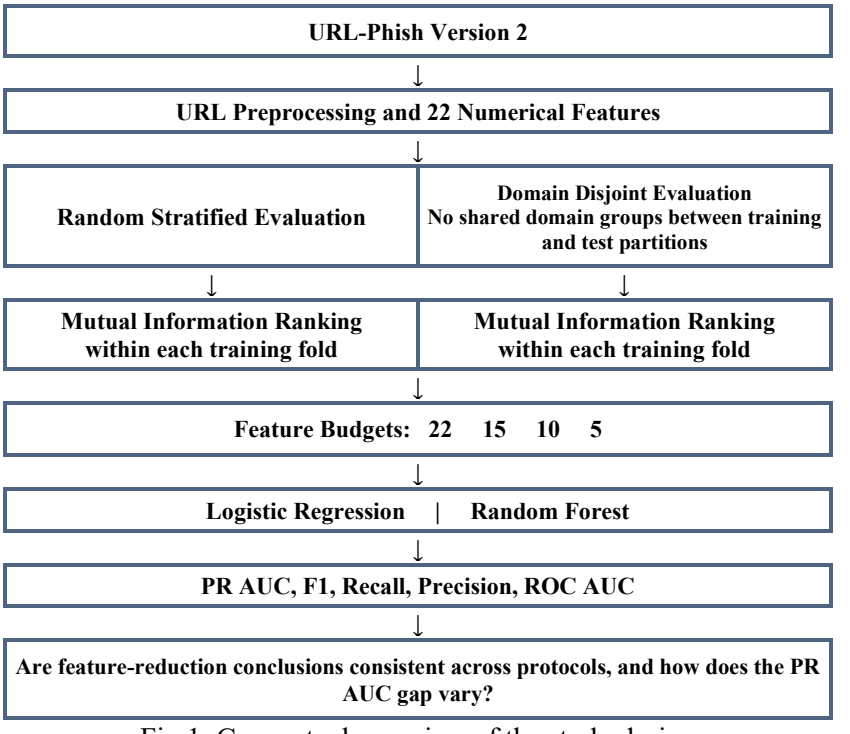

Fig.1. Conceptual overview of the study design

http://www.doi.org/10.62341/istj-vol39-1-ad48

Section 2 reviews related work and defines the research gap, Section 3 describes the methodology, and Section 4 presents and discusses the results. Sections 5 and 6 address the limitations and conclusions, respectively.

2. Related Work

2.1 Feature Reduction in Phishing URL Detection

Machine learning based phishing detection has been investigated using URL, webpage, and hybrid feature representations (Aljofey et al., 2022), while feature reduction has received particular attention for decreasing model complexity while preserving useful discriminatory information. Bustio-Martínez et al. (2022) showed that a compact subset of URL features could retain strong classification performance, while Kapan and Sora Gunal (2023) demonstrated that performance depends on both the selected feature representation and the classifier. More recent studies have explored optimization based feature selection, including Genetic Algorithm methods (Kocyigit et al., 2024), Rough Set Theory based selection (Setu et al., 2025), and optimized feature frameworks for URL based detection (Kaur and Jain, 2025).

Vajrobol et al. (2024) are particularly relevant to the present work because they combined Mutual Information with Logistic Regression and evaluated progressively reduced feature subsets. Their results showed that strong performance could be achieved with a small number of selected features. This provides an important reference for feature reduction, but their study did not investigate whether the conclusion remains stable when domain overlap between training and testing data is prevented.

2.2 Generalization and Evaluation Strategies

Another line of research has focused on the generalization of phishing detection models. Rashid et al. (2024) showed that performance can decrease considerably when models are evaluated across different datasets, indicating that results obtained within a single data distribution may not fully represent performance under distributional change.

Ahamed et al. (2026) extended this concern by evaluating phishing detectors under domain aware partitioning, adversarial URL modification, unseen attack families, and external validation. Their

http://www.doi.org/10.62341/istj-vol39-1-ad48

study demonstrated the importance of evaluation design when assessing robustness and generalization. However, their objective was broader robustness evaluation rather than determining how feature reduction behaves under different partitioning strategies. Nagunwa (2024) also examined practical generalization through real time and zero day phishing detection, emphasizing the role of feature representation and evaluation conditions in model performance.

2.3 Research Gap and Study Positioning

Previous research therefore establishes two important findings. First, reduced feature sets can provide strong phishing detection performance. Second, model performance can change when evaluation conditions become more demanding. The studies reviewed above do not examine these two issues together as a controlled feature reduction problem.

These studies also differ in what they treat as evidence of generalization. Feature reduction studies primarily evaluate predictive performance within a fixed partitioning scheme, whereas generalization oriented studies vary datasets, domains, or robustness conditions without isolating feature budget effects. Consequently, the existing evidence does not establish whether a favorable conclusion about a reduced representation is preserved when the partitioning strategy changes while the dataset, feature ranking procedure, feature budgets, and classifiers are held fixed.

The present study focuses on this intersection. It compares the same Mutual Information ranking procedure, feature budgets, and classifiers under random stratified and domain disjoint evaluation. The purpose is not to introduce a new feature selection method or classifier, but to determine whether conclusions about reduced feature performance remain consistent when domain overlap between training and testing data is removed.

Table 1 summarizes the studies most closely related to this research and highlights their relationship to the present evaluation design.

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

Table 1. Studies most closely related to the present evaluation design

Study	Approach	Main finding	Relation to this study
Bustio-Martínez et al. (2022)	URL feature reduction	Strong performance with a compact subset	Feature reduction focus
Kapan and Sora Gunal (2023)	Feature groups and classifiers	Performance depended on representation and classifier	Supports controlled feature evaluation
Kocyigit et al. (2024)	Genetic Algorithm feature selection	Reduced unnecessary features while maintaining performance	Selection optimization focus
Vajrobol et al. (2024)	Mutual Information with Logistic Regression	Strong performance with reduced subsets	Closest feature reduction comparator
Rashid et al. (2024)	Cross dataset generalization	Performance decreased across datasets	Supports stricter generalization evaluation
Ahamed et al. (2026)	Domain aware robustness evaluation	Evaluation conditions affected observed robustness	Closest evaluation protocol comparator

3. Methodology

3.1 Dataset and Preprocessing

The experiments used URL-Phish Version 2 obtained from Mendeley Data (Linh and Hung, 2026). The downloaded Version 2 CSV contained 116,600 records, including 100,000 benign and 16,600 phishing URLs, together with lexical and structural URL attributes. The associated Data in Brief article describes the earlier dataset release and its feature construction (Linh and Hung, 2025). The present study uses the later Mendeley Version 2, and all sample counts reported here refer specifically to that version.

This study used the 22 numerical features provided in the dataset. The string reference columns url, dom, and tld were not used as classifier input features. The original URL string was used only for preprocessing and domain group construction. Exact duplicate URL strings were removed while retaining the first occurrence, resulting

http://www.doi.org/10.62341/istj-vol39-1-ad48

in 115,231 samples, including 98,641 benign and 16,590 phishing URLs. No conflicting labels were found among the duplicated URLs. The original dataset file was retained unchanged for reproducibility.

According to the associated dataset documentation, the benign and phishing samples originate from different source populations. Benign URLs were collected from trusted and institutional sources, whereas phishing URLs were obtained from PhishTank (Linh and Hung, 2025). This source separation may introduce source related class confounding and should therefore be considered when interpreting the reported performance. Domain disjoint evaluation prevents domain groups from appearing in both training and testing folds, but it does not remove this source related limitation. This issue is considered further in Section 5.

3.2 Domain Group Construction

Domain groups were derived from the original URL string. Each hostname was normalized and mapped to its effective top level domain plus one, using the packaged public suffix information available through tldextract version 5.3.0. IP based URLs were grouped using the canonical IP address. When a valid public suffix was unavailable, the normalized hostname was used as a fallback. This procedure produced 90,859 unique domain groups from the 115,231 retained URLs.

3.3 Evaluation Protocols

Two evaluation protocols were compared. The first used StratifiedKFold with five folds, shuffle=True, and random_state=0, while preserving the class distribution across folds.

The second protocol used StratifiedGroupKFold with five folds, **shuffle = True**, and **random_state = 0**, where the domain group was treated as the grouping variable. All URLs belonging to the same domain group were therefore assigned entirely to either training or testing data within each fold. This produced zero domain group overlap between training and testing partitions.

Both protocols used the same dataset, models, feature ranking procedure, feature budgets, and evaluation metrics. The partitioning strategy was the only experimental factor varied by design. Because feature ranking was recomputed within each training fold, however,

changing the partitioning could also alter the selected feature subsets through changes in training composition. The extent of this effect is examined in Section 4.5.

3.4 Feature Ranking and Feature Budgets

Mutual Information was selected as a classifier independent filter method so that the same feature ranking procedure could be applied consistently to both classifiers. It was used to rank the 22 numerical URL features with random_state=0. Feature ranking was performed separately within the training portion of each fold to prevent information from the corresponding test partition from influencing feature selection. The resulting rankings were evaluated using four feature budgets containing the top 22, 15, 10, and 5 features.

The four feature budgets were predefined to represent the full 22 feature set and progressively more compact representations, rather than being selected retrospectively to maximize performance. No feature ranking calculated from the complete dataset was used during model evaluation.

3.5 Classification Models

Two established classifiers were selected to represent complementary learning behaviours rather than to provide an exhaustive comparison of machine learning algorithms. Logistic Regression was included as a linear probabilistic baseline, while Random Forest represented a nonlinear tree based ensemble. Logistic Regression was implemented within a pipeline containing feature standardization, while Random Forest was trained using 200 trees. Both models used balanced class weights to account for the unequal distribution of benign and phishing samples.

The decision threshold for class labels was fixed at 0.5. No oversampling or synthetic sample generation was applied.

To quantify the practical computational effect of feature reduction, classifier training time was measured in a separate benchmark using the same cross validation splits, feature ranking procedure, feature budgets, and model parameters as the main experiment. Wall clock time was measured for model fitting only and summarized as mean ± standard deviation across the five folds. Mutual Information ranking time was excluded because the ranking was computed once per training fold using all 22 features and was common to the four

http://www.doi.org/10.62341/istj-vol39-1-ad48

feature budgets. The timing results are intended for relative comparison within the same Google Colab runtime rather than as hardware independent performance measurements.

3.6 Evaluation Metrics

Precision Recall Area Under the Curve, PR AUC, was selected as the primary metric because the dataset contains substantially fewer phishing than benign samples. F1 score, recall, precision, and ROC AUC were used as secondary measures.

Because domain disjoint partitioning produced some variation in phishing prevalence across test folds, ROC AUC and a prevalence normalized form of PR AUC were also examined as descriptive robustness checks. These additional analyses were used only to assess whether prevalence differences could reasonably explain the observed protocol gaps and were not treated as replacements for the primary PR AUC results.

For each PR AUC protocol gap, descriptive uncertainty was summarized using SE = sqrt ($SD\_random^2/5 + SD\_domain^2/5$), where the standard deviations were calculated across the five folds of each protocol. A descriptive 95% uncertainty interval was then calculated as the observed gap ± 1.96 SE. These intervals were used only to summarize the magnitude of fold-level dispersion around the observed protocol gaps. Formal hypothesis tests were not applied to the five fold scores because k-fold cross-validation folds are statistically dependent through their overlapping training partitions, and the random stratified and domain disjoint protocols do not generate matched test folds that can be treated as paired independent observations. Bengio and Grandvalet (2004) showed that no general unbiased estimator of the variance of k-fold cross-validation is available. Consequently, conventional tests applied directly to these fold scores could give an unjustified impression of inferential precision. The reported intervals are therefore interpreted as descriptive summaries rather than formal confidence intervals or tests of statistical significance.

Prevalence normalized PR AUC was calculated as:

$$\boldsymbol{Normalized\ PR\ AUC = \frac{PR\ AUC - Prevalence}{1 - Prevalence}}$$

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

Table 2 summarizes the complete experimental configuration. Overall, the experimental design consisted of two evaluation protocols, two classifiers, four feature budgets, and five folds, resulting in 80 model fits.

Table 2. Experimental configuration

Item	Setting
Dataset	URL-Phish Version 2
Original samples	116,600
Samples after URL deduplication	115,231
Numerical features	22
Domain grouping	eTLD+1, canonical IP, hostname fallback
Feature ranking	Mutual Information within training folds
Feature budgets	22, 15, 10, 5
Evaluation protocols	Random Stratified, Domain Disjoint
Cross validation	5 folds
Classifiers	Logistic Regression, Random Forest
Random Forest trees	200
Class handling	Balanced class weights
Decision threshold	0.5
Primary metric	PR AUC
Secondary metrics	F1, Recall, Precision, ROC AUC
Total model fits	80

4. Results and Discussion

4.1 Domain Overlap under Random Evaluation

Before comparing classification performance, the amount of domain overlap produced by conventional random partitioning was examined (Figure 2). Across 20 preliminary random splits, the median overlap rate was 14.93% for benign URLs and 78.48% for phishing URLs. The much larger phishing overlap indicates that many phishing observations belong to domain groups represented by more than one URL. Consequently, a random split can place different URLs from the same domain group on both sides of the training and testing boundary.

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

The same pattern was observed in the final five fold random evaluation, where between 2,129 and 2,197 domain groups were shared between training and testing in each fold. In contrast, the domain disjoint protocol produced zero shared domain groups in every fold. This difference establishes the intended experimental contrast. It should not be interpreted as proof that random evaluation is invalid or that its performance is necessarily inflated. Rather, random and domain disjoint evaluation answer different generalization questions, since the latter requires predictions on domain groups not represented during training.

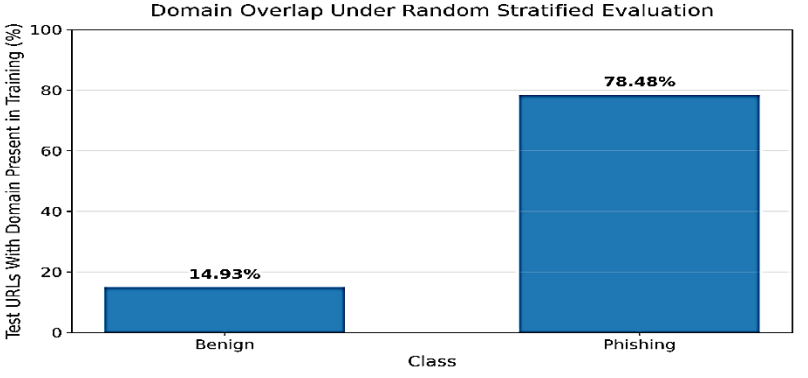

Fig.2. Median domain overlap under random stratified evaluation.

4.2 Performance across Feature Budgets

Figure 3 and Table 3 show the primary PR AUC results across feature budgets. The use of all 22 features produced the highest mean PR AUC for both classifiers under both protocols. Logistic Regression achieved 0.8984 under random evaluation and 0.8581 under domain disjoint evaluation. Random Forest performed more strongly, reaching 0.9723 and 0.9345, respectively.

Reducing the feature budget decreased PR AUC in both models. For Logistic Regression, the mean PR AUC declined from 0.8984 to 0.7775 under random evaluation and from 0.8581 to 0.7288 under domain disjoint evaluation when the feature budget was reduced from 22 to 5. Random Forest showed a larger reduction, from 0.9723 to 0.8061 under random evaluation and from 0.9345 to 0.7009 under domain disjoint evaluation. Therefore, within the

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection

http://www.doi.org/10.62341/istj-vol39-1-ad48

present dataset, Mutual Information ranking procedure, and the two classifiers evaluated, the reduced feature representations did not preserve the PR AUC of the full 22 feature representation**.**

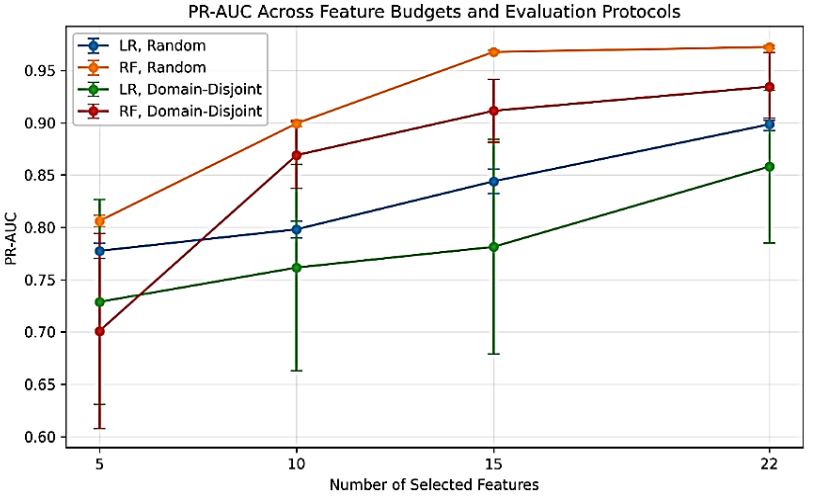

Fig.3. PR AUC across feature budgets under both evaluation protocols.

Table 3. Main performance across feature budgets

Model	Features	Random PR AUC	Domain PR AUC	Random F1	Domain F1
LR	22	0.8984 ± 0.0059	0.8581 ± 0.0731	0.7832 ± 0.0019	0.7445 ± 0.0848
LR	15	0.8440 ± 0.0118	0.7813 ± 0.1026	0.6936 ± 0.0119	0.6499 ± 0.1220
LR	10	0.7981 ± 0.0080	0.7615 ± 0.0985	0.6380 ± 0.0025	0.6006 ± 0.1190
LR	5	0.7775 ± 0.0073	0.7288 ± 0.0979	0.6322 ± 0.0020	0.5604 ± 0.0871
RF	22	0.9723 ± 0.0016	0.9345 ± 0.0323	0.9247 ± 0.0006	0.8717 ± 0.0333
RF	15	0.9676 ± 0.0015	0.9114 ± 0.0299	0.9115 ± 0.0008	0.8318 ± 0.0460
RF	10	0.8993 ± 0.0030	0.8690 ± 0.0318	0.8126 ± 0.0057	0.7818 ± 0.0339
RF	5	0.8061 ± 0.0055	0.7009 ± 0.0931	0.6946 ± 0.0039	0.5980 ± 0.0770

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection

http://www.doi.org/10.62341/istj-vol39-1-ad48

F1 (Figure 4) followed the same broad pattern, generally decreasing as the feature budget was reduced and under domain disjoint evaluation. The largest loss occurred for Random Forest at five features, consistent with the PR AUC results.

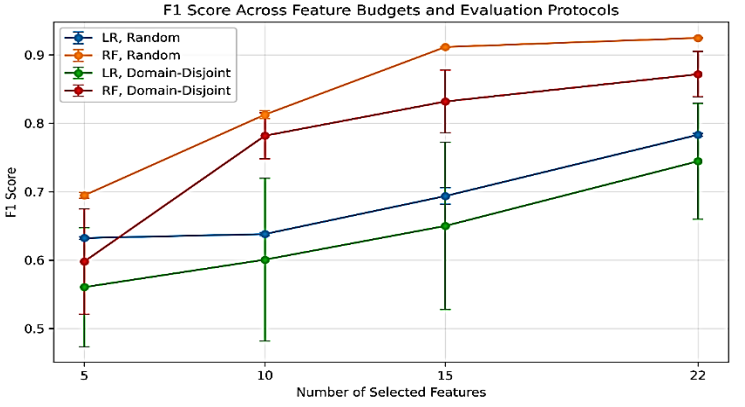

Fig.4. F1 score across feature budgets under the two evaluation protocols.

Recall (Figure 5) showed the same model contrast. For Logistic Regression, recall changed from 0.9260 to 0.9017 under random evaluation and from 0.9060 to 0.8334 under domain disjoint evaluation as the feature budget fell from 22 to 5. Random Forest declined more sharply, from 0.9077 to 0.7574 and from 0.8293 to 0.6595, respectively.

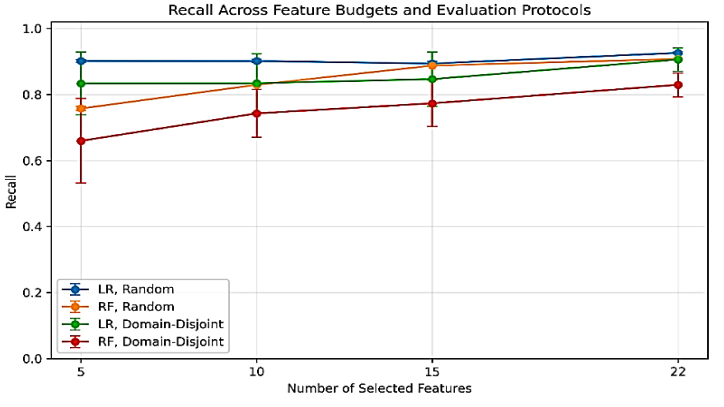

Fig.5. Recall across feature budgets under the two evaluation protocols.

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

Precision is reported separately in Figure 6 to complete the operating point view of the classifiers. Precision, recall, and F1 depend on the fixed decision threshold of 0.5 and are therefore interpreted as secondary evidence. The main conclusions of this study rely primarily on PR AUC and are subsequently checked using ROC AUC, both of which evaluate ranking performance independently of a single classification threshold.

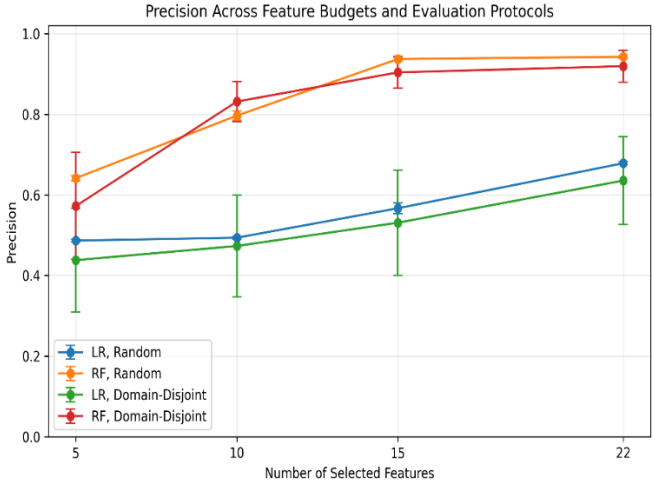

Fig.6. Precision across feature budgets under the two evaluation protocols.

The computational benchmark showed that the practical benefit of feature reduction was model dependent. For Logistic Regression, reducing the feature budget from 22 to 5 decreased mean training time from 0.7713 ± 0.0859 s to 0.1964 ± 0.0291 s under random evaluation, a reduction of 74.5%. Under domain disjoint evaluation, training time decreased from 0.6889 ± 0.0930 s to 0.1474 ± 0.0538 s, a reduction of 78.6%. In contrast, Random Forest training time changed only from 11.5013 ± 0.0897 s to 11.1727 ± 0.2439 s under random evaluation and from 11.1337 ± 0.4912 s to 10.7459 ± 0.5860 s under domain disjoint evaluation, corresponding to reductions of only 2.9% and 3.5%, respectively. Thus, reducing the number of input features produced substantial training-time savings for

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

Logistic Regression but did not translate into comparable savings for Random Forest.

4.3 Protocol Gap and Descriptive Uncertainty

Although the mean PR AUC was higher under random evaluation for every model and feature budget, the magnitude of this difference was not constant. The Random Forest gap was 0.0378 with 22 features, 0.0563 with 15 features, 0.0303 with 10 features, and 0.1052 with 5 features. The five feature configuration therefore produced the largest observed protocol sensitivity. However, the intermediate budgets do not follow a monotonic sequence. It would consequently be inappropriate to claim that the protocol gap necessarily increases whenever the feature budget becomes smaller. Logistic Regression behaved differently. Its random minus domain disjoint gaps ranged from 0.0366 to 0.0626 and were accompanied by much larger fold dispersion under domain disjoint evaluation. As shown in Figure 7, all four descriptive uncertainty intervals for Logistic Regression include zero. By contrast, the four Random Forest intervals remain above zero. These intervals are descriptive rather than inferential and should not be interpreted as conventional statistical significance tests. They indicate that the observed Random Forest gaps were more consistently separated from zero relative to fold-level dispersion, whereas the Logistic Regression gaps were less clearly separated from that variability.

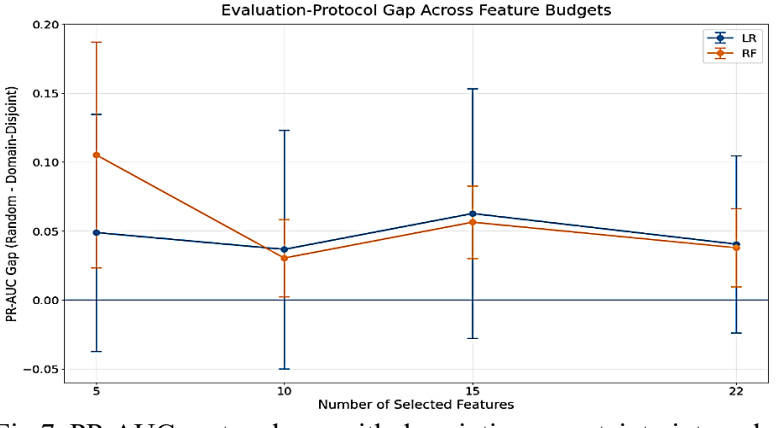

Fig.7. PR AUC protocol gap with descriptive uncertainty intervals.

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

Table 4 summarizes the protocol gaps and the accompanying robustness diagnostics.

Table 4. Protocol gap and robustness diagnostics

Model	Features	PR AUC gap	Descriptive 95% interval	ROC AUC gap	Normalized PR AUC gap
LR	22	0.0404	−0.0239 to 0.1046	0.0054	0.0414
LR	15	0.0626	−0.0279 to 0.1531	0.0138	0.0663
LR	10	0.0366	−0.0500 to 0.1232	0.0085	0.0361
LR	5	0.0488	−0.0373 to 0.1348	0.0156	0.0538
RF	22	0.0378	0.0095 to 0.0662	0.0067	0.0416
RF	15	0.0563	0.0300 to 0.0825	0.0171	0.0657
RF	10	0.0303	0.0023 to 0.0583	0.0037	0.0335
RF	5	0.1052	0.0234 to 0.1869	0.0385	0.1210

The distinction between the two classifiers is important. The results do not support a universal statement that changing the evaluation protocol has the same effect on every classifier. Instead, protocol sensitivity is model dependent, with the clearest difference observed for Random Forest, particularly under aggressive feature reduction.

4.4 Robustness to Fold Prevalence

Domain disjoint grouping constrains how samples are distributed across folds. Phishing prevalence was approximately 14.40% in each random test fold but ranged from 8.72% to 26.47% across domain disjoint folds. Because PR AUC is prevalence sensitive, this variation may have contributed to the greater dispersion under domain disjoint evaluation.

Two diagnostics were examined. ROC AUC remained higher under random evaluation for every model and feature budget, while prevalence normalized PR AUC gaps were positive in all eight comparisons. For Random Forest with five features, the PR AUC, ROC AUC, and normalized PR AUC gaps were 0.1052, 0.0385, and 0.1210, respectively.

These checks do not rule out an influence of prevalence, but they indicate that prevalence variation alone does not appear sufficient to explain the protocol differences. As noted in Section 3.6, these

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

comparisons are descriptive and are used to assess directional consistency rather than formal inference.

4.5 Fold Variability and Feature Selection Consistency

Fold to fold variation was substantially greater under domain disjoint evaluation (Figure 8). For Random Forest, PR AUC standard deviation increased from 0.0016 to 0.0323 at 22 features and from 0.0055 to 0.0931 at five features. Logistic Regression showed the same pattern at 22 features, increasing from 0.0059 to 0.0731.

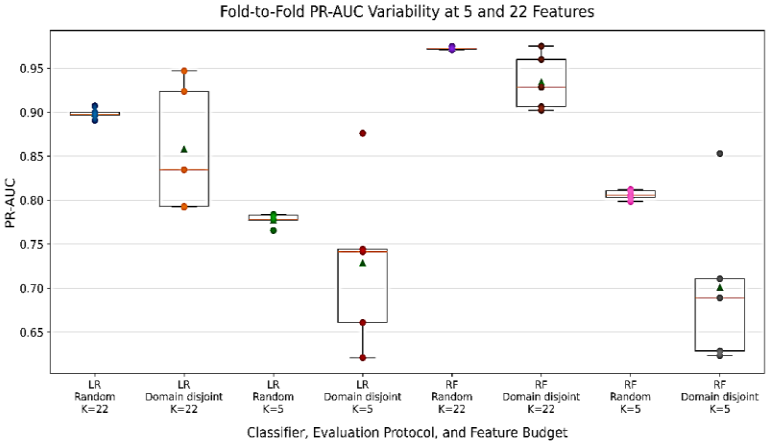

Fig.8. Fold variability in PR AUC at feature budgets of 5 and 22.

This greater variation should not automatically be interpreted as learning algorithm instability. Each domain disjoint fold contains a different set of unseen domain groups, so part of the dispersion reflects the evaluation setting itself. With only five folds, however, these standard deviations should be interpreted as observed dispersion rather than precise estimates of population variability.

The Mutual Information results (Figure 9) also showed substantial selection consistency. Entropy, digit ratio, digit count, and special character ratio appeared in the top five in all ten training folds. HTTPS status appeared in nine folds, while URL length appeared only in the remaining domain disjoint fold. The top five feature set was therefore identical in nine of the ten folds.

Feature Reduction Sensitivity to Evaluation Protocol in Phishing URL Detection
http://www.doi.org/10.62341/istj-vol39-1-ad48

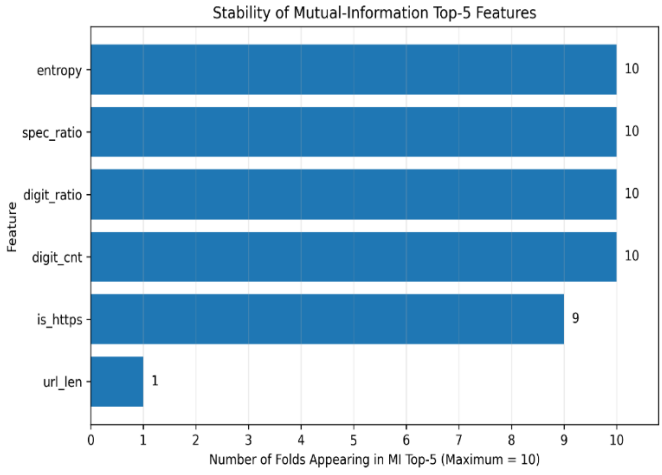

Fig.9. Stability of top five Mutual Information features across folds.

Selection was also stable at the intermediate budgets. At 10 features, nine appeared in all ten training folds. The tenth differed by protocol: all five random folds selected dash_cnt, while the domain disjoint folds selected dom_len in three cases, tld_len in one, and dash_cnt in one. This shows that training composition affected the marginal ranking at this budget. At 15 features, twelve features appeared in all ten folds, with the remaining variation concentrated near the selection boundary.

Overall, the selected subsets were largely consistent across folds, with differences concentrated among lower ranked features near the selection boundary. The observed performance differences are therefore unlikely to be explained simply by entirely different feature sets under the two protocols, since the most consistently selected features were preserved.

4.6 Discussion and Implications

The central finding is not that random evaluation is incorrect or that domain disjoint evaluation should replace it. Within URL-Phish Version 2, using Mutual Information ranking and the two classifiers evaluated here, reducing the feature budget lowered PR AUC under both protocols. The qualitative conclusion about feature reduction therefore remained unchanged within this experimental setting. What varied was the random minus domain disjoint PR AUC gap, which was largest for Random Forest at five features.

http://www.doi.org/10.62341/istj-vol39-1-ad48

Previous studies have shown that compact URL representations can achieve strong performance under particular datasets and experimental settings (Bustio-Martínez et al., 2022; Vajrobol et al., 2024). The present results do not contradict those findings, but they show why the evaluation protocol should accompany claims about compact representations. Random Forest achieved the strongest overall PR AUC with larger feature budgets but also the clearest protocol difference at five features. Logistic Regression had lower overall PR AUC, while its protocol differences were less decisive relative to domain disjoint fold variability. The observed gap therefore appears to depend jointly on the feature representation, classifier, and composition of the domain groups.

The computational results also show that reduced representations should not automatically be equated with proportional computational savings. Reducing the feature budget from 22 to 5 produced substantial reductions in Logistic Regression training time, with decreases of 74.5% under random evaluation and 78.6% under domain disjoint evaluation. In contrast, the corresponding reductions for Random Forest were only 2.9% and 3.5%. These results indicate that the practical computational benefit of feature reduction depends on the learning algorithm as well as the number of input features**.**

The domain overlap analysis provides a plausible explanation for the protocol differences. Under random splitting, URLs from the same domain group can appear in both training and testing data, whereas domain disjoint evaluation requires predictions on groups not represented during training. This is consistent with previous findings that phishing detection performance can change across datasets or under stricter generalization settings (Rashid et al., 2024; Ahamed et al., 2026). However, domain overlap alone does not establish information leakage, and the observed differences should not be interpreted as a pure leakage effect.

A practical implication is that studies reporting strong performance from reduced URL feature sets should describe how training and testing samples were partitioned. Where domain information is available, domain disjoint results can complement conventional random evaluation by showing how a reduced representation

http://www.doi.org/10.62341/istj-vol39-1-ad48

behaves on unseen domain groups. The computational findings further suggest that claims about lightweight or efficient representations should be supported by direct computational measurements rather than inferred from feature count alone.

These recommendations follow directly from the observed results: the feature reduction trend remained qualitatively consistent across protocols, while the size of the protocol gap and the computational benefit varied with classifier and feature budget.

Finally, the results do not support a universal rule that smaller feature budgets produce larger protocol gaps. The largest gap occurred for Random Forest at five features, while the intermediate budgets were less regular. The direction of the feature reduction result was consistent across protocols, but the size of the protocol gap varied with classifier and feature budget.

To summarize the relationship between the study questions and the experimental findings, Table 5 maps each research question to its corresponding result.

Table 5. Research questions and corresponding findings

Research question	Corresponding finding
RQ1. Does the overall conclusion about feature reduction remain consistent across evaluation protocols?	Within this experimental setting, yes. PR AUC decreased as the feature budget was reduced under both protocols and for both classifiers
RQ2. How does the random minus domain disjoint PR AUC gap vary across classifiers and feature budgets?	The gap was positive in all eight comparisons but varied nonmonotonically. It was largest for Random Forest at five features (0.1052), while Logistic Regression gaps were less clearly separated from fold-level variability.

5. Limitations

This study has several limitations that should be considered when interpreting the results. First, the experiments were conducted using a single phishing URL dataset and two classification models, Logistic Regression and Random Forest. The findings describe the behaviour observed for URL-Phish Version 2, Mutual Information based feature ranking, and the two evaluated classifiers, and should not be generalized to other phishing datasets, classifiers, or feature

http://www.doi.org/10.62341/istj-vol39-1-ad48

selection procedures without further validation**.** In addition, the benign and phishing samples originate from different source populations. According to the dataset documentation, the benign subset was derived from trusted sources, including Research Organization Registry based data, whereas the phishing subset was obtained from PhishTank. This separation may introduce source related class confounding. Domain disjoint evaluation prevents domain groups from appearing in both training and testing folds, but it does not remove this source related limitation. Because the same source construction is retained under both evaluation protocols, this limitation should be considered when interpreting both the absolute performance values and the protocol differences reported here.

The domain disjoint folds also produced variation in phishing prevalence because complete domain groups had to remain intact. ROC AUC and prevalence normalized PR AUC were examined as descriptive robustness checks, but they cannot eliminate all effects of fold composition. Furthermore, only five folds were used, so the reported standard deviations and uncertainty intervals should be interpreted descriptively rather than as formal statistical inference. Accordingly, the study does not make claims of statistical significance for differences between evaluation protocols or feature budgets. Finally, precision, recall, and F1 were calculated using a fixed decision threshold of 0.5, meaning that these measures reflect one operating point of each classifier. For these reasons, the conclusions are restricted to the evaluation protocol sensitivity observed within the controlled conditions of this study.

6. Conclusions

This study examined whether conclusions about feature reduction in phishing URL detection remain stable when evaluation changes from random stratified to domain disjoint cross validation. Within URL-Phish Version 2, using Mutual Information ranking and the Logistic Regression and Random Forest models evaluated here, PR AUC decreased as the feature budget was reduced under both protocols. The qualitative conclusion was therefore consistent within this experimental setting: the reduced feature sets did not preserve the performance of the full 22 feature representation. What

http://www.doi.org/10.62341/istj-vol39-1-ad48

changed was the size of the random minus domain disjoint PR AUC gap across the two classifiers and feature budgets.

The clearest protocol difference was observed for Random Forest, where the PR AUC gap reached 0.1052 at five features. The descriptive uncertainty intervals for the Random Forest gaps remained above zero at all four feature budgets, whereas all four Logistic Regression intervals included zero. The Logistic Regression differences were therefore less decisive relative to fold variability. Additional checks using ROC AUC and prevalence normalized PR AUC followed the same overall direction and indicated that unequal phishing prevalence across folds did not appear to fully explain the observed differences. These comparisons remain descriptive, however, and the five fold design limits how precisely the protocol differences can be estimated.

The practical implication is that claims about the performance of compact phishing URL representations should be reported together with the evaluation protocol used to obtain them. Where domain information is available, domain disjoint evaluation can complement conventional random evaluation by showing how a reduced representation behaves when test domain groups are not represented during training. Future work should examine whether the pattern observed here persists across additional datasets, classifiers, and domain grouping rules.

References

Ahamed, T., Kakon, S. C., Farid, F. A., Uddin, J., and Abdul Karim, H. B., 2026, “An integrated evaluation protocol for adversarial robustness, generalization, and explanation stability in URL-based phishing detection,” *Frontiers in Computer Science*, 8, 1834407. DOI: 10.3389/fcomp.2026.1834407.

Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., and Wang, Y., 2022, “An effective detection approach for phishing websites using URL and HTML features,” *Scientific Reports*, 12, 8842. DOI: 10.1038/s41598-022-10841-5.

Bengio, Y., and Grandvalet, Y., 2004, “No unbiased estimator of the variance of k-fold cross-validation,” *Journal of Machine Learning Research*, 5, 1089–1105.

Bustio-Martínez, L., Álvarez-Carmona, M. A., Herrera-Semenets, V., Feregrino-Uribe, C., and Cumplido, R., 2022, "A lightweight data representation for phishing URLs detection in IoT environments," *Information Sciences*, 603, 42–59. DOI: 10.1016/j.ins.2022.04.059.

Kapan, S., and Sora Gunal, E., 2023, "Improved phishing attack detection with machine learning: A comprehensive evaluation of classifiers and features," *Applied Sciences*, 13(24), 13269. DOI: 10.3390/app132413269.

Kaur, K., and Jain, A. K., 2025, "OptiPhish: URL-based phishing detection using an optimised feature framework," *International Journal of Security and Networks*, 20(2), 93–109. DOI: 10.1504/IJSN.2025.146759.

Kocyigit, E., Korkmaz, M., Sahingoz, O. K., and Diri, B., 2024, "Enhanced feature selection using genetic algorithm for machine-learning-based phishing URL detection," *Applied Sciences*, 14(14), 6081. DOI: 10.3390/app14146081.

Linh, D. M., and Hung, T. C., 2025, "A feature-engineered dataset of benign and phishing URLs for machine learning and large language models evaluation," *Data in Brief*, 63, 112162. DOI: 10.1016/j.dib.2025.112162.

Linh, D. M., and Hung, T. C., 2026, "URL-Phish: A Feature-Engineered Dataset for Phishing Detection," *Mendeley Data*, Version 2. DOI: 10.17632/65z9twcx3r.2.

Nagunwa, T., 2024, "AI-driven approach for robust real-time detection of zero-day phishing websites," *International Journal of Information and Computer Security*, 23(1), 79–118. DOI: 10.1504/IJICS.2024.136735.

Rashid, F., Doyle, B., Han, S. C., and Seneviratne, S., 2024, "Phishing URL detection generalisation using Unsupervised Domain Adaptation," *Computer Networks*, 245, 110398. DOI: 10.1016/j.comnet.2024.110398.

Setu, J. H., Halder, N., Islam, A., and Amin, M. A., 2025, "RSTHFS: A Rough Set Theory-Based Hybrid Feature Selection Method for Phishing Website Classification," *IEEE Access*, 13, 68820–68830. DOI: 10.1109/ACCESS.2025.3561237.

Vajrobol, V., Gupta, B. B., and Gaurav, A., 2024, "Mutual information based logistic regression for phishing URL detection," *Cyber Security and Applications*, 2, 100044. DOI: 10.1016/j.csa.2024.100044.

Wazirali, R., Ahmad, R., and Abu-Ein, A. A. K., 2021, "Sustaining accurate detection of phishing URLs using SDN and feature selection approaches," *Computer Networks*, 201, 108591. DOI: 10.1016/j.comnet.2021.108591.