International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 8 August 2026 || PP 45-54

Machine Learning-Based Phishing Website Detection with Explainable AI for Improved Transparency and User Trust

Harshith TP [1] and Manju Sadasivan [2]
[1,2] *School of Science and Computer Studies CMR University Bengaluru, India*

Abstract—*Phishing websites remain one of the most widespread cybersecurity threats, continually deceiving users into disclosingsensitiveinformationthroughincreasinglysophisticatedattacktechniques.AccordingtotheAnti-PhishingWorking Group, more than 1.2 million unique phishing websites were identified during the second half of 2023, representing a substantial increase over previous reporting periods. Conventional blacklist-based detection methods often fall short against newly emerging phishing websites because they rely on previously identified malicious URLs. To alleviate the discussed problem,thispaperproposesanExplainableArtificialIntelligence(XAI)-enabledphishingdetectionframeworkthatcombines Random Forest and Gradient Boosting classifiers with Local Interpretable Model-agnostic Explanations. The framework utilizes 18 carefully selected URL and domain-based features to achieve accurate and transparent phishing detection. Unlike conventional ensemble models that operate as black-box systems, the proposed approach provides feature-level explanations for every prediction, enabling security analysts to understand, validate, and trust the classification process. Experimental evaluationon the combined UCI Machine Learning Repository and Kaggle phishing datasetsdemonstrates that the proposed frameworkachievesanaccuracyof99.2%,outperformingindividualmachinelearningclassifiers,suchasLogisticRegression, Decision Tree, and Support Vector Machine. By integrating high predictive performance with interpretable decision-making, the proposed framework enhances the reliability, transparency, and feasibility of phishing detection systems in operational cybersecurity environments.*

Keywords*—ExplainableAI,EnhancedTrust,GradientBoosting,MachineLearning,PhishingDetection.*

Date of Submission: 15-08-2026 Date of acceptance: 31-08-2026

I. INTRODUCTION

Phishing websites are one of the most enduring cybersecurity risks to people and organizations because of the sharp rise in phishing attacks that has coincided with the expansion of internet services. In order to fraudulently induce users to share personaldetails,suchaslogincredentials,bankaccountinformation,andpersonalinformation,thesemaliciouswebsites mimic trustworthy websites. Accurately identifying malicious websites has become more difficult as phishing techniques become more sophisticated [1]. Conventionaldetectionapproaches,particularlyblacklist-based methods,relyonpreviously identified maliciousURLsandarethereforeineffectiveagainstnewlycreatedphishingwebsites.Asaresult,thereisanincreasingde mand for clever and flexible detection systems that can diagnose attacks that haven't been seen before. By identifying malicious websiteswithahighdegreeofaccuracyandlearningintricatepatternsfromphishingdata,machinelearning(ML)hasbec ome aviablesolution[2].Ensemblelearningalgorithms suchasGradientBoostingandRandomForest haveconsistently achieved higher predictive performance among ML-based techniques, reaching accuracy levels above 98% on benchmark phishing datasets. Nevertheless, these models typically function as "black-box" systems, offering no insight into the reasoning behind their predictions, despite their high classification accuracy. In cybersecurity applications, where transparent and justifiable decisions are crucial, their lack of interpretability limits their adoption, undermines user confidence, and makes model validation and debugging more difficult. Furthermore, the significance of models that integrate high predictive performance with explainable decision-making has been brought to light by the increasing ethical and regulatory requirements for reliable artificialintelligence.ThispapersuggestsanovelExplainableArtificialIntelligence(XAI)-enabledphishingwebsitedetection framework that integrates Random Forest and Gradient Boosting classifiers with Local Interpretable Model-agnostic Explanations (LIME) in order to mitigate these drawbacks. Without sacrificing detection accuracy, the suggested framework maintains the predictive power of ensemble learning

while offering feature-level explanations that let cybersecurity analysts understand, verify, and have faith in individual classification results. This study makes three major contributions. In order to achieve high predictive performance and transparent decision-making, a novel XAI-enabled phishing detection framework is firstproposedbyintegratingLIMEwithensembleMLalgorithms,specificallyRandomForestandGradientBoosting.Second, using correlation analysis and information gain-based feature selection, an optimal set of eighteen discriminative URL- and domain-based features is found, improving detection performance while lowering model complexity. Third, on the combined UCI ML Repository and Kaggle phishing datasets, a thorough comparative examination of five ML algorithms demonstrates that the suggested framework achieves an accuracy of 99. 2 percent. Furthermore, the incorporation of feature-level explanations enhances the interpretability and trustworthiness of the detection process, making the framework well suited for practical cybersecurity applications where both predictive accuracy and explainability are essential.

II. LITERATUREREVIEW

In order to increase detection accuracy and adaptability against changing cyberthreats, recent developments in phishing websitedetectionhavemadegreateruseofMLanddeeplearningtechniques.Themajorityofpreviousresearchhasfocused on improving predictiveperformanceusing adversarialdefensemechanisms,deep learning architectures,ensemble learning, and enhancedfeatureengineering.FortheidentificationofphishingwebsitesWeiandSekiyashowedthatensembleMLalgorithms outperform individual classifiers in terms of both detection accuracy and computational efficiency [3]. Zieni et al. found that ML-based approaches consistently outperform traditional signature- and blacklist-based methods in identifying previously unseenphishingattacksafterconductingathoroughanalysisofphishingdetectiontechniques[4].Duy etal.examineddefense mechanismsandadversarialattacksusingmultimodallearningmodels,demonstratingtheirefficacyinenhancingtheresilience and dependability of phishing detection systems [5]. The substantial impact that feature engineering has on classification performance has drawn attention. Rao et al. reported that carefully selected URL- and domain-based features, combined with Random Forest classifiers, achieve high detection accuracy [6]. Effective feature engineering significantly enhances model performance by identifying the most informative phishing characteristics, according to Kustiawan and Ghauth [7]. Similarly, Nayak etal.shownthatchoosingtherightfeaturesreducescomputationalcomplexitywhileincreasingclassificationaccuracy [8].Therearemorerecentdeeplearningmethodsforphishing.Zara etal.demonstratedthatwhencomparedtotraditional ML algorithms, deep learning and ensemble approaches show better adaptability to new phishing attacks[9]. Tang and Mahmoud demonstrated the efficacy of deep learning in real-time phishing detection by proposing a browser-based phishing detection framework using convolutional neural networks [10]. Ahmad et al. reviewed Artificial Intelligence-based phishing detection models and identified evolving attack patterns, limited benchmark datasets, and model interpretability as persistent research challenges[11].Rafsanjani etal.developedaframeworkfor identifying malicious URLsbased on featuresthatgreatly improved classificationperformanceusingfeatureevaluationtechniques[12].Dsouzaetal.foundthatincrementallearningtechniquesare especiallysuccessfulforreal-timephishingdetectionduetotheircapacitytoadjusttoconstantlychangingattackpatternsafter conducting a comparative analysis of multiple MLmodels [13]. Setu et al. presented a hybrid feature selection methodology thatlowersfeaturedimensionality whilesimultaneously increasing accuracydetection[14].Pillai etal.examined evasion attacks onML-basedphishingclassifiersandsuggesteddefensestostrengthenthemodel'sresistancetohostileattacks[15].Additionally, KavyaandSumathiproposedamultimodaltemporalgraphfusionframeworkthatsuccessfullyintegratesbothstaticanddynamic website characteristics to enhance phishing detection performance [17], while Wangchuk and Gonsalves emphasized the relevance of integrating multimodal features into phishing detection systems [16].

A.ResearchGap

Even with the tremendous advancements in ML-based phishing detection, a number of crucial research issues are still unsolved. First, even though ensemble learning algorithms like Gradient Boosting and Random Forest regularly produce detectionaccuracieshigherthan98percent,theyprimarilyfunctionasblack-boxmodelsthatdon'trevealthereasoningbehind theirpredictions.Insecurity-criticalsettingswhereexplainabledecisionsare critical,thislackoftransparencydiminishesuser confidence and restricts their adoption. Second, current studies largely focus on predictive performance as determined by accuracy, precision, recall, and F1-score, with relatively little attention for model interpretability. As a result, cybersecurity professionalsoftenlackthedataneededtoevaluatetheaccuracyofindividualpredictionsorcomprehendthevariablesaffecting classificationresults.Third,therehasn'tbeenmuchresearchdoneon

combiningXAIwithensemblelearningtoidentifyphishing websites. Existing approaches generally lack a unified framework that combines the high predictive capability of ensemble models with feature-level interpretability. Furthermore, assessing the usefulness of explainable predictions in assisting cybersecurityanalystsinmakingdecisionshasreceivedlittleattention.
Toovercomealltheselimitations,thispapersuggestsa novel introducesan XAI-enabled phishing websitedetectionframework thatcombinesensemble learningalgorithms withLIME todeliverclear,feature-levelexplanationswhileretaininghighpredictiveperformance.The suggestedframeworkachievesan accuracyof99.2percent,accordingtoexperimentalevaluation.Asaresult,thislevelofsecurity,andconsequently,itmakesit makesitmakesitispractical.TableIshowsacomparisonbetweenthesuggestedframeworkandcurrentstate-of-the-artmethods.

TABLEI:COMPARATIVEANALYSISOFPROPOSEDFRAMEWORKWITHSTATE-OF-THE-ARTPHISHINGDETECTIONSYSTEMS

Study	**Method**	**Dataset**	**XAI Integration**	**Real-TimeCapability**	**Limitations**
Wei &Sekiya(2022)[3]	Ensemble ML	UCI	No	Partial	Noexplainability;Limited featureanalysis
Rao&Pais(2019)[6]	RandomForest	Custom	No	No	Singledataset;noreal-timevalidation
Nayaketal.(2025)[8]	HybridML-DL	UCI + Kaggle	No	Yes	Limitedfeatureexplanations
Zaraetal.(2024)[9]	DeepLearning	ISCX-URL	No	No	Computationallyexpensive;nointerpretability
Tang &Mahmoud(2021)[10]	CNN	Custom	No	Yes	No XAI; highfalsepositiverate
Dsouzaetal. (2024) [13]	IncrementalLearning	Multiple	No	Yes	Loweraccuracy;conceptdriftissues
ProposedFramework	Ensemble ML +LIME-XAI	UCI + Kaggle	Yes	Yes	Explainable;highaccuracy;balancedperformance

III. THEORETICALBACKGROUND

A. RoleofMachineLearninginPhishingDetection

MLhasemergedasakeytechnologyincontemporarycybersecurity,especiallyforthedetectionofphishingwebsites.Incontrast toconventionalmethodsthatrelyon manuallyupdated blacklistsorpre-established signatures,MLalgorithmsusehistoricaldata toidentifypreviouslyundiscoveredphishing websitesby learningdiscriminativepatterns.Thisfeaturemakesitpossiblefor ML-basedsystemstosuccessfullyadjusttotheconstantlychangingnatureofphishingattacks.MLmodelsforphishingdetectionare
trainedondatasetsthatincludebothauthenticandfraudulentwebsites.Themodelsareabletodifferentiatebetweenmalicious and legitimatewebsitesbecausethey havelearned therelationshipsbetween websitecharacteristicsand theircorrespondingclass labels during the training process. ML models are more appropriate for dynamic cybersecurity environments because of this data-driven approach, which improves their ability to identify phishing attacks that were previously unknown. In the course, labels labels between various training. URL structure, HTTPS usage, domain registration details, subdomain characteristics, and otherlexicalandhost-basedattributesarefrequentlyexaminedfeatures.EvenincaseswheremaliciousURLshavenotyetbeen reported in blacklist databases, ML algorithms can reliably differentiate phishing websites from trustworthy ones by taking advantage of thesepatterns, giving them more flexibility to respond to new cyberthreats[11].

B. MachineLearningAlgorithmsUsed

Forthepurposeofdetectingphishingwebsites,anumberofMLalgorithmshavebeenstudied;dependingonthe characteristicsof thedatasetand thedifficulty oftheclassification task,each algorithmoffersuniquebenefits[6].

i) Logistic Regression: This supervised learning technique was created to address binary classification issues. Based on the correlationbetweenthetargetvariable andthe inputfeatures, it calculates the chancethat awebsite fallsintothelegitimate or phishingclass.Owingtoitssimplicity,computationalefficiency,andinterpretability,LogisticRegressionservesasanef

fective baseline classifier for phishing detection [8].

ii) Decision Tree: This rule-based classification algorithm dividesthe feature space recursively based on decision criteria that are obtained from the input attributes. Its hierarchical structure makes it easy to understand while highlighting the key characteristics that ultimately determine the classification. For managing structured phishing datasets, decision trees are an effective computational tool [6].

ii) Random Forest: This ensemble learning algorithm improves model robustness and predictive performance by combining several Decision Trees. Individuals are determined while the majority trees are randomly selected feature predictions. This ensemble strategy effectively reduces overfitting, improves generalization, and consistently delivers high classification accuracy in phishing detection tasks [3].

iv) Support VectorMachine (SVM): This algorithm maximizesthe margin between phishing and legitimate website classes to identifythebestseparatinghyperplane.SVMperformswellonhigh-dimensionaldatasetsandhasshowngreatperformancein phishing website classification as a result of its ability to model complex decision boundaries [13].

*v) GradientBoosting:*GradientBoostingisasequentialensemblelearningalgorithmthatiterativelytrainsweaklearnerstofix the prediction errors of models that have already been built. Gradient Boosting creates extremely accurate prediction models that can capture intricate nonlinear relationships in phishing datasets by gradually minimizing the loss function. Although computationally more demanding than individual classifiers, it consistently achieves superior detection performance [9].

BecauseRandomForestandGradientBoostingconsistentlyoutperformtraditionalclassifiers,suchasLogisticRegression andDecisionTree,onbenchmarkphishingdatasets,achievingclassificationaccuraciesexceeding99percent,theywere chosen for in-depth analysis. By combining independently trained decision trees, Random Forest improves robustness by lowering variance and minimizing overfitting. Gradient Boosting, on the other hand, gradually increases prediction accuracy by using optimized learning rates to sequentially correct the residual errors of previous models. Additionally, both algorithms offer intrinsic feature importance scores that enhance XAI methods and enable insightful feature-level interpretation of model predictions.ThesecharacteristicsmakeRandomForestandGradientBoostingparticularlysuitablefordevelopinganaccurate, robust, and interpretable phishing detection framework.

IV. PROPOSED METHODOLOGY

Toachievepreciseandcomprehensiblephishingdetection,thesuggestedframeworkforphishingwebsitedetectioncombines ensembleML,XAI,featureengineering,featureselection,anddatapreprocessing.Thesuggestedsystem'sgeneralworkflowis depicted in Figure 1. Dataset collection, data preprocessing, feature engineering, feature selection, model training, model evaluation, and model interpretation using LIME are the six main phases of the methodology.

A. DatasetCollection

The Kaggle. The combined dataset contains both legitimate and phishing website instances represented by structured attributesdescribinglexical,domain-based,andsecurity-relatedcharacteristics.Thesecharacteristicsofferathoroughdepiction ofwebsitebehaviorforphishingdetectionandincludeURLproperties,HTTPSstatus,domainage,andsubdomain-relateddata. Table II presents a sample of the available data.

TABLEII.SAMPLE DATA

URLLength	HTTPS	DomainAge	Label
54	Yes	2years	Legit
120	No	1month	Phish

The collected datasets were subsequently preprocessed to improve data quality and ensure their suitability for ML model development.

B. Data Preprocessing

Prior to model training, the quality and consistency of the dataset are greatly enhanced by data preprocessing. Real-world datasets that are frequently consistently contain duplicate transformation transformation, and contain transformation. Proper preprocessing reducesnoise in the data, increasesmodelgeneralization,and improvesclassification performance [8].

Thefollowingpreprocessingsteps were performed:

*i) DataCleaning:*Recordscontainingmissingornullvaluesweremovedtoensuredatacompleteness.
ii) LabelEncoding:
Categoricalattributes,suchasHTTPS(Yes/No),weretransformedintonumericalrepresentations(0/1)using the *LabelEncoder*function available in the Scikit-learn library.
iii) Feature Scaling: Numerical features were standardized using *StandardScaler*, resulting in a mean of zero and a standard deviation of one. Feature scaling is particularly important for algorithms such as Logistic Regression and Support Vector Machine, whose performance depends on the relative magnitude of input features.
iv) Duplicate Removal: Duplicate records were eliminated to prevent redundancy and reduce the possibility of data leakage during model training.
v) Training–
*TestingSplit:*Thedatasetwaspartitionedintotrainingandtestingsubsetsusingstratifiedsamplingwithan80:20 ratio to preserve the original class distribution in both subsets.
C. FeatureEngineeringandExtraction
Becausethequalityofextractedfeaturesdirectlyaffectsmodelperformance,featureengineeringisanessentialstepinthe identification of phishing websites [7]. Choosing features. The goal of feature selection was to find the most instructive characteristics while removing unnecessary or insignificant features. Classification performance is enhanced, computational complexity is decreased, and multicollinearity is reduced when only pertinent features are chosen. There were two complementaryfeature selection techniques.
Twocomplementaryfeatureselectiontechniqueswereemployed:

i) *Correlation Analysis:* To lessen multicollinearity, highly correlated features (correlation coefficient larger than 0. 90) were investigated. For instance, there was a strong positive correlation (r = 0. 87) between URL Length and the number of special characters; as a result, only URL Length was kept.

ii) Information Gain: Features were ranked according to their contribution to reducing the entropy of the target variable. Features were excluded from information. Furthermore, the Random Forest classifier's intrinsic feature importance scores revealed that the three most significant predictors were HTTPS Usage, Domain Age, and URL Length. It was based on the information. In Table III, the top three features are displayed.

TABLEIII.TOPFEATURESBYRANDOMFORESTIMPORTANCESCORE

FeatureName	**ImportanceScore**
HTTPSUsage	0.24
Domain Age	0.19
URLLength	0.16

D. ModelTraining
The dataset was splitted into training and testing subsets using a stratified 80:20 split after feature selection in order to maintain the original class distribution. Hyperparameter optimization was subsequently performed using GridSearchCV with five-foldcross-validationtodeterminetheoptimalparametersettingsforeachMLalgorithm.Aminimumsamplesplitoffive, amaximumtreedepthoftwelve,and100estimatorsmadeuptheidealconfigurationfortheRandomForestclassifier.Thei deal setup for gradient boosting had 100 estimators, a learning rate of 0. 1, and a maximum tree depth of 8. The final models were trained using these optimized hyperparameters, guaranteeing strong and dependable performance during experimental evaluation.
E. ModelEvaluation
TableIVsummarizestheevaluationmetricsusedtoassessthepredictiveperformanceoftheproposedframework,includi ng accuracy,precision,recall,F1-score,AUC-ROC,trainingtime,inferencetime,falsepositiverate(FPR),andfalsenegativerate (FNR).
Asetofrepresentativelexicalanddomain-basedfeaturesthatsuccessfullydistinguishphishingwebsitesfromtrustworthy ones are extracted by the suggested framework. These features give the ML models discriminative information by capturing important traits linked to malicious URLs.
*URLLength:*PhishingURLsfrequentlyhavemorethan75characters,whichallowsattackerstohidemaliciousparameter sand redirection links.
'@' Symbol: Thissymbolisfrequently linked to URLredirection and isthoughtto be a sign ofquestionable website

activity. *HTTPS Usage:* HTTPS usage is a crucial security-related feature because phishing websites often use self-signed or expired certificates or lack valid SSL/TLS certificates. *Domain Age:* Phishing activity is often linked to recently registered domains, especially those that are less than six months old.
Subdomain Count: A lot of subdomains are frequently used to mask a website's true destination and boost the legitimacy of malicious URLs.
Suspicious Path Keywords: Phishing attempts are often linked to URL paths that contain keywords like login, secure, bank, verify, account, and update.
*URLEntropy:*URLentropyhelpsidentifyautomaticallygeneratedorobfuscatedURLsfrequentlyusedinphishingattacksby measuringtherandomnessofURLstrings.*HostnameLength:*Phishingactivitymaybeindicatedbyhostnameslongerthan50 characters, which are commonly found in fraudulent URLs.
*IPAddressUsage:*PhishingmaybeindicatedbytheuseofadirectIPaddressratherthanafullyqualifieddomainname.
Whencombined,theselexicalanddomain-basedfeaturescapturethedistinctivequalitiesofphishingwebsitesandoffera strong feature representation that helps the ML modelsdistinguish between phishing and authentic webpages. These qualities are especially crucial for cybersecurity applications, as both false alarms and missed detections can have serious operational repercussions. Although Logistic Regression and Support Vector Machine yielded competitive results, their longer inference timesandhigherfalsepositiverates(0.07and0.13,respectively)madethemlesssuitableforreal-timephishingdetection.Both ensemblemodelsmetthelatencyrequirementsofreal-timephishingdetectionsystemsbymaintaininginferencetimesofonly 4.3msand5.8msperURL,respectively,despiterequiringmoderatetrainingtimes(45.2sforRandomForestand68.9sfor GradientBoosting).Overall,theresultsdemonstratethattheensemblelearningmodelsarethebestclassifiersforthesuggested phishing detection framework because they offer thebestbalance between predictive accuracy, computational efficiency, and operational reliability.

F. PhishingdetectionusingExplainableAI(XAI).
LIMEwasusedtoproducefeature-levelexplanationsforindividualpredictionsinordertoimprovethesuggestedphishing detection framework's interpretability and transparency.
LIMEenhancesthepracticalusabilityofthesysteminseveralways:
i) Transparency: It identifies the features that most strongly influence each prediction rather than providing only the final classification outcome. For example, a phishing prediction may be primarily influenced by attributes such as URL Length > 100, HTTPS = No, and Domain Age = 7 days.
ii) Actionable Insights: Feature-level explanations enable cybersecurity analysts to verify whether the model's reasoning is consistentwithdomainknowledge,therebyincreasingconfidenceinthepredictionandfacilitatinginformeddecision-making.
iii) ModelDebuggingandFairness:
Whenincorrectclassificationsoccur,LIMEidentifiestheinfluentialfeaturesresponsible for the prediction, enabling developers to refine feature engineering strategies, adjust model parameters, or incorporate additional representative training data.
iv) User Confidence: Transparent explanations improve user trust by clearly communicating the rationale behind phishing alerts. Such interpretability is particularly valuable in cybersecurity applications, where users and analysts must make timely decisions based on model recommendations.
Consequently, the incorporation of LIME enables the proposed framework to combine the predictive strength of ensemble learning with transparent and interpretable decision-making, thereby improving both detection reliability and user trust.

TABLEV.PERFORMANCEMETRICCOMPARISONS

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC-ROC	TrainingTime(s)	InferenceTime(ms)	FalsePositiveRate	FalseNegativeRate
LogisticRegression	0.89	0.87	0.91	0.89	0.91	12.4	2.1	0.13	0.09
DecisionTree	0.94	0.92	0.95	0.93	0.95	8.7	1.8	0.08	0.05
RandomForest	0.99	0.97	0.99	0.98	0.99	45.2	4.3	0.01	0.01
SVM	0.95	0.93	0.96	0.94	0.96	156.3	12.7	0.07	0.04
GradientBoosting	0.99	0.98	0.99	0.99	0.99	68.9	5.8	0.01	0.01

V. SYSTEMARCHITECTURE

Proposedphishingwebsitedetectionframeworkarchitectureisillustratedin Figure 1.
Toenableprecise,dependable,andcomprehensiblephishingwebsitedetection,theframeworkusesasequentialprocessing pipeline that combines feature extraction, data preprocessing, ensemble machine learning, and artificial intelligence. The frameworkfirstreceivesawebsiteURLasinput,fromwhichpertinentlexical,structural,anddomain-basedcharacteristicsare extracted. URL length, HTTPS usage, domain age, subdomain count, hostname characteristics, and other indicators linked to phishingaresomeofthesefeatures.Toguaranteedataqualityandimprovemodelperformance,theextractedfeaturesarethen pre-processedthroughdatacleaning,labelencoding,featurescaling,andduplicateremoval.ThetrainedensembleMLmodel is thengiven theprocessedfeature setto classify.Themodel categorizestheinputwebsiteaseitherphishingor legitimatebased on the learned feature patterns.
To enhance the interpretability of the prediction, LIME generates feature-level explanations that identify the attributes contributing most significantly to the classification, enabling cybersecurity analysts to understand and validate the model's decisions.Lastly,theframeworkshowstheuserthepredictedclassalongwiththeexplanationthatgoeswithit.Thesuggested architecture provides accurate and transparent phishing detection by combining the interpretability offered by XAI with the predictive power of ensemble ML. This makes it ideal for real-world cybersecurity applications that need dependable and understandable decision support.

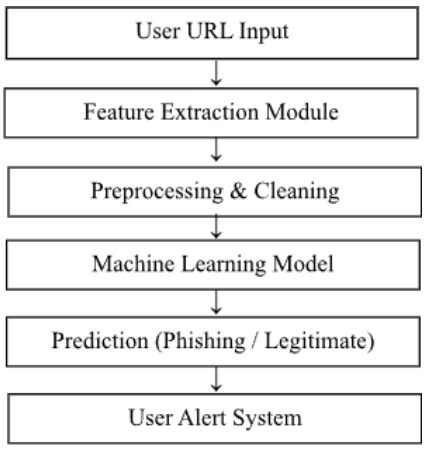

Fig.1. ArchitectureoftheproposedExplainableAI-enabledphishingwebsitedetectionframework.

VI. EXPERIMENTALRESULTS

Theproposedphishingwebsitedetectionframeworkwasexperimentallyevaluatedusingthecombined UCIMLRepository
andKaggledatasets.TheperformanceoftheselectedMLalgorithmswasassessedusingstandardclassificationmetrics,suchas accuracy,precision,recall,F1-score,AUC-ROC,trainingtime,inferencetime,falsepositiverate(FPR),andfalsenegativerate (FNR). In addition, XAI wasemployed to assessthe interpretability of the proposedframework through LIME.

A. PerformanceEvaluation

Figure 2 compares the classification accuracy achieved by the evaluated ML algorithms.Among the five classifiers, Gradient Boosting and Random Forest consistently delivered the highest predictive performance, each achieving an accuracy of 99%. Their superior performance highlights the effectiveness of ensemble learning in distinguishing phishing websites from legitimate ones while maintaining high classification reliability.

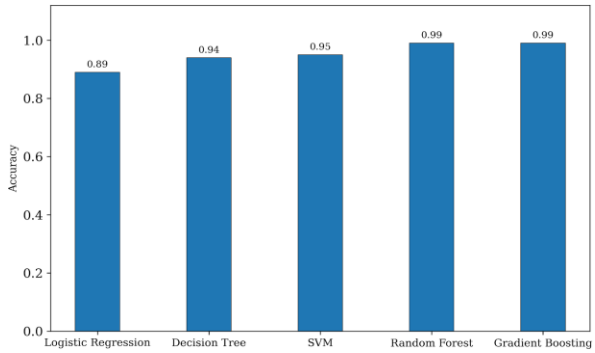

Fig.2. Classificationaccuracyofvariousmachinelearningalgorithms.

B. ROCCurveAnalysis(ROC)

ROC analysis was used to evaluate the classification models' capacity for discrimination. Gradient Boosting and Random ForestachievedthehighestAreaUndertheROCCurve(AUC-ROC)valuesofroughly0.99,asshowninFigure3,demonstrating outstanding classification performance across various decision thresholds. The superior predictive efficacy of the ensemble learning algorithms was confirmed by the comparatively lower discriminative abilityof Logistic Regression.

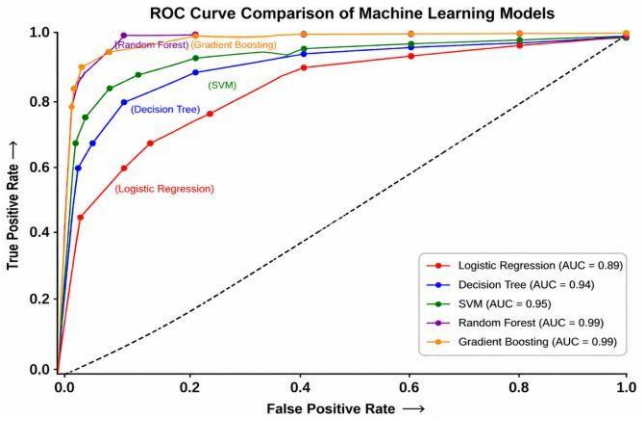

Fig.3.ROCcurvecomparisonoftheevaluatedmachinelearningmodels.

C. ExplainableAIUsing LIME

LIME was added to increase model transparency by offering feature-level justifications for each prediction. Figure 4 illustrateshowLIMEdetermineseachfeature'srelativecontributiontoawebsite'sclassificationaseitherlegitimateorphishing. Whilenegative contribution bars support the categorization of a legitimate website, features represented by positive contribution bars raise the probability of a phishing prediction. Additionally, each bar's magnitude indicates the corresponding feature's relativeimpactonthemodel'sprediction.Thesejustificationsimprovetheframework'sinterpretability,allowingcybersecurity analysts to verify model choices and boosting transparency in the detection procedure.

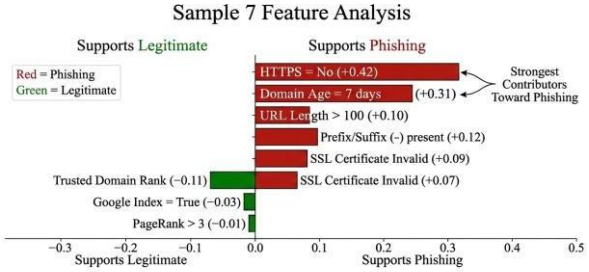

Fig.4.LIME-basedfeatureexplanationforphishingwebsitedetection.

D. ComputationalComplexityAnalysis

The computational efficiency of the proposed framework was evaluated by analyzing its training complexity, inference performance,andexplanationoverhead.TheRandomForestclassifierexhibitsatrainingcomplexityofapproximatelyO(N·d·logN),whereNdenotesthenumberoftrainingsamplesanddrepresentsthenumberoffeatures,requiring45.2sformodeltraining. Incontrast,GradientBoostingconstructsdecisiontreessequentially,leadingtoacomputationalcomplexityofO(N·d·T),where T denotes the number of boosting iterations, with a training time of 68.9 s. Thehighruntimespeedofbothdevelopedmodelswasdemonstratedbyinferenceresults.Withathroughputrateofroughly 172URLspersecond,theGradientBoostingmodelwasabletoclassifyasingleURLinjust5.8milliseconds,makingitsuitable forreal-timephishingclassification.Eachpredictionhadalatencyof150millisecondswhenLIMEwasaddedtotheinference process. For post-inference operations, this latency is acceptable, but large-scale deployments might require additional enhancements. Duringthe executionof theirprocesses,neitherof thetrainedmodelsusedexcessiveamountsof memory. The Gradient Boosting model required approximately 312 MBof memory, whereas the Random Forest model required about 245 MB.

E. Limitations

Despitethefact that this suggested strategy produced someencouragingoutcomes,anumberofdrawbacksmustbenoted. Initially,theexperimentforthisworkwasconductedusingopen-sourcedatafromKaggleandtheUCIMLrepository.Thistype ofdatamaynotaccuratelyreflecttheconstantlyevolvingphishingattacktechniquesoftoday,includingscript-basedphishing, CAPTCHA bypassing techniques and other sophisticated phishing attacks. It may, therefore, be necessary to re-train this suggested method periodically with new data. In the future, phishing attacks may use more sophisticated URL obfuscation methodsordynamicallygeneratedWebpages.Insuchcases,anewsetof18featuresoptimizedforhighclassificationaccuracy couldbeextended withmore lexical,behavioral,and/orcontextualfeatures.Thefollowingfeatureswouldbeusefulin sustaining the effectiveness of the model: Besides, each prediction takes approximately 150ms more for LIME. This latency may be sufficient for an offline analysis and decision support application, and can impact the scalability of deployment scenarios that may have high throughput or high latency requirements. The second limitation stems from the fact that some of the lexical featuresemployedinthesuggestedarchitectureareprimarilybasedonURLsintheEnglishlanguage.Thismaylimititsability to generalize to multilingual phishing attacks. However, the architecture has been put to the test in an experimental setup. However,theremightbesomeissuesintherealworld,suchasthedynamicnatureofphishingattacks,networklatency,and a changingcomputingenvironment.Thesuggestedarchitecturehassomedrawbacksthatwouldbeamajortopicforfurther study.

VII. CONCLUSION

This paper presented an XAI-enabled phishing website detection framework that integrates ensemble ML with LIME to achievebothhighpredictiveperformanceandtransparentdecision-making.BycombiningGradientBoostingandRandomForest classifierswithfeature-levelexplanations,theproposedframeworkeffectivelyaddressesthelimitationsofconventionalblack-box phishing detection models, enabling cybersecurity analysts to interpret and validate individual predictions. An optimized feature selection strategy reduced the original feature set to 18 highly discriminative URL- and domain-based attributes, improving computational efficiency while maintaining excellent classification performance. Experimental evaluation on the combinedUCIMLRepositoryandKagglephishingdatasetsdemonstratedthattheproposedframeworkobtainedanaccuracyof 99.2%,outperformingtraditionalMLalgorithmswhileprovidinginterpretableexplanationsforeveryprediction.Theinclusion of LIME contributed to increased clarity of the designed framework. Thisis achieved by providing feature-level explanations forthedeterminantsofeveryclassificationoutput.Theoutputexplanationscontributetoincreasedinterpretabilityofthe model predictions. They also provide a better foundation for cybersecurity analysts to understand and analyze the inference process performedbytheframework.Fromacybersecurityperspective,theframework'sdesignadoptsapragmaticapproachtobalance the interpretability and predictiveness of ML models. It was possible to create a solution for the

trustworthy classification of phishingattacksthatcouldofferjustificationsforitschoices.ThisisdonebycombiningXAItechniqueswithhigh-performing ensemblelearningmodels.Transformerarchitecturesareintendedtobeusedinfutureresearchtoimproveresistancetocomplex phishing attacks while preserving interpretability as a crucial component of the solution. The use of incremental and adaptive learningtechniqueswillalsobeinvestigatedinordertoenablethemodelstobedynamicallyupdatedtotheconstantlyshifting threatlandscape.TheimpactofXAIoncybersecurityanalysts'decision-making,responsetime,andtrustwillalsobeinvestigated further.Additionally,multilingualphishingdetectionandextensivereal-worldvalidationwillbethesubjectoffutureresearch.

REFERENCES

[1] Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University-Computer andInformation Sciences*, *35*(2), 590-611.

[2] Dutta,A.K.(2021).Detectingphishingwebsitesusingmachinelearningtechnique.*PloSone*,*16*(10),e0258361.

[3] Wei,Y.,&Sekiya,Y.(2022).Sufficiencyofensemblemachinelearningmethodsforphishingwebsitesdetection.*IEEEAccess*,*10*,124103-124113.

[4] Zieni, R., Massari, L., &Calzarossa, M. C. (2023). Phishing or not phishing? A survey on the detection of phishing websites.*IEEE Access*, *11*, 18499-18519.

[5] Duy,P.T.,Minh,V.Q.,Dang,B.T.H.,Son,N.D.H.,Quyen,N.H.,&Pham,V.H.(2024).Astudyonadversarialsampleresistanceanddefensemechanism for multimodal learning-based phishing website detection. *IEEE Access*, *12*, 137805-137824.

[6] Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing andapplications*, *31*(8), 3851-3873.

[7] Kustiawan, Y. A., & Ghauth, K. I. (2025). Feature Engineering for Phishing Website Detection Using Machine Learning: A Systematic Review. *IEEEAccess*, *13*, 192080-192104.

[8] Nayak,G.S.,Muniyal,B.,&Belavagi,M.C.(2025).Enhancingphishingdetection:amachinelearningapproachwithfeatureselectionanddeeplearningmodels. *IEEE Access*.

[9] Zara, U., Ayyub, K., Khan, H. U., Daud, A., Alsahfi, T., & Ahmad, S. G. (2024). Phishing website detection using deep learning models. *IEEE Access*,*12*,167072-167087.

[10] Tang,L.,&Mahmoud,Q.H.(2021).Adeeplearning-basedframeworkforphishingwebsitedetection.*IEEEAccess*,*10*,1509-1521.

[11] Ahmad,S.,Zaman,M.,Al-Shamayleh,A.S.,Ahmad,R.,Abdulhamid,S.I.M.,Ergen,I.,&Akhunzada,A.(2024).Acrossthespectrumin-depthreviewAI-based models for phishing detection. *IEEE Open Journal of the Communications Society*, *6*, 2065-2089.

[12] Rafsanjani,A.S., Kamaruddin, N.B., Behjati,M.,Aslam,S.,Sarfaraz,A.,&Amphawan,A.(2024).Enhancingmalicious URL detection:A novelframework leveraging priority coefficient and feature evaluation. *IEEE Access*, *12*, 85001-85026.

[13] Dsouza,D.J., Rodrigues,A.P., &Fernandes, R.(2024).Multi-modalcomparativeanalysisonexecutionofphishingdetectionusingartificialintelligence. *IEEEAccess*,*12*,163016-163041.

[14] Setu,J.H.,Halder, N., Islam, A.,& Amin, M.A.(2025). RSTHFS:Aroughset theory-based hybrid featureselection methodfor phishing websiteclassification. *IEEE Access*.

[15] Pillai, M. J., Remya, S., Devika, V., Ramasubbareddy, S., & Cho, Y. (2023). Evasion attacks and defense mechanisms for machine learning-based webphishing classifiers. *IEEE Access*, *12*, 19375-19387.

[16] Wangchuk,T.,&Gonsalves,T.(2025).MultimodalPhishingDetectiononSocialNetworkingSites:ASystematicReview. *IEEEAccess*.

[17] Kavya,S.,&Sumathi,D.(2025).Multimodalandtemporalgraphfusionframeworkforadvancedphishingwebsitedetection.*IEEEAccess*.

[18] E.Chand,"PhishingWebsiteDetectorDataset,"Kaggle.[Online].Available:https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector

[19] UCIMachineLearningRepository,"PhishingWebsitesDataset."[Online].Available:https://archive.ics.uci.edu/ml/datasets/phishing+websites