iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81565

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

AI-Powered Phishing URL Detection System Using Machine Learning

N.Jaya Santhi, Manubrolu Ravali, Marri Prathyusha, Shaik Asma
Department of Computer Science and Engineering Bapatla Women's Engineering College

Abstract: ***Phishing attacks are considered to be one of the major cybersecurityconcerns, whereusersaretrickedint orevealingcritical information using phishing websites. Blacklisting techniques are not effective in detecting new phishing URLs, especially in zero-day attacks. This paper proposes a novel phishing URL detection system using artificial intelligence techniques, including machine learning. ThesystemextractS30featuresfromURLs,includinglexicalfeatures, using machine learning techniques. The system uses a Gradient Boosting Classifier to classify legitimate URLs and phishing URLs. The systemisimplementedusingawebapplicationusingtheFlasklibrary, which can be used by users to check the safety of URLs in realtime. Experimentalresultsshowtheeffectivenessofthesystemintermsof accuracy and prediction time.***

Index Terms—Phishing Detection, Machine Learning, Cybersecurity, Feature Extraction, Gradient Boosting Classifier, Web Application Security.

I. INTRODUCTION

Therapidgrowthoftheinternethasrevolutionizedtheway people communicate, shop, bank, and share their health and personal information. With the rise of internet users running intothetensofbillions,digitallifehasbecomeatreasuretrove ofsensitiveinformation,rangingfromfinancialinformationto medical records and social media. However, this has also become a fertile ground for cybercriminals, and web-based cybercrimes have become more frequent, sophisticated, and devastatingthanever.Amongthese,phishinghasbecomeone ofthemost commonandsuccessfultypesofsocialengineering attacks.

Phishing is a practice in which cyber attackers deceive people into revealing sensitive information, like login credentials,creditcard details,orSocialSecurity numbers,by impersonatingatrustworthysourceindigitalcommunication. Inmostinstances, attackersluretheir victimswithconvincing emails,textmessages,orsocialmedialinksthatdirectthemto fake websites that are made to resemble the real website.

Once the victim enters their sensitive information on these fake websites, the attackers get the information in real time, leading to serious consequences, like identity theft, financial loss,anddamagetothereputationoftheorganizationswhose customers are victimized by these attacks. In fact, the overall cost of phishing is huge, with billions of dollars lost each year by businesses and individuals, and therefore, it is one of the most important issues in the context of web-based cybercrimes.

Phishing is a menace that has survived the test of time becauseitisbasedonthepsychologyofpeople,andtheattack isbasedonthepsychology ofpeople,andtheattackmethod, the URL, is a technical aspect that can be studied and prevented. Traditionally, the most common method of preventingphishingisbasedonatechniquecalledblacklisting. In blacklisting, a huge database of URLs is maintained, and whenever a user wants to access a website, the URL is comparedwiththelistofURLsinthedatabase.IftheURLisin the database, access is denied. Blacklisting is a very effective method and does not cause much delay in the process, but it has a major drawback. For example, phishing gangs now use machines to generate thousands of short-lived, recently registered domains on a daily basis. By the time a suspicious URL is detected, recorded, investigated, and included in a blacklist, the campaign has probably done its work and the domain has become inactive.

Toovercomethedelaysandlimitationsofblacklists,people have resorted to heuristic detection. Heuristic detection involves looking at the composition of a webpage, its HTML, someofthewordsitcontains,andthehostitison,andsoon. Heuristicdetectionisapreventivemeasure,butitisalsoaslow and computationally intensive process, difficult to do in real timeasyousurf,andalsogeneratesalotoffalsealarms.False alarms cause users to dismiss the system altogether.

That is why there is an immediate need to have a smarter, dynamic,andproactivewayofjudgingURLsinstantlywithout any need to check their reputation beforehand.This is where MachineLearning(ML)hassteppedinasasolution.Unlike traditionalblacklistsorrules-basedsystems,MLmodelslearn from the complex patterns within the huge amount of data available about the URLs themselves.

©IJRASET: All Rights are Reserved

By learning the numericalandlinguistic characteristicsofthe URLs,suchasthe total length oftheURL, thepresenceofunusual characters in theURL(e.g.,@or-),thepresenceofanIPaddressratherthan a domain name, or unusual subdomain structures, the ML model is able to make judgments about whether the URL is legit or malicious.

This paper presents the design, development, and deployment of a robust AI-based Phishing URL Detection System that overcomes the limitations of traditional security tools. This system is based on a wide range of 30 unique structural and lexical features of the URLs and is based on a Gradient Boosting Classifier due to its high performance with tabular data, low overfitting, and ability to show the importance of each feature. This is then packaged into a lightweight Flask web app that offers real-time URL analysis throughasimpleand accessibleuserinterface. Thisproactive system is a robust defensive measure that blocks phishing threatsattheexacttimewhentheuserisinteractingwiththe URL.

A. Contributions of this Paper

Themaincontributionsofthispaperincludethefollowing:

- The paper develops a phishing detection system using machinelearningtechniqueswithlexicalfeaturesofURLs.
- Thepaperidentifies30features,coveringbothstructural and lexical characteristics, for accurate classification of phishing emails.
- A Gradient Boosting Classifier is used to improve the detection of phishing emails.
- The classifier is then used to create a web application usingtheFlaskwebdevelopmenttool,providingasimple interface for instant phishing detection.

II. LITERATURE REVIEW

In order to mitigate the increasing problem of phishing, a number of detection techniques have been proposed.

1) Kumar et al. investigated thefeasibility ofusing machine learning techniques in the detection of phishing sites usingSVMandRandomForesttechniques.Goodaccuracy wasobtained,butthefeatureextractionphaserequireda
2) lotofcomputations,slowingdownthepredictionphase.
3) Sharma et al. presented a hybrid model using a combination of blacklisting and heuristic detection techniques. However, the major drawback was the large number of false positives during the detection of benign and newly registered domains, which had no reputation or history.
4) SmithfocusedonthelexicaldetectionofmaliciousURLs. Goodresultswereobtained,butthetechniquewas limited to detecting only those phishing sites that could beidentifiedusinglexicalfeatures,failingtodetectthose phishing sites that are similar in name to the actual domain.

The proposed system is based on the idea of selecting the optimal set of 30 features, which are computationally inexpensive and using a Gradient Boosting technique.

III. PROBLEM STATEMENT

Internet users have become more exposed to phishing attacksduetothefactthatthephishingattacksevolvequicker than the traditional detectors. Some of the major challenges that exist in the detection of phishing attacks include the following:

- Blacklistingdoesnotcoverzero-dayphishingattacks.
- Too many false positivesmakethesystem inconvenient.
- Thoroughanalysisusingheuristicstakestoolong.

Themaingoaloftheproposedsystemistobuildanefficient web-based system that can intelligently extract the most importantfeaturesfromanunknownURLonthefly,alongwith theresultsoftheestimationofthelikelihoodoftheURLbeing a phishing site.

IV. PROPOSED SYSTEM

The proposed Phishing URL Detection system combines a smart backend with a user-friendly web frontend.

A. System Components

- FeatureExtractionModule:GivenarawURL,thismodule, through mathematical calculations, extracts 30 features (e.g., presence of an IP address, total URL length, presenceofURLshorteningservices,presenceofthe”@” character, and the presence of the HTTPS protocol).

- Machine Learning Module: A pre-trained Gradient BoostingClassifiermodel,storedinmodel.pkl,willaccept the30featuresandoutputapredictionandaprobability score.
- Web Interface (User Module): A simple and clean user interfacewhereuserswillinputsuspiciousURLsandview the safety rating in real time.

B. Main Features

- Real-time URL Parsing: Quick parsing of lexical and host-based features.
- Probability Scoring: A percentage score indicating the probability of the website's safety.
- VisualIndicators:Instantvisualfeedback,displayedonthe webpage, indicating the website's safety status in real time.

C. Technologies Used

- Frontend:HTML,CSS
- BackendFramework:PythonwiththeFlaskframework
- MachineLearningLibrary:Scikit-Learn,Pandas,NumPy
- WebScraping/Parsing:BeautifulSoup,python-whois

V. SYSTEM ARCHITECTURE

The phishing URL detector is a simple end-to-end pipeline thatlinksaweb frontendandamachinelearningclassifier.It is a system of connected elements that accept user input, extract meaningful features, and output a prediction.

A. User Input:

AuserentersaURLusingthewebinterface'sinputform.The web interface is constructed using the HTML and CSS languages, enabling any user to input a suspicious URL to check.

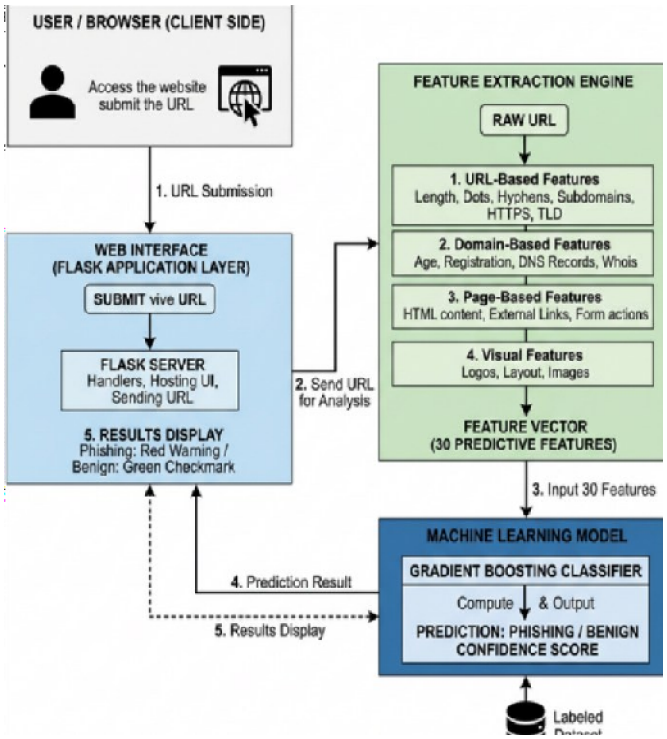


Fig.1.SystemArchitectureoftheAI-PoweredPhishingURLDetectionSystem

B. ApplicationLayer:

TheinputURLispassedtothebackend,wheretheFlaskweb application processes the HTTP POST request. The main application layer is located in the app.py file, which manages the entire application, including communication between different modules.

C. FeatureExtractionModule:

TheURLispassedtothefeature.pyfile,whichisthefeature extractionmodule.Inthismodule,theURLisexaminedand30 lexical and structural featuresareextracted and encoded in a numerical featurevector, which isa set ofnumerical features common in phishing URLs.

D. ClassificationModule:

The extracted features are reshaped to a 1x30 matrix and passed to thepre-trained Gradient Boosting Classifier, stored inthemodel.pklfile.Inthismodule,thefeaturesareevaluated and a prediction made on whether the URL is legitimate or a phishing URL.

E. OutputGeneration:

Theoutputandtheprobabilityofthepredictionarepassed back to the user interface, and the webpage displays the predictionusingtheindex.htmlfile,enablingtheusertoknow whether the URL is legitimate or a phishing URL.

VI. METHODOLOGY

The phishing URL detector was developed following a welldefined workflow, which includes the following steps:

A. DataAcquisition:

The experiment began with the acquisition of a labeled list of URLs, including legitimate and phishing sites, which was saved in the file named phishing.csv. Each record is provided withasetofpredefinedstructuralandlexicalfeaturesthatare useful in the classification process.

B. ModelTraining:

The experiment involved Exploratory Data Analysis to understand the nature of the data and the features' interrelation. Various machine learning models were tested, with the Gradient Boosting Classifier proving to be the best. The trained model was saved in the file named model.pkl.

C. FeatureExtractionDevelopment:

The experiment involved the development of a Python modulethatcancalculate30lexicalandstructuralfeaturesfor anygivenURLinrealtime,whichareknowntobelinkedtothe phenomenon of phishing.

D. SystemIntegration:

The experiment involved the development of a web interface that integrates the trained model with the user interface using the Flask web development library.

E. TestingandValidation:

The experiment involved the testing of the system with legitimate and phishing URLs to gauge the system's performanceintermsofspeed.Theexperimentrevealedthat thesystemiscapableofanalyzingthenatureoftheURLinreal time with minimum delay.

VII. DATASET DESCRIPTION

The phishing detection model is trained using a dataset collected from the UCI Machine Learning Repository, which contains a collection of phishing websites. The dataset containslabeled phishingand legitimateURLs, each associated with a set of pre-computed features.

The details of the dataset used for the phishing detection model are given below:

- TotalSamples–11,000URLs
- PhishingURLs–6,000URLs
- LegitimateURLs–5,000URLs
- Features–30structuralandlexicalfeatures

TableIoutlinesasubsetofthecriticalfeaturesextractedfor classification.

TABLEI
SELECTEDFEATUREDESCRIPTIONS

Feature	Description
URL Length	MeasurestotalcharactersinURL
HTTPS	ChecksifHTTPSprotocolis used
IPAddress	DetectsifURLusesIPinsteadofdomain
'@' Symbol	Identifiespresenceof'@'in URL
URLShortening	Detectsuseofshorteningservices

VIII. IMPLEMENTATION

TheapplicationrunsunderaPythonvirtualenvironment.

- Framework: Flask isresponsibleforroutingandhandling incoming HTTP requests at the root path ("/").
- Data Handling: NumPy is used to create arrays of extracted features in the format expected by the Scikitlearn model.

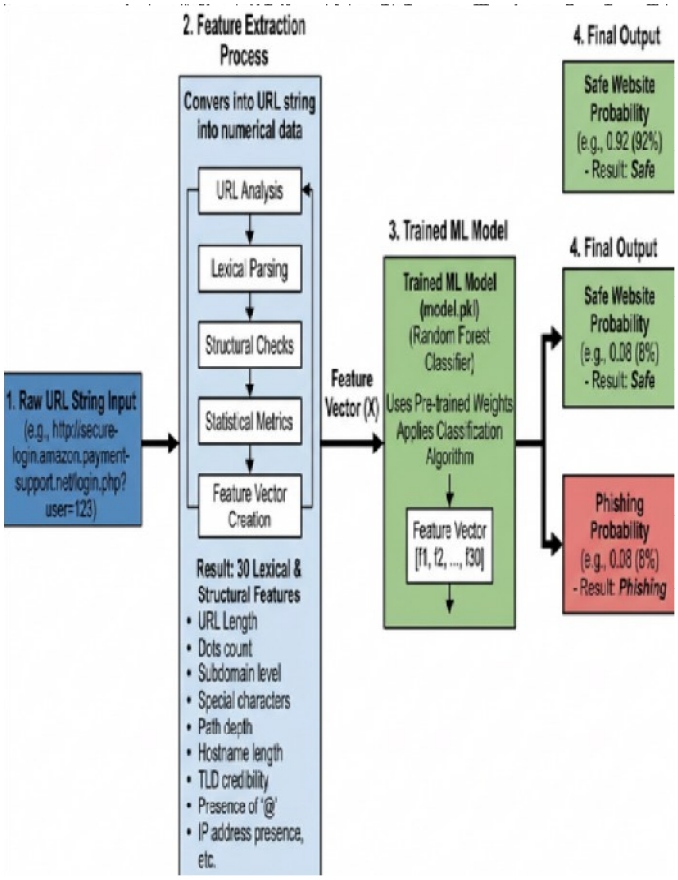


Fig.2.DataFlowDiagramoftheSystem

- Model Integration: Upon application startup, the pretrained model model.pkl is deserialized using the pickle library, ensuring that predictions occur within milliseconds when a request is received.
- User Interface: The user interface is implemented using Jinja2templating,whererender_templateisusedto dynamicallydisplaythecalculatedsafetypercentage directly below the input field.

IX. RESULTS AND EVALUATION

The phishing URL detector demonstrates satisfactory performance in identifying malicious URLs. It processes the inputURLandmakesthepredictionalmostinstantly.Thisisin accordancewiththerequirementsofreal-timeapplicationsin the field of cybersecurity.

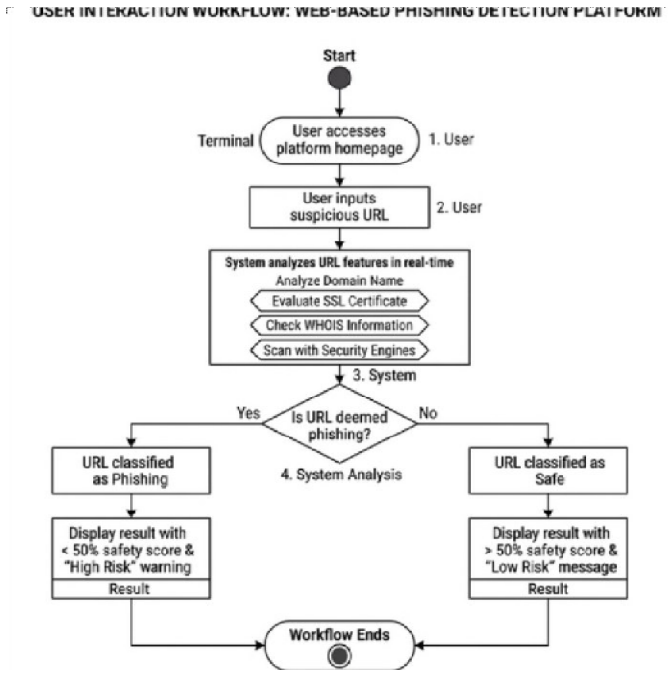


Fig.3.PlatformWorkflowforUserInteraction

- Performance:Thepredictionisgeneratedalmostinstantly after the input URL is provided. This allows the user to check the URL's safety before proceeding to visit the potentially malicious website.
- User Interaction: The user interface is quite clear in communicating the safety status of the provided URL, whether it is safe or not.
- Benefits:Theproposed method isbeneficial in thesense that it is not limited to the list of known phishing URLs, unlike other methods that are based on blacklisting. Instead, it uses the features of the URL. This is beneficial in identifying the new phishing websites that are not yet known to the system.
- PerformanceEvaluation:Theperformanceofthetrained model is determined using various metrics.

TABLEII
MODELEVALUATIONMETRICS

Metric	Value
Accuracy	96.2%
Precision	95.8%
Recall	96.5%
F1 Score	96.1%

Table II presents the evaluation metrics obtained for the proposed phishing URL detection model.

X. COMPARISON WITH EXISTING SYSTEMS

The paper compares the new machine learning approach with existing phishing detection methods. It reveals that the existing methods, which rely on blacklisting, fail to protect againstzero-dayattacks, whiletheexistingheuristicmethods are also slower. However, the new system is more adaptable and reliable.

TABLEIIICOMPARISONWITHEXISTINGSYSTEMS

Feature	Blacklist	Heuristic	ProposedML
Zero-DayDetection	Poor	Moderate	High
ProcessingSpeed	Fast	Slow	Fast
Adaptability	Low	Moderate	High

TableIIIillustratestheadvantagesoftheproposedmachine learning system compared to existing phishing detection approaches.

XI. BENEFITS OF THE PROPOSED SYSTEM

The phishing URL detection system proposed in this paper offers the following advantages over existing approaches:

1) Dynamic Analysis:ThesystemanalyzestheURLbasedon its pattern rather than depending only on blacklists or URLsthathavebeenreportedtobemaliciousinthepast.

2) High Accuracy: The proposed system uses a Gradient Boosting Classifier, ensuring high accuracy in the prediction results based on the features of URLs.

3) User-Friendly Interface: The system provides a simple user interface through the web interface, ensuring that anyonecancheckthesafetyofaURLwithouttheneedfor any technical expertise.

4) Lightweight Architecture: The proposed system is developed using the Flask framework, ensuring efficient usage of system resources and faster prediction results.

XII. FUTURE SCOPE

Theproposedsystemhasbeenabletoachievehighaccuracy in detecting phishing URLs, but there is always room for improvement:

1) Browser Extension Development: The system could also bedevelopedasabrowserextensionforbrowserssuchas Google Chrome, Mozilla Firefox, etc., to check the URL before the page is loaded in the browser.

2) Deep Learning Techniques: The proposed system could also make use of deep learning techniques that make predictions based on the raw URL content itself.

3) Analysis of Content on the Webpage: The feature extractionmodulecouldalsobeenhancedtoanalyzethe contentofthewebpageitselftodetectphishing URLsthat have been designed to look similar to legitimate URLs.

4) hortened URLs: The system could also be enhanced to handle shortened URLs before the feature extraction module is applied to the URL, ensuring that the attacker cannot shorten the URL before sending it to the victim.

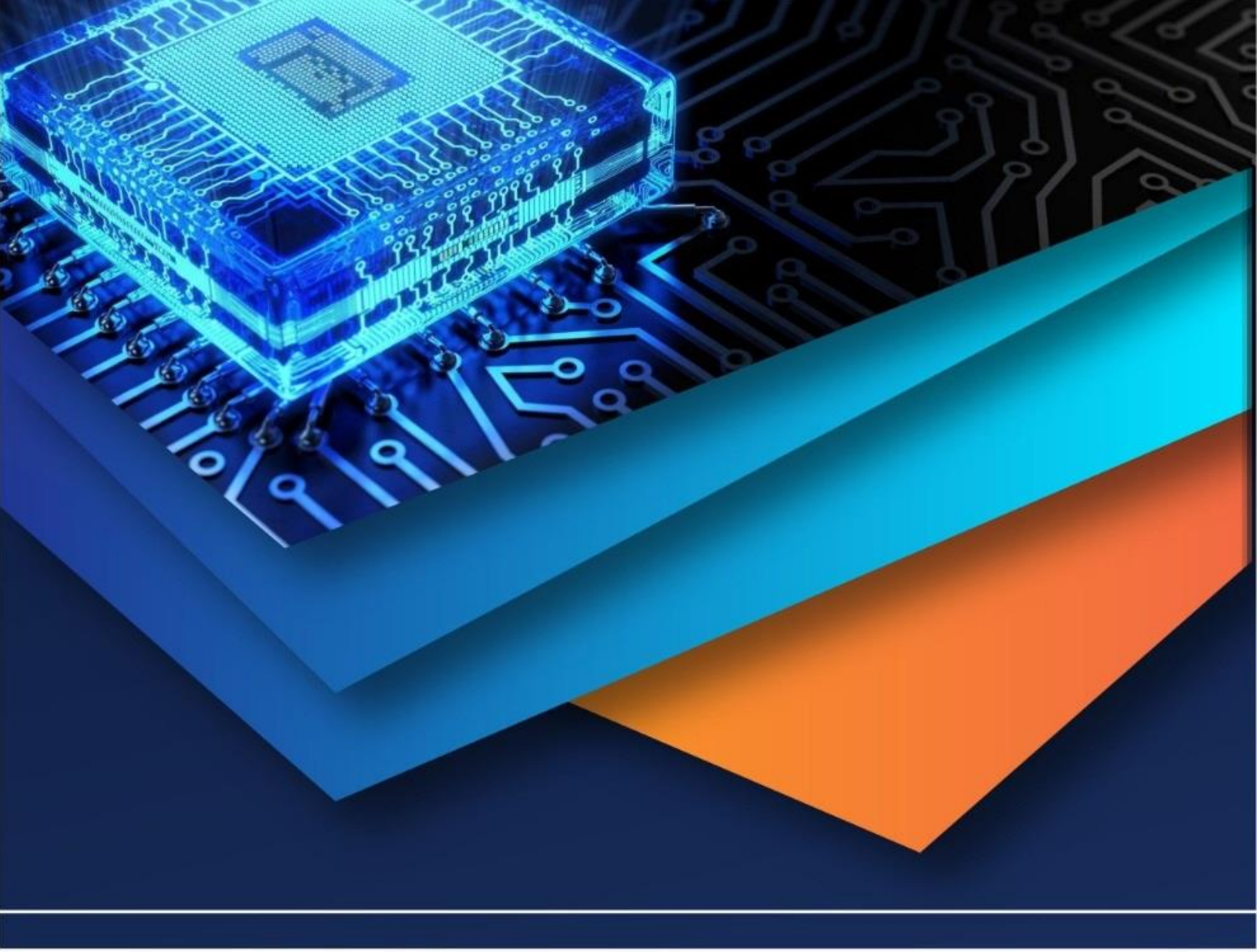
XIII. CONCLUSION

Phishing is still one of the major concerns for cybersecurity today.ThispaperproposesanAI-basedphishingURLdetector, which addresses the limitations of conventional blacklisting techniques for detecting phishing URLs. This technique involves a 30-feature extraction technique and Gradient Boosting,enablingthedynamicevaluationofURLsandmaking accurate decisions on the legitimacy of URLs. The proposed technique is implemented as a web-based application, and users can check the safety of URLs in real time using this application.Theresultsindicatethepotentialofthistechnique in providing an effective tool for increasing user protection against phishing attacks.

REFERENCES

[1] A. Jain and B. B. Gupta, “A machine learning based approach forphishing detection using hyperlinks information,” Journal of AmbientIntelligence and Humanized Computing, vol. 10, no. 5, pp. 2015–2028,2019.
[2] R. S. Rao and A. R. Pais, “Jail-phish: An improved search engine basedphishingdetectionsystem,”Computers&Security,vol.83,pp.246–267,2019.
[3] M.N.Al-Naymat,M.Al-Kasassbeh,andS.Al-Hawari,“Phishingwebsitedetection using machine learning methods,” in Proc. Int. Conf.Information Technology (ICIT), Amman, Jordan, 2021, pp. 317–321.
[4] P. Yang, G. Zhao, and P. Zeng, “Phishing website detection based onmultidimensionalfeaturesdrivenbydeeplearning,”IEEEAccess,vol.7,
[5] pp.15196–15209,2019.
[6] E.Benavides,W.Fuertes,S.Sanchez,andM.Sanchez,“Classificationofphishing attack solutions by employing deep learning techniques: Asystematic literature review,” IEEE Access, vol. 8, pp. 117456–117473,2020.
[7] S.Patil,Y.S.Kumar,K.R.C.Prasad,andT.M.K.Reddy,“Machinelearningbased phishing URLs detection,” in Proc. Int. Conf. ComputerCommunicationandInformatics(ICCCI),Coimbatore,India,2023,pp.1–6.
[8] Y. A. Alsirhani, M. Bampouli, and J. C. Malyon, “A survey of phishingdetectionusingmachinelearningtechniques,”inProc.Int.Conf.DigitalForensics and Cyber Crime (ICDF2C), 2016.
[9] W. Ali and A. Ahmed, “Hybrid intelligent phishing website predictionusing deep neural networks with genetic algorithm-based featureselection and weight optimization,” IEEE Access, vol. 7, pp. 146765–146779,2019.
[10] A. K. Jain and B. B. Gupta, “Phishing detection: Analysis of visualsimilarity based approaches,” Security and Communication Networks,vol. 2017, Article ID 5406087, 2017.
[11] S. Parekh, D. Parikh, S. Kota, and S. Sankhe, “Phishing attack detectionusingmachinelearninganddeeplearning,”inProc.Int.Conf.SmartCityand Emerging Technology (ICSCET), Mumbai, India, 2018, pp. 1–4.
[12] M. A. Adebowale, K. T. Lwin, M. A. Hossain, and T. T. V. Nguyen,“Intelligentweb-phishingdetectionandprotectionschemeusing integrated features of images, frames and text,” Expert Systems withApplications, vol. 115, pp. 300–313, 2019.
[13] H. Shirazi, K. Haefner, and I. Ray, “Understanding the vulnerability ofgradientboostingmodelstoad versarialsampleinjection, ”inProc.IEEEInt.Conf.Trust,SecurityandPrivacyinComputingandCommunications(TrustCom),2020.

[14] V. S. Urdhwareshe, M. S. Roopalakshmi, and C. S. Anoop, "A review ofURL based phishing detection using machine learning classifiers," inProc.Int.Conf.EmergingTechnology(INCET),Belagavi,India,2020,pp.1–6.
[15] A.Al-AhmadiandA.S.Al-MalaiseAl-Ghamdi, "Acomparativestudyonmachine learning algorithms for phishing detection," Journal of KingSaudUniversity-ComputerandInformationSciences,vol.34,no.8,pp.5831–5840,2022.
[16] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishingwebsitesbasedonself-structuringneuralnetwork,"NeuralComputingand Applications, vol. 25, no. 2, pp. 443–458, 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)