iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74059

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 13 Issue IX Sep 2025- Available at www.ijraset.com

PhishGuard: A Smart Approach for Malicious URL Identification with ML

Miss. Varalakshmi H[1], Mrs. Sharvani V[2]
Department of MCA, Ballari Institute of Technology & Management, Ballari, Karnataka, India

Abstract: ***With the rise of mobile device usage and increasing reliance on the internet, most real-world activities—banking, shopping, and communication—have transitioned online. While this digital shift enhances convenience, it also raises cybersecurity concerns, particularly phishing attacks. Phishing involves deceptive websites mimickinglegitimateonestostealsensitive user data like passwords and credit card numbers. Traditional security tools often fail to detect such sophisticated, zero-day attacks. This study suggests a machine learnings based phishing detections system that uses algorithms such as Random Forests, SVM, Decision Tree, Naïve Bayes and Neural Networks. These models analyze URL features to classifysitesaslegitimateorphishing.The system is deployed on a free, ad-free, non- profit website that also allows users to report suspicious URLs, enhancing accuracy over time. Tested on diverse datasets, the models achieved over 90% accuracy.Thisresearchhighlightsmachine learning'sroleineffectivelycombating phishing threats and strengthening cybersecuritydefenses.***

Keywords: Phishing Detection, Machine Learning, Cybersecurity, URL Features, Decision Tree, Support Vector Machine (SVM), Naïve Bayes, Random Fores,NeuralNetworks,Zero-dayAttacks.

I. INTRODUCTION

Phishing has emerged as among the most significant and prevalent cybersecurity concerns in today's digital environment.As more people use the internet for activities like banking, shopping, paying bills, and recharging mobile accounts, the chances of being targeted by phishing attacks have gone up a lot. These attacks takeadvantageofhow individuals act instead of looking for technical weaknesses, making them very effective. While cybersecurityexperts can usuallyspot fake websites,theaverageusermightnotbecapableof tell the difference. This can result in sensitive information like login details, passwords, and personal data being stolen without their knowledge.

Phishing is a form of online fraud where attackers pretend to be trustworthysources orpeopleto trickusersintosharingprivate information.

Reports show that phishing causes a lot of financial harm around the world.

IntheUnitedStates,Phishing attempts, for instance, havecaused losses over $2 billion each year. A 2014 study by Microsoft's Safer Computing Index estimated that the global damage from phishing could be as highas$5billionannually.Evenwithmore awareness,phishingremainsabigproblem because it can influence people's behavior and attackers can easily copyreal websites using tools like HTTrack.

Traditional ways of detecting phishing mainly rely on blacklists—lists of known bad websites and IP addresses kept by antivirus programs.

However, attackers have now developed better methods like hiding URLs, using fast-changing hosting, and automatically creating domains to get past these lists.

Thus, depending only on blacklists is no longer enough to catch new or unknown phishingsites.Heuristicmethodstrytofind suspicious websites by looking at certain signs, butThey frequently offer too much false alarms and can't be very accurate consistently.

To fix these issues, Machine learning (ML) is now being used by researchersfor better phishing detection.Machine learning is part In artificial intelligence (AI), which enables systems to generate predictions and learn from data or decisions without needing detailed instructions.

In phishing detection, ML Training is done with datamodels sets that include both phishing and real websites. These algorithms have the ability to recognize patterns that in differentiating between phishing and authentic websites.

The length and structure of URLs, whether they use HTTPS, the domain's age, and other characteristics are frequently employed in phishing detection. IP addresses being present in the URL, the odd terms used and the website's resemblance to a legitimate one.

©IJRASET: All Rights are Reserved

Supervised learning techniques like Decision Trees, Support Vector Machines (SVM), Naïve Bayes, and Neural Networks are used to identify if a URL is phishing or legitimate.

One big challenge in detecting phishing attempts with machine learningis not having enough large, good quality public data sets.

Not enough data can reduce the detection accuracy and weaken the model training.

So, models must be trained on varied and complete pieces of data and evaluated with fresh data that hasn't been seen before. Also, educating users is important, as even the best system can't protect against attacks if users aren't careful or informed.

As stated by Interisle Consulting Group, phishing attacks jumped by 61% between 2020and2022,withtheuseoffakedomain names increasing by 82%.

55% of phishing websites impersonated well-known businesses, including Amazon, Google, Facebook, WhatsApp, Netflix, and Apple, according to a 2020 study by F5 Labs.

These attacks are hard to spot and stop because they keep changing.

To deal with this, this project introduces a web-based, user-friendly, and ad-free machine learning-based application to detect phishing URLs.

Users can type in a URL, and the app determines the website's security using a machine learning algorithm.

It also has a way to report suspicious URLs,helpingmaketheonlineworldsafer.

Thoughthesystemhassomelimits,likeits reliance on the caliber of the model and data used have a tremendous deal of potential to lessen the damage that phishing assaults create.

By combining user awareness with smart automateddetection,thisprojectaimsto make the internet a safer place for everyone.

II. PROJECT AIM AND OBJECTIVE

This project aims to establish a smart system that can quickly and accurately use a machine learning to identify phishing websites helpingprotectusersfromonlinefraudand unsafe websites.The system evaluates 32 important features from website URLs, their content, and their behavior to assess if a website is dangerous or authentic.By assisting users in identifying phony websites that mimic authentic ones, this effort seeks to increase internet security, thus lowering the chances of them being scammed through phishing attacks. The solution includes a simple and easy-to-use interface with a login, a dashboard, prediction results, and performance charts, enabling users to get fast and straightforward phishingdetection in the moment

- Reviewvariousautomatedmethods used for phishing detection to understand existing solutions.
- Selectthebest methods for machine learning and build a dependable detection model.
- Select and get ready a suitable dataset for the model's testing and training.
- To evaluate the model's performance, use metrics like accuracy, exactness, recall, and false positive rate.

III. PROPOSED SYSTEM

It is suggested that a combined approach will be utilized to detect phishing websites moreefficiently.'Attheoutset,itexamines URLs by comparing them to a regularly updated blacklist to identify known phishingthreats.It serves as a quick and effective initial line of defense. It can assist in identifying new or unidentified phishing websites by searching for unblocked URLs using a Random Forest model By combining these two techniques, detection may be done quickly and intelligently. To safeguard users from phishing techniques,thesystem updatesits blacklist regularly to ensure ongoing security.

Indepthpreprocessing

- HighAccuracy.
- HighEfficiency.
- FastProcessing.

IV. BACKGROUND

A. Phishing website discovery

URLAnalysisisacrucialsystemused in identifying websites that are phishing. It involveslookingnearlyatthestructure ofawebsite’sURLto spot anystrange or suspicious features that could show a phishing attack. Cybercriminals frequently make URLs that looknearly the same as real websites, but with small changes like misspelled words, redundant figures, added subdomains, or uncommonsphere consummations.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

For case, a fake website might use a URL like www.paypa1. comrather of the real www. PayPal. com. The letter" l" is replaced withthe number" 1" in this illustration, which is easy to miss, particularly for those who aren't
veritably tech-expertise.

Automated systems,frequently using Machine literacy (ML)ways, can overlookand classify URLs grounded on known phishing patterns. These models look at severalof the URL,similar as its length, the use of special characters, whether it includes an IP address, or how numerous blotches are in thesphere name. When these systems work with over- to- date lists of known phishing spots, they form a strong first step in guarding against phishing attempts.

ContentAnalysisisanotherimportant system used in phishing discovery.

It involves checking the internal corridor of a website,similar as textbook, images, links, and how the runner is structured, to find signs of fraud.

This process looks foreffects like suspicious words, like “corroborate your account”inadequately designed layouts, or mismatched brand rudiments.

ML tools can automatically check these aspects and compare them with real websites to find differences.

Contentanalysismightalsocheckfora high number of external links, the presenceofloginforms,beddedscripts, and the similarity of images.

All thesesuggestions help in determining whether a website is genuine or if it has been created with bad intentions.

Together, URL and Content Analysis greatly enhance the capability of phishing discovery systems to directly identify fake websites.

B. Phishing approach

- List- Grounded Approach Uses two lists — whitelist forlicit URLs and blacklist for phishing URLs. Access is allowedonly when the URL is availablein the white list (5), while the blacklist blocks knownviciousspots (6).
- Heuristic- Grounded Approach Analyzes the structure and pattern of phishing URLs grounded on preliminarily linked phishing spots. Bracket depends on the URL's compliance with these patterns (7).
- Visual Similarity- Grounded Approach Compares the visual appearance of websites using image processing. Garçon- side views are anatomized to spot small differences between real and fakespots (8).
- Content- Grounded Approach Excerpts and analyzesrunner content and metadata using hunt machines and DNSwaiters.Brand-relatedkeywords andensigns are used to assess authenticity (9)(10).
- Fuzzy Rule- Grounded Approach Handlesnebulous features usingfuzzy sense and predefined expert rules. Smaller, applicable features lead to advanceddelicacy (11).

C. Structure

Category	Feature	Description
URL-Based Features	URLLength	LongerURLs may indicate phishing.
	Presence of “@”Symbol	Phishers use “@”toredirect URLs.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Category	Feature	Description
	Presenceof "//" after Protocol	Indicates redirection.
	UseofIP Address insteadof Domain	SuspiciousifIP is used instead of domain name.
	Number of Subdomains	More subdomains= suspicious.
	HTTPS (SSL Certificate) Used	LackofHTTPS is a red flag.
	Presenceof Hyphen in Domain Name	Legitdomains rarely use hyphens.
	URL Shortening Services	Like bit.ly or tiny urlindicate hiding real URL.
	Favicon Domain Consistency	Different faviconhostis suspicious.
	Non- Standard PortUsage	Unusualports may indicate phishing.
	URL Contains Suspicious Words	Wordslike "login," "secure," "bank."
	Prefix/Suffi xinDomain	Unusual additions to knowndomains.
	PathLength	Long or complicated paths are suspicious.
	Use of Special Characters	%,&,$,=,etc. in URLs raise suspicion.
	HTTPS Tokenin URL	Using"https"in a non-secure URL.

Category	Feature	Description
Domain-Based Features	DomainAge	New domains aremorelikely phishing.
	Domain Expiry Period	Short expiry suggestsfraud.
	DNSRecord Availability	NoDNS = suspicious.

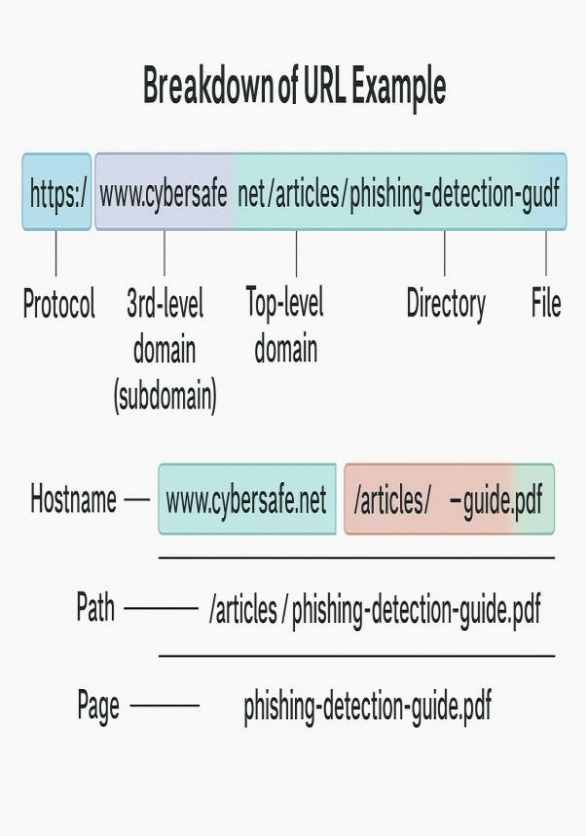
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

	WHOIS Data Availability	HiddenWHOIS info may be malicious.
	Registration Length	Lessthan 1year =highrisk.
	Name Server Consistency	Frequent changesare suspicious.
	Number of MXRecords	LimitedMX records = suspicious.
	Alexa Ranking	Low/no rank = likelyphishing.
	Web Traffic	Veryloworno traffic = red flag.
HTML/ Java Scrip t / Content-Based Features	Presence of I FrameTags	Used to hide maliciouscode.
	MouseOver Behavior	Changesstatus bar on hover.
	Right-Click Disabled	Prevents user frominspecting elements.
	Form Handlingin JavaScript	MaliciousJS form submission.
	Numberof External Links	Highnumber= suspicious.
	Website Popups	Excessive popups = phishingtrait.

Category	Feature	Description
	Anchor Tagswith Empty/# Links	Usedtomislead users.
	Script Length or Obfuscation	Long/obfuscate dscriptssuggest phishing.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com



V. MACHINE LEARNING

Sites that are phishing may be identified by the accurate and efficient application processing of machine learning algorithms. Among the often utilized algorithms are Naive Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbor (KNN). The algorithms are trained using both authentic and phishing data. website features to learn patterns and classify new URLs efficiently. Machine learning has become a popular detection method because of its capacity to identify new forms of phishing and deliver dependable results. Its capacity to evaluate enormous volumes of data and generate accurate forecasts is unmatched by traditional techniques.

A. DecisionTreeAlgorithm[5]

A Decision Tree Algorithm is one of the easiest and most common methods in machinelearning.Itworkslikeaflowchart, where each question splits the data into smaller groups. Each node (point) asks a question, and each branch shows the possible answer. Finally, theleafnodesgivetheresultorclass.To decide the best questions to ask, the algorithm uses measures like Information GainortheGiniIndex,whichhelppickthe most crucial characteristics for data separation.

A more sophisticated variant of this is Random Forest, which creates stronger and more accurate predictions by combining many decision trees. A more complex variant of this is a randomly generated forest, which combines many decision trees to provide more precise and dependable forecasts.

B. RandomForestAlgorithm[6]

The Random Forest algorithm is an effective machine learning technique that's builtonthe idea of decision trees. Instead of using just onetree,itcreatesmanytrees(a“forest”),and more trees usually mean better accuracy.

Tobuildthesetrees,itusesamethodcalled bootstrapsampling,whererandomsamples andfeaturesarechosen(withreplacement) fromthedataset.Foreachsplitinatree,the algorithm picks the best feature using measureslikeGiniIndexorInformation Gain.Thisprocesscontinuesuntiltheforest has the required number of trees.

Each tree makes a prediction, and the final answer is decided by majority voting—the resultthatmosttreesagreeonistakenasthe prediction.

©IJRASET: All Rights are Reserved

C. Algorithm for Support Vector Machines

Support Vector Machine (SVM) is one efficient classification techniquein machine learning. It works by plotting data points in an n-dimensional space and finding the best separating line, called a hyperplane, to divide the data into two classes.

Thealgorithmlooksforthesupportvectors (the closest data points to the hyperplane) andplacesthehyperplanein a manner thatthe margin (distance between the hyperplane and the support vectors) is as wide as possible, which improves classification accuracy.

The kernel trick is a technique used by SVM to convert complicated or non-linear data into a higher-dimensional space where separation is possible, even if the data cannot be separated with a straight line.

D. K-NEARESTNEIGHBORS

Computers may easily make choices by analyzing historical data using the K-Nearest Neighbors (KNN) technique. Let's say you need to know how to classify a new point that you have encountered.Mostof the neighbors in the "K" area are analyzed by KNN. Afterward, it grants the group a new point. Its method of measuring closenessisoftenbasedondistance,suchas Euclidean distance.

VI. SCOPE

The scope of URL phishing detection involves creating and deploying advanced algorithms and techniques to identify and block malicious activities using deceptive URLs. It focuses on real-time detection With machine learning's assistance , AI, and behavioral analysis, along with integration into web browsers and security tools. The scope also includes educating users, ensuring compatibility across devices, minimizingfalsepositives,andmaintaining compliancewithcybersecurityregulations.

Continuous monitoring and updates are essential to adapt to evolving threats. Ultimately, the goal is to deliver comprehensive protection, safeguarding users'personaldataandonlinetransactions from phishing attacks effectively.

VII.LITERATURE REVIEW

Phishingremainsoneofthemostpervasive threats in the cybersecurity landscape, especially as cybercriminals increasingly use sophisticated techniques to evade detection. One of the main strategies in modern phishing detection is the application of machine learning (ML) techniques. These models can identify patterns, extract pertinent information, and categorize URLs as either authentic or fraudulent.

A notable study implemented an Extreme Learning Machine (ELM) model using 30 unique features extracted through machine learning techniques. Since many phishing websites now use HTTPS to appear trustworthy, detection methods must look beyond just the protocol. Generally, phishing detection is carried out in three ways:analyzingURLstructures,evaluating domain authority, and verifying content authenticity. The study combined highly cor-relatedfeaturesfromtwodifferent datasets—URL-based, domain-based, and content-based—and tested multiple machinelearningsmodels.Amongthem,the Random Forest (RF) algorithm delivered the best results.

Models like SVM, Neural Networks, Decision Trees, and XGBoost were used to analyze URLs supplied by users. Decision Tree received an accuracy score of 82.4%, whereas Random Forest received 87.0%. Another research proposed a method to cut false positives by 30% by combining WHOIS data with classification and association algorithms.

Research conducted in 2017 used Random Forest to analyze phishing characteristics andreachedanaccuracyof98.8%using26 feature combinations. Browser-based protectionmechanismswerealsoproposed. Similarly, in another study, four models were compared to explain phishing behavior and improve detection performance. However, the achieved accuracy was not always reported.

A report by Threat Labs highlighted a significant increase in phishing incidents between 2020 and 2021, with sectors like retail,government,andeducationbeingthe most affected. To address this, researchers have frequentlyrelied on machine learnings models. According to one research, Random ForestoutperformedSVMandDecision Tree in accuracy and false-positive rate, achieving 97.14% detection success.

Another study utilized WEKA tools and public datasets, reporting that Random Forest reached 97.36% accuracy in phishing detection. Feature selection was emphasized in several works, as it helps reducedataredundancyandimprovemodel performance.Amoreadvancedframework proposed by Korkmaz et al. incorporated hybrid models for phishing detection, achieving94.59%accuracythrough10-fold cross-validation.

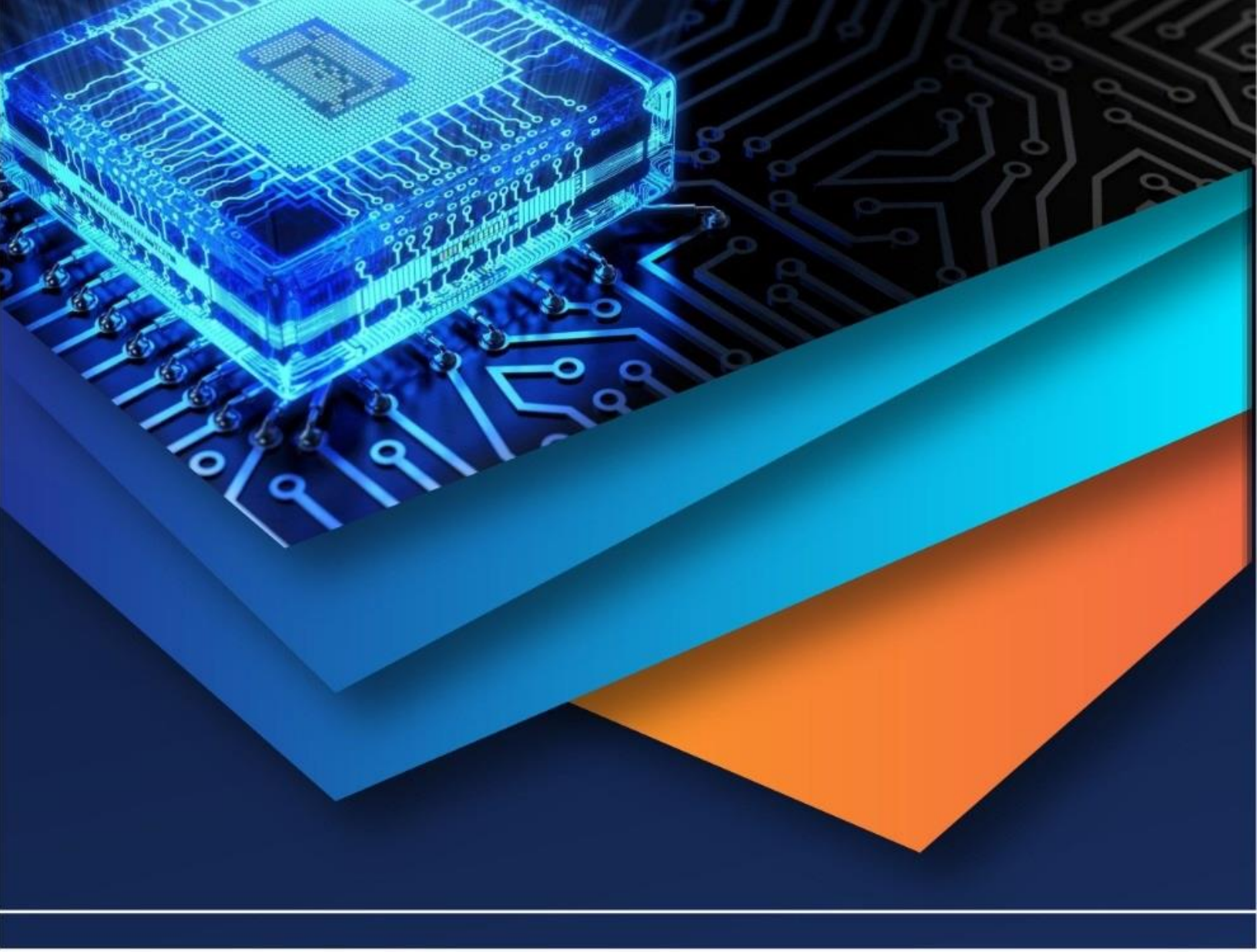
Observation

Using machine learning to identify fraudulent websites helps automatically spot fake or harmful websites by studying patterns in web addresses, domain details, and site behavior.

Models such as SVM, Random Forest and Neural Networks are trained to tell the difference between real and phishingsiteswithhighaccuracy.Asthese modelslearnfromnewdata,theygetbetter atcatchingnewertypesofphishingattacks. This approach not only speeds up the detection process but also reduces errors, making it easier to protect users and businesses from online scams. In short, machinelearningmakesphishingdetection smarter, faster, and more reliable.

REFERENCES

[1] G. Karabatis and A. AlEroud, "Bypassing Detection of URL-Based Phishing Attacks Using Generative Adversarial Deep Neural Networks," in Proc. 6th Int. Workshop on Security and Privacy Analytics, 2023.
[2] M.Arivukarasi,A.Manju,R.Kaladevi, S. Hariharan, M. Mahasree, and A. B. Prasad, "Efficient Phishing Detection and Prevention Using SVM Algorithm," in Proc. IEEE Int. Conf. on Communication Systems and Network Technologies (CSNT),Bhopal,India,pp.545–548,2023.
[3] A. Mohamed, G. Özdemir, and S. Alrefaai, "Detecting Phishing Websites Using Machine Learning," in Proc. Int. CongressonHuman-ComputerInteraction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2022.
[4] B.Sabir,M.A.Babar,R.Gaire,andA. Abuadbba, "Reliability and Robustness Analysis of Machine Learning Based Phishing URL Detectors," IEEE Transactions on Dependable and Secure Computing, 2022.
[5] M.Almousa,T.Zhang,A.Sarraf Zadeh, and M. Anwar, "Phishing Website Detection: How Effective Are Deep Learning-Based Models and Hyperparameter Optimization?" Security and Privacy, vol. 5, no. 6, p. e256, 2022.
[6] "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," IEEE Access, 2022, by N. Q. Do, A. Selamat, O.Krejcar, E. Herrera-Viedma, and H. Fujita.
[7] M. D. Bhagwat, P. H. Patil, and T. S. Vishawanath, "A Methodical Overview on Detection, Identification and Proactive PreventionofPhishingWebsites,"in Proc. 3rd Int. Conf. on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021.
[8] Z.Alkhalil,C.Hewage,L.Nawaf,and I. Khan, "Phishing Attacks: Recent Comprehensive Study and a New Anatomy,"Computer Science Frontiers vol. 3, p. 6, 2021.
[9] H. H. Chinaza Uchechukwu and J. Ding, "A Survey of Machine Learning Techniques for Phishing Detection," IEEE Access, August 2020.
[10] M. Korkmaz, O. K. Sahingoz, and B. Diri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," in Proc. Int. Conf. on Computing,Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1–7, 2020.
[11] "Phishing Website Classification and Detection Using Machine Learning," with J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Bindhumadhava, in Proc. Int. Conf. on Coimbatore, India: Computer Communication and Informatics (ICCCI), pp. 1–6, 2020.
[12] OFS-NN: A Successful Phishing Website Detection Model Using Neural Networks and Optimal Feature Selection, IEEE Access, vol. 7, pp. 73271–73284, 2019; E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu.
[13] IEEE Access, "Detection of Phishing Websites Using Multidimensional Features Driven by Deep Learning," vol. 7, pp. 15196–15209, 2019.
[14] Int. J. of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 7, 2019, A. Kulkarni and L. L. Brown III, "Detection of Phishing Websites Through Machine Learning."
[15] According to R. Mahajan and I. Siddavatam, "Detecting Phishing Websites Using Machine Learning Algorithms," International Journal of Computer Applications, vol. 181, no. 23, pp. 1–7, 2018.



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)