IJAICET
International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies
ISSN: 3139-8413 | Vol. 1 No. 2 (2026) | Original Research Article | www.ijaicet.com | DOI : https://doi.org/10.68307/a3mqv686

Robust Explainable and Calibrated Machine Learning for Phishing URL Detection under Constrained Feature Perturbations

Dhyanendra Singh Patel[1] Prakash Kumar Maravi[2] Kamlesh Raghuvanshi[3] Surabhi Karsoliya[4]
[1]M Tech Research Scholar [2]Assistant Professor Department of AIML [3]Head of Department [4]Management Member
Technocrats Institute of Technology-Excellence Bhopal Madhya Pradesh India

Abstract: Phishing detectors can achieve high benchmark accuracy while remaining vulnerable to feature manipulation and difficult to assess as trustworthy decision systems; this study evaluates leakage controlled discrimination calibration explanation stability and constrained adversarial robustness on the PhiUSIIL dataset through a normalized domain grouped development protocol and one frozen test assessment; five model families were compared under conservative and full sensitivity feature conditions while calibration permutation importance Tree SHAP explanations and greedy feature space evasion were analysed; conservative uncalibrated HistGradientBoosting achieved locked test recall of 0.999897 F1 of 0.999949 and Brier score of 0.000041 although attack success reached 0.7042 at three changes and 0.8462 at five changes while mean SHAP cosine similarity fell to minus 0.350; the study also presents a six layer operational architecture that connects evidence governance prediction explanation robustness controls and monitoring; the evidence shows that clean grouped performance and probability reliability can coexist with severe representation level vulnerability and therefore phishing model evaluation should report predictive calibration explanation architectural and adversarial evidence together.

Index Terms: phishing detection; trustworthy artificial intelligence; calibration; explainable artificial intelligence; adversarial robustness; SHAP; domain grouped evaluation

I. INTRODUCTION

Phishing detection research uses lexical, domain, content and learned representations because malicious pages can imitate legitimate visual and linguistic cues while changing infrastructure quickly; URL based machine learning supports fast screening [2,5,7,9,20] while combined URL and HTML systems can capture page structure and reference behaviour [4,6,22,25], yet high reported accuracy does not by itself show that confidence estimates are reliable or that evaluation groups prevent entity leakage [10,17].

Recent work has compared base classifiers, ensembles and deep architectures [3,8,11,12,13,18,19,23,24,26,27]; however temporal change [14], external validation [15], explanation quality [15,16] and problem space resistance [21] remain unevenly integrated, so a detector can appear decisive on a fixed split while relying on collection specific proxies or manipulable webpage attributes.

Trustworthy evaluation requires several linked forms of evidence because probability calibration concerns whether predicted confidence corresponds to observed outcomes [29], SHAP provides additive feature attributions for global and local interpretation [28] and adversarial evaluation asks whether feasible test time manipulation can evade the classifier [30]; this study combines these elements within one reproducible protocol rather than treating accuracy as the sole endpoint.

The study contributes a normalized domain grouped partition with explicit feature governance; a comparison of five model families under conservative and full sensitivity conditions; separate group probability calibration; cross fold permutation stability and SHAP explanation agreement; a constrained greedy feature space attack with explanation drift analysis; and a one time locked test evaluation with 1,000 paired domain cluster bootstrap replicates.

In addition the study proposes a deployment oriented architecture that separates evidence acquisition governance prediction explanation robustness control and continuous monitoring; this separation is important because a trustworthy detector must support traceable decisions while preventing raw identifiers high risk proxies and unverified model updates from bypassing the registered evaluation process.

II. RELATED WORK

A. Feature Representations and Classifiers

Lexical approaches derive length, character, token and domain patterns directly from URLs [2,5,18,20] while page aware systems add HTML structure, resource references, forms, scripts and visual or title correspondence [4,6,22,25]; the PhiUSIIL framework extends this direction through a diverse security profile with similarity and incremental learning features [1], which makes it useful for broad modelling but also requires sensitivity analysis when derived similarity fields may act as strong proxies.

Tree ensembles and stacking frequently improve benchmark performance because they model nonlinear interactions among heterogeneous indicators [3,4,12,26,27] while convolutional, attention and temporal architectures learn sequential URL representations [7,8,9,13,20,24]; nevertheless model complexity increases the need for explanation and efficiency evidence because comparable headline accuracy can conceal different confidence behaviour and operational cost.

B. Reliability Explainability and Robustness

Calibration research distinguishes discrimination from probability reliability and shows that post hoc methods must be evaluated rather than assumed to help [29]; phishing studies rarely report Brier score, log loss and expected calibration error together, so this study compares uncalibrated, sigmoid and isotonic probabilities through domain disjoint training, calibration and validation subsets.

Explainable phishing research increasingly uses SHAP and hybrid feature selection [15,16] while the SHAP framework

IJAICET
International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies
ISSN: 3139-8413 | Vol. 1 No. 2 (2026) | Original Research Article | www.ijaicet.com | DOI : https://doi.org/10.68307/a3mqv686

defines consistent additive attributions for complex predictions [28]; this study supplements global SHAP magnitude with permutation loss, cross fold rank correlation, top feature overlap and representative error explanations because a single importance chart cannot establish stability.

Adversarial learning research demonstrates that security classifiers can be evaded through test time manipulation [30] while DeltaPhish examines phishing detection in compromised websites [21]; therefore the present threat model restricts changes to registered URL deployment and webpage feature fields, reports the feature space limitation explicitly and measures both predictive degradation and explanation drift.

III. MATERIALS AND METHODS

A. Dataset Audit and Governance

The official UCI PhiUSIIL file contained 235,795 rows and 56 columns with 100,945 phishing records encoded as 0 and 134,850 legitimate records encoded as 1; integrity checks found no missing cells or exact duplicate rows but identified 7,588 rows involved in normalized URL duplication, 43 normalized URLs with conflicting labels and 22,563 rows within repeated normalized domains [1].

TABLE I. DATASET AUDIT AND REGISTERED PARTITION

Measure	Development	Test	Total
Phishing records	81,498	19,447	100,945
Legitimate records	107,721	27,129	134,850
All records	189,219	46,576	235,795
Normalized domain overlap	0 across partitions	0 across partitions	Not applicable

The split hashed normalized domains with seed 20261003 and a 0.20 test threshold so every record from one normalized domain remained in one partition; the test partition stayed sealed through model selection, calibration, explainability and adversarial analysis while its protocol was committed before the single permitted access.

B. Feature Conditions and Models

The conservative condition excluded filename, raw URL, domain, title and top level domain identifiers plus URLSimilarityIndex, CharContinuationRate, TLDLegitimateProb, URLCharProb, DomainTitleMatchScore and URLTitleMatchScore; the resulting 44 predictors formed the primary analysis while the 50 predictor full sensitivity condition restored the six derived fields and remained secondary because near perfect results could reflect collection specific proxy information.

TABLE II. REGISTERED MODELS AND EVALUATION ROLES

Model	Key configuration	Role
Logistic Regression	Balanced classes; standardized predictors	Reference baseline
Decision Tree	Depth 12; minimum leaf 20	Interpretable baseline
Random Forest	300 trees; balanced subsamples	Bagging ensemble
Extra Trees	300 trees; balanced classes	Randomized tree ensemble
HistGradientBoosting	250 iterations; learning rate 0.08	Selected ensemble candidate

Five fold Stratified Group K Fold validation used normalized domains as groups and treated phishing as the positive class; preprocessing was fitted inside each training fold while the registered metrics were accuracy, precision, recall, specificity, F1, ROC AUC, PR AUC, false negative rate, false positive rate, Brier score, log loss, expected calibration error, training time and inference time.

C. Proposed Trustworthy Detection Architecture

The proposed architecture organizes the detector as six connected layers consisting of URL and page evidence acquisition governance and feature control predictive and reliability estimation parallel decision and explanation services a robustness and policy gate and continuous monitoring; the arrangement converts the experimental protocol into an operational design while retaining a clear boundary between measured evidence and deployment policy.

The ingestion layer extracts registered lexical domain HTML and reference features while raw URL domain title filename and top level domain values remain available only for grouping provenance and audit; the governance layer validates the schema normalizes the registered domain blocks overlap and removes high risk proxy fields before any model call which makes the conservative feature condition an enforceable interface rather than an informal training choice.

The predictive layer applies the frozen HistGradientBoosting model and emits a phishing probability while the decision service applies the registered threshold and an organizational abstention rule; the explanation service produces local SHAP evidence and concise reason codes from the same governed feature vector so the probability and explanation cannot silently refer to different preprocessing states.

IJAICET
International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies
ISSN: 3139-8413 | Vol. 1 No. 2 (2026) | Original Research Article | www.ijaicet.com | DOI : https://doi.org/10.68307/a3mqv686

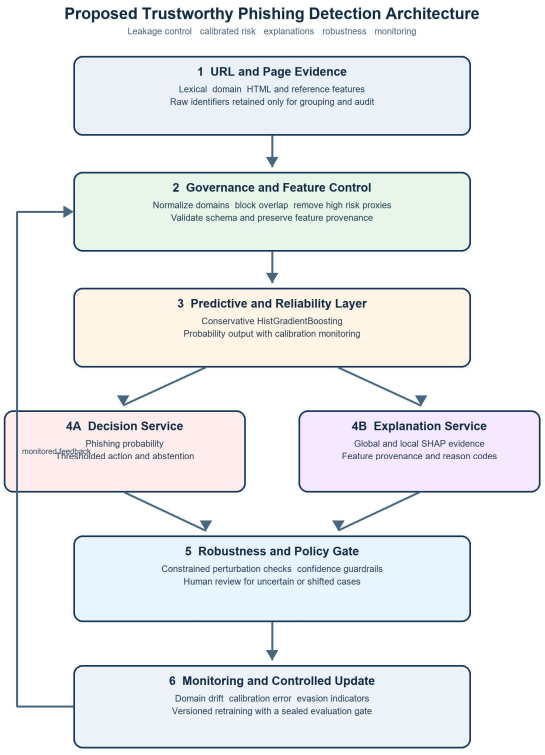


Fig. 1. Proposed trustworthy phishing detection architecture

TABLE III. ARCHITECTURE COMPONENTS AND TRUST CONTROLS

Component	Primary function	Trust control
Evidence acquisition	Extract URL and page indicators	Schema and provenance validation
Feature governance	Normalize and filter predictors	Domain grouping and proxy exclusion
Prediction	Estimate phishing probability	Frozen model and registered threshold
Explanation	Generate local and global evidence	Shared governed feature vector
Robustness gate	Check manipulation sensitivity	Budget tests and confidence guardrails
Monitoring	Observe drift and reliability	Versioned update and sealed reevaluation

The robustness gate evaluates whether influential mutable attributes exhibit combinations associated with the constrained attack study and routes uncertain shifted or policy sensitive cases to review; the monitoring layer records domain drift calibration error feature attribution movement and evasion indicators while every model update returns through a versioned development and sealed evaluation gate.

D. Calibration and Model Selection

The ensemble comparison ranked conservative Random Forest, Extra Trees and HistGradientBoosting by F1, recall, ROC AUC, Brier score, expected calibration error and training time with an unweighted sum of ranks; the selected ensemble was then retrained within every outer fold on a domain disjoint subtraining subset while sigmoid and isotonic calibrators used a separate 20 percent group split and the outer validation domains remained unseen.

This separation prevents the calibrator from observing validation outcomes while allowing a direct comparison with the same fitted base estimator; calibration was adopted only if it improved the joint reliability profile because a small threshold metric gain could otherwise mask worse probability quality.

E. Explainability and Stability

Permutation importance used five repeats on 5,000 validation records per fold and measured the decrease in phishing F1 while global Tree SHAP analysed 2,000 records from the final validation fold; stability was summarized by pairwise Spearman correlation and top ten overlap while method agreement used the overlap between the global permutation and SHAP top ten sets [28].

Local explanations represented the true positive, true negative, false positive and false negative cases closest to the decision threshold; exports contained only model features and probabilities while raw URLs, domains, titles, top level domains and filenames were excluded.

F. Constrained Adversarial Evaluation

The white box feature space threat model sampled 1,000 correctly detected phishing records from each grouped validation fold and greedily replaced distinct mutable feature values with legitimate class medians learned from that training fold; budgets of one, three and five changes were evaluated across fifteen deployment and webpage variables with the objective of minimizing predicted phishing probability [30].

Attack success meant a previously correct phishing prediction crossed below the fixed 0.5 threshold while explanation drift at budget five was measured through SHAP cosine similarity, top ten overlap and sign flip rate; the perturbations are a controlled representation level stress test and do not prove that every vector is a functional deployable page.

G. Frozen Test Evaluation and Statistics

The analysis protocol froze conservative uncalibrated HistGradientBoosting, conservative Logistic Regression, the 0.5 threshold and all metrics before test access; each model was fitted once on all 189,219 development rows and evaluated on 46,576 test rows while 1,000 normalized domain cluster bootstrap replicates produced 95 percent percentile intervals and common resamples produced paired differences.

IJAICET
International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies
ISSN: 3139-8413 | Vol. 1 No. 2 (2026) | Original Research Article | www.ijaicet.com | DOI : https://doi.org/10.68307/a3mqv686

IV. RESULTS

A. Development Model Comparison

Conservative Logistic Regression exceeded the conservative Decision Tree with mean F1 of 0.999184 compared with 0.996995 while HistGradientBoosting produced the strongest registered ensemble profile with mean F1 of 0.999896, recall of 0.999828, Brier score of 0.000076 and mean training time of 9.07 seconds; it ranked first on all six ensemble selection criteria and therefore advanced to calibration and explanation analysis.

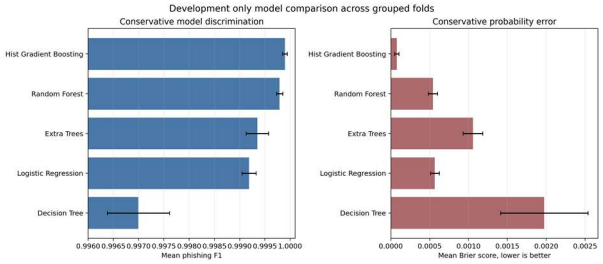


Fig. 2. Development model comparison under the conservative feature condition

TABLE IV. CONSERVATIVE DEVELOPMENT MODEL COMPARISON

Model	Mean F1	Mean recall	Mean Brier	Training seconds
Logistic Regression	0.999184	0.998969	0.000566	1.405
Decision Tree	0.996995	0.997411	0.001974	2.570
Random Forest	0.999791	0.999730	0.000542	35.418
Extra Trees	0.999349	0.998847	0.001056	30.225
HistGradientBoosting	0.999896	0.999828	0.000076	9.070

The full sensitivity condition approached perfect discrimination across models and HistGradientBoosting reached mean F1 of 0.999994; these values demonstrate the predictive contribution of the six derived fields but strengthen the proxy risk concern, so they were not used to replace the conservative primary path.

B. Probability Calibration

Uncalibrated HistGradientBoosting produced mean Brier score 0.000093, log loss 0.000523 and expected calibration error 0.000113 after the separate group split; sigmoid calibration slightly increased F1 but worsened all three probability measures while isotonic calibration produced the highest Brier score and log loss, so the uncalibrated model was retained.

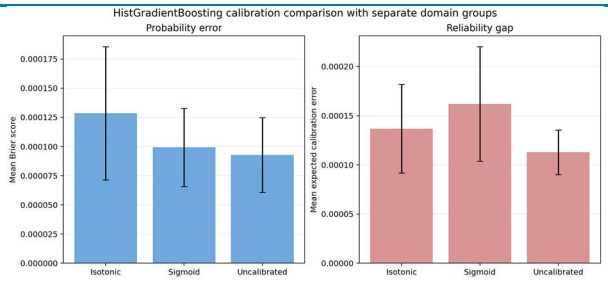


Fig. 3. Separate group probability calibration comparison

C. Global and Local Explanations

Permutation importance ranked IsHTTPS, LineOfCode, NoOfSubDomain and NoOfSelfRef as the four strongest features while mean absolute SHAP ranked LineOfCode, NoOfSelfRef, NoOfExternalRef and IsHTTPS first; nine of the top ten features overlapped across methods while cross fold permutation ranks achieved mean pairwise Spearman correlation 0.854 and mean top ten overlap 0.84.

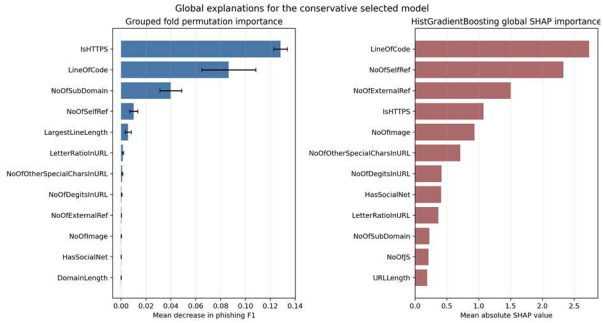


Fig. 4. Global permutation and SHAP explanations for conservative HistGradientBoosting

Local error explanations showed that the representative false positive was driven toward phishing by special character count and URL length despite opposing page evidence while the representative false negative combined strong legitimate directed contributions from external references and image count with opposing self reference and source length effects; these patterns motivated the mutable feature groups used in the adversarial experiment.

D. Constrained Adversarial Robustness

A single greedy feature change yielded mean attack success 0.0566 while three changes raised success to 0.7042 and reduced phishing recall to 0.2958; five changes raised success to 0.8462 and reduced recall to 0.1538 while mean phishing probability fell from 0.999948 to 0.161253.

IJAICET
International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies
ISSN: 3139-8413 | Vol. 1 No. 2 (2026) | Original Research Article | www.ijaicet.com | DOI : https://doi.org/10.68307/a3mqv686

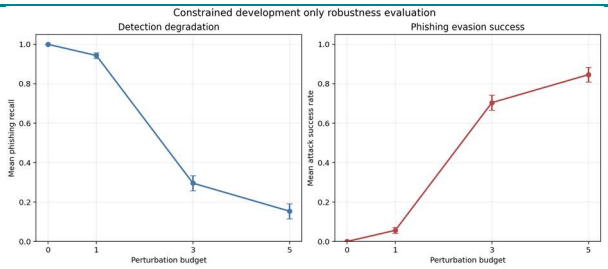


Fig. 5. Detection degradation and evasion success under constrained perturbations

NoOfSelfRef, LineOfCode, NoOfImage, NoOfExternalRef and IsHTTPS dominated greedy selections while budget five explanations retained mean top ten overlap 0.864 but reached mean cosine similarity minus 0.350 and sign flip rate 0.140; the stable feature set therefore concealed a substantial reversal in attribution direction during evasion.

E. Locked Test Performance

The selected model achieved locked test accuracy 0.999957, precision 1.000000, phishing recall 0.999897, specificity 1.000000, F1 0.999949, Brier score 0.000041 and log loss 0.000205; the registered threshold produced two false negatives and zero false positives across the 46,576 test records.

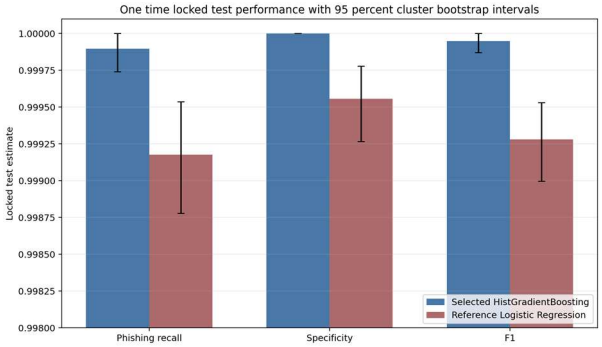


Fig. 6. One time locked test performance with normalized domain cluster intervals

TABLE V. LOCKED TEST PERFORMANCE AND CLUSTER BOOTSTRAP INTERVALS

Metric	Selected estimate	95 percent interval	Reference estimate
Phishing recall	0.999897	0.999740 to 1.000000	0.999177
Specificity	1.000000	1.000000 to 1.000000	0.999558
F1	0.999949	0.999870 to 1.000000	0.999280
Brier score	0.000041	0.000000 to 0.000105	0.000535
Log loss	0.000205	0.000009 to 0.000527	0.002190
Expected calibration error	0.000045	0.000009 to 0.000110	0.000520

Every oriented paired difference favoured HistGradientBoosting and each 95 percent interval excluded zero; the recall improvement was 0.000720 with interval 0.000362 to 0.001137 while the Brier score reduction was 0.000493 with interval 0.000353 to 0.000644.

V. DISCUSSION

A. Interpretation of Experimental Evidence

The clean evaluation supports the use of grouped conservative HistGradientBoosting for the registered dataset because it improves discrimination and probability error over linear and tree references without relying on raw identifiers or the six highest risk derived features; the result agrees with phishing research that finds value in nonlinear ensemble interactions [3,4,12,22,26,27] while extending prior comparisons through calibration, paired uncertainty and an immutable test protocol.

The calibration experiment shows why post hoc methods should not be adopted mechanically because sigmoid and isotonic transformations worsened the joint reliability profile despite the established value of calibration assessment [29]; when the base model already separates the development groups with very small probability error, a limited calibration subset can introduce variance without a compensating reliability gain.

The explanation results connect predictive dependence with attribution magnitude because permutation and SHAP shared nine leading variables but ranked them differently; this distinction is consistent with the different questions answered by performance permutation and additive attribution [28], while the cross fold stability estimates prevent a single sample explanation from being presented as universal.

The adversarial experiment changes the practical conclusion because near perfect clean performance did not protect the model against coordinated movement toward legitimate feature values; this outcome follows the broader security lesson that a classifier must be evaluated under an explicit attacker model [30] and complements phishing research concerned with compromised sites, temporal adaptation and robust ensembles [14,21,27].

The primary implication is that a phishing detector should not be selected from clean accuracy alone; model governance should preserve feature provenance, separate high risk derived predictors, monitor influential mutable attributes, report calibrated probability measures and test representation level evasions before operational deployment.

B. Architectural and Deployment Implications

The proposed architecture treats the conservative feature boundary as a runtime contract and therefore blocks the accidental reintroduction of raw identifiers or high risk derived proxies after model selection; this design also allows the explanation service to expose the same governed evidence used by the classifier while the robustness gate can apply stricter review when highly mutable webpage features dominate a decision.

IJAICET
International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies
ISSN: 3139-8413 | Vol. 1 No. 2 (2026) | Original Research Article | www.ijaicet.com | DOI : https://doi.org/10.68307/a3mqv686

Continuous monitoring should distinguish ordinary input drift from reliability drift and possible evasion because these conditions require different responses; domain and feature shifts motivate investigation while calibration degradation motivates threshold or probability review and repeated manipulation indicators motivate feature extractor hardening threat model revision or temporary escalation to human analysis.

A production implementation should retain model version feature schema normalization code threshold calibration state explanation method and monitoring configuration as one signed release unit; separating these artifacts can produce apparently reproducible scores while changing the effective decision system and therefore update approval should require the same grouped validation robustness assessment and locked evaluation logic used in this study.

C. Comparison with Existing Evaluation Practice

Many phishing studies emphasize discrimination through accuracy F1 or area under the curve while fewer combine entity separated evaluation probability reliability explanation stability and constrained attack evidence within one registered analysis; the present design does not claim that these dimensions exhaust trustworthiness but demonstrates that they can materially alter the conclusion drawn from the same high performing model.

The architecture also clarifies that explainability is not a decorative output added after prediction because explanation evidence supports review drift analysis and adversarial diagnosis; however explanations remain model dependent associations and must not be described as causal proof that a URL or webpage is malicious.

VI. LIMITATIONS

The study uses one public dataset and a hash based domain split from the same collection, so the locked test result does not establish temporal or cross dataset transfer; external validation and later acquisition periods remain necessary because phishing infrastructure and page design change [14,15,19].

The perturbation study operates in feature space and targets training fold legitimate medians, so it does not guarantee syntactically or functionally valid websites after every combined change; future work should construct problem space pages, enforce joint semantic constraints and measure feature extractor consistency before describing operational attack success.

SHAP and permutation importance can redistribute credit among correlated variables while expected calibration error depends on binning; the paper therefore reports agreement and stability measures but does not interpret any feature attribution as causal evidence.

VII. CONCLUSION

A leakage controlled and precommitted evaluation selected conservative uncalibrated HistGradientBoosting and confirmed strong locked test discrimination with low probability error; however constrained feature perturbations reduced mean phishing recall to 0.1538 at budget five and reversed explanation direction, so the central finding is that trustworthy phishing detection requires clean performance, reliability, explanation stability and adversarial evidence to be considered together rather than inferred from one benchmark score.

DATA AND CODE AVAILABILITY

The repository records dataset provenance, hashes, preprocessing rules, experiment configurations, aggregate results, figures and reproducible scripts while raw URLs, row level predictions and the dataset file remain excluded from version control; the PhiUSIIL data are available from the official UCI record under its stated licence [1].

REFERENCES

[1] Arvind Prasad and Shalini Chandra. PhiUSIIL A Diverse Security Profile Empowered Phishing URL Detection Framework Based on Similarity Index and Incremental Learning. Computers and Security. 2024. 10.1016/j.cose.2023.103545.

[2] Ozgur Koray Sahingoz and Ebubekir Buber and Onder Demir and Banu Diri. Machine Learning Based Phishing Detection from URLs. Expert Systems with Applications. 2019. 10.1016/j.eswa.2018.09.029.

[3] Ammara Zamir and Hikmat Ullah Khan and Tassawar Iqbal and Nazish Yousaf and Farah Aslam and Almas Anjum and Maryam Hamdani. Phishing Web Site Detection Using Diverse Machine Learning Algorithms. The Electronic Library. 2020. 10.1108/EL-05-2019-0118.

[4] Yukun Li and Zhenguo Yang and Xu Chen and Huaping Yuan and Wenyin Liu. A Stacking Model Using URL and HTML Features for Phishing Webpage Detection. Future Generation Computer Systems. 2019. 10.1016/j.future.2018.11.004.

[5] Brij B Gupta and Krishna Yadav and Imran Razzak and Konstantinos Psannis and Arcangelo Castiglione and Xiaojun Chang. A Novel Approach for Phishing URLs Detection Using Lexical Based Machine Learning in a Real Time Environment. Computer Communications. 2021. 10.1016/j.comcom.2021.04.023.

[6] Dong Jie Liu and Guang Gang Geng and Xiao Bo Jin and Wei Wang. An Efficient Multistage Phishing Website Detection Model Based on the CASE Feature Framework Aiming at the Real Web Environment. Computers and Security. 2021. 10.1016/j.cose.2021.102421.

[7] Wei Wei and Qiao Ke and Jakub Nowak and Marcin Korytkowski and Rafal Scherer and Marcin Wozniak. Accurate and Fast URL Phishing Detector A Convolutional Neural Network Approach. Computer Networks. 2020. 10.1016/j.comnet.2020.107275.

[8] Xi Xiao and Dianyan Zhang and Guangwu Hu and Yong Jiang and Shutao Xia. CNN MHSA A Convolutional Neural Network and Multi Head Self Attention Combined Approach for Detecting Phishing Websites. Neural Networks. 2020. 10.1016/j.neunet.2020.02.013.

[9] Farid Tajaddodianfar and Jack W Stokes and Arun Gururajan. Texception A Character Word Level Deep Learning Model for Phishing URL Detection. IEEE ICASSP. 2020. 10.1109/ICASSP40776.2020.9053670.

[10] Ayman El Aassal and Shahryar Baki and Avisha Das and Rakesh Verma. An In Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs. IEEE Access. 2020. 10.1109/ACCESS.2020.2969780.

[11] Seok Jun Bu and Hae Jung Kim. Optimized URL Feature Selection Based on Genetic Algorithm Embedded Deep Learning

IJAICET
International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies
ISSN: 3139-8413 | Vol. 1 No. 2 (2026) | Original Research Article | www.ijaicet.com | DOI : https://doi.org/10.68307/a3mqv686

for Phishing Website Detection. Electronics. 2022. 10.3390/electronics11071090.

[12] Anjaneya Awasthi and Noopur Goel. Phishing Website Prediction Using Base and Ensemble Classifier Techniques with Cross Validation. Cybersecurity. 2022. 10.1186/s42400-022-00126-9.

[13] Nguyet Quang Do and Ali Selamat and Ondrej Krejcar and Takeru Yokoi. Phishing Webpage Classification via Deep Learning Based Algorithms An Empirical Study. Applied Sciences. 2021. 10.3390/app11199210.

[14] Asif Ejaz and Adnan Noor Mian and Sanaullah Manzoor. Life Long Phishing Attack Detection Using Continual Learning. Scientific Reports. 2023. 10.1038/s41598-023-37552-9.

[15] Norah Alsuqayh and Areej Alhogail and Abdulrahman Mirza. Hybrid Feature Selection with SHAP Based Explainable AI for Interpretable Phishing Website Detection. Journal of King Saud University Computer and Information Sciences. 2026. 10.1007/s44443-026-01281-6.

[16] Hoc Minh Le and Van Thuong Nguyen. A Secure and Explainable Phishing Detection Framework Using Deep Learning and LLM Based Reasoning. Discover Informatics. 2026. 10.1007/s44564-026-00020-3.

[17] Asadullah Safi and Satwinder Singh. A Systematic Literature Review on Phishing Website Detection Techniques. Journal of King Saud University Computer and Information Sciences. 2023. 10.1016/j.jksuci.2023.01.004.

[18] A Suryan and C Kumar and M Mehta and R Juneja and A Sinha. Learning Model for Phishing Website Detection. EAI Endorsed Transactions on Scalable Information Systems. 2020. 10.4108/eai.13-7-2018.163804.

[19] Paulius Vaitkevicius and Virginijus Marcinkevicius. Comparison of Classification Algorithms for Detection of Phishing Websites. Informatica. 2020. 10.15388/20-INFOR404.

[20] Abdullah Al Alyan and Saad Al Ahmadi. Robust URL Phishing Detection Based on Deep Learning. KSII Transactions on Internet and Information Systems. 2020. 10.3837/tiis.2020.07.001.

[21] Igino Corona and Battista Biggio and Matteo Contini and Luca Piras and Roberto Corda and Mauro Mereu and Guido Mureddu and Davide Ariu and Fabio Roli. DeltaPhish Detecting Phishing Webpages in Compromised Websites. Computer Security ESORICS 2017. 2017. 10.1007/978-3-319-66402-6_22.

[22] Ali Aljofey and Qingshan Jiang and Abdur Rasool and Hui Chen and Wenyin Liu and Qiang Qu and Yang Wang. An Effective Detection Approach for Phishing Websites Using URL and HTML Features. Scientific Reports. 2022. 10.1038/s41598-022-10841-5.

[23] Erzhou Zhu and Yuyang Chen and Chengcheng Ye and Xuejun Li and Feng Liu. OFS NN An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. IEEE Access. 2019. 10.1109/ACCESS.2019.2920655.

[24] Mohamed Abdelkarim Remmide and Boumahdi Fatima and Narhimene Boustia and Chalabia Lilia Feknous and Romaissa Della. Detection of Phishing URLs Using Temporal Convolutional Network. Procedia Computer Science. 2022. 10.1016/j.procs.2022.10.209.

[25] Haifeng Jiang and Yu Chen and Yingjie Zhu and Xin Xu and Yuejia Song and Qiwei Chen. D PhishNet A Dual Branch Network for URL and HTML Feature Fusion in Phishing Webpage Detection. Computer Networks. 2025. 10.1016/j.comnet.2025.111648.

[26] Catherine Olivia Winanda and Andreas Liujaya Wiranata and Azani Cempaka Sari. Phishing Website Detection Using a Stacking Ensemble Classifier of Random Forest LightGBM and CatBoost. Procedia Computer Science. 2026. 10.1016/j.procs.2026.08.222.

[27] Abdellah Rezoug and Mohamed Bader el den. A Robust Ensemble Learning Approach to URL Based Phishing Webpage Detection. Big Data and Cognitive Computing. 2026. 10.3390/bdcc10050136.

[28] Scott M Lundberg and Su In Lee. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30. 2017. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[29] Chuan Guo and Geoff Pleiss and Yu Sun and Kilian Q Weinberger. On Calibration of Modern Neural Networks. Proceedings of Machine Learning Research 70. 2017. https://proceedings.mlr.press/v70/guo17a.html.

[30] Battista Biggio and Igino Corona and Davide Maiorca and Blaine Nelson and Nedim Srndic and Pavel Laskov and Giorgio Giacinto and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. ECML PKDD 2013. 2013. https://hdl.handle.net/11584/105260.

[31] M. K. Bagwani, A. Gangwar, K. Vishwakarma and V. K. Tiwari, “Real-time signature-based detection and prevention of DDOS attacks in cloud environments,” International Journal of Science and Research Archive, vol 12, no 2, pp 2929–2935, 2024, doi: 10.30574/ijsra.2024.12.2.1608. Source

[32] A. Khare, N. Choudhary and M. K. Bagwani, “Trustworthy and Adaptive Intrusion Detection for Imbalanced IoT Traffic,” International Journal of Artificial Intelligence, Cybersecurity and Emerging Technologies, vol 1, no 1, pp 80–86, 2026, doi: 10.68307/k9pxqe13. Source

[33] P. G. K. S. Mahesh Kumar Bagwani, "Performance comparison of REST API and GraphQL in a micro services architecture," International Conference on Data Science, Artificial Intelligence and Advanced Computing, 2024.