IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

Detecting Phishing Website Using Machine Learning

Prof.P.Rengasamy
Associate Professor, Department of CSE (Cyber security)
Sengunthar Engineering College (Autonomous), Tiruchengode,India
prengasamy.cse@scteng.co.in
https://orcid.org/0009-0005-0537-8373
Dhikshitha R,Harshini PR,Mythili S,Subiksha M
UG Students, Department of Computer Science & Engineering
Sengunthar Engineering College (Autonomous), Tiruchengode,India
ramasamydharsan@gmail.com,raveendaranharshini@gmail.com
Mythili6469@gmail.com,Subiksha1404@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10118
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10118
Received: 30, January 2026, Revised: 13, February 2026, Accepted: 28 February 2026 Published Online: 25 March 2026
https://www.irjcs.com/volumes/Vol13/iss-03/39.CSMR26.MRCS10118.pdf
Article Citation: Prof.Rengasamy,Dhikshitha,Harshini,Mythili,Subiksha(2026),Phishing Website Detection Using Machine Learning, IRJCS: International Research Journal of Computer Science, Volume 13,Issue 03 of 2026 pages 327-334 **Doi:>** https://doi.org/10.26562/irjcs.2026.v1303.39 **BibTeX Key Prof.Rengasamy@2026Phishing**
Orcid: https://orcid.org/0009-0004-9398-7488
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1303.39 The journal's official abbreviation is IRJCS.

Abstract: Phishing sites expect to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we measured and compared the performance of the classifier in terms of accuracy.

I. INTRODUCTION

Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that the end goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers [1], [2]. In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing end eavours in messages and identifying phishing substance on sites, phishes think of new and half breed strategies to go around the accessible programming and systems [3], [4].Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to be originating from honest to goodness sources like money related foundations, ecommerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email. The misleading sites are intended to emulate the look of a genuine organization site page.[5], [6].Employing so as to phishing invader' strap clients diverse social building strategies, for example, debilitating to suspend client accounts on the off chance that they don't finish the account upgrade process, give other data to approve their records or a few different motivations to get the clients to visit their satirize page [7],[8]. Supervised learning (Classification Technique) accommodate slavishly improved precision while unsupervised learning accommodates a quick and dependable way to deal with infer information from a dataset. That's why we used supervised learning in our work[9], [10]. Studies have demonstrated the effectiveness of various ML algorithms including Random Forest, Support Vector Machines (SVM), and deep neural networks in classifying network flows and detecting APT-related activities such as command-and-control communication, lateral movement, and data exfiltration [11],[12]. Network flow analysis, which examines metadata such as source/destination IPs, ports, protocols, packet counts, and flow durations, provides a scalable and privacy-preserving alternative to deep packet inspection [13], [14].

I. LITERATURE REVIEW

(Mao,J.,Bian,J.,Tian,W.,Zhu, S., Wei, T., Li, A., & Liang, Z. 2018)[1] In this article, we intend to further develop understanding abilities through AI. Specifically, it is suggested that the exploration technique be founded on the page design choice strategy utilized for page search. The consequences of the review show that our techniques are exact and helpful in deciding the phishing sheet. (Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary, Dr. Swapna Borde/ 2021) [2] This page analyzes the apparatuses used to learn and get machines.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

Phishing is known for its gate crashers since duping somebody is straighter forward to hit on a terrible line than beating a safeguard framework. The negative connections in the principle body of the message are expected to show that these corporate images and other real items are utilized to arrive at degenerate associations.(Ishant Tyagi; Jatin Shad; Shubham Sharma; Siddharth Gaur; Gagandeep Kaur/2018) [3] This page centers around different AI calculations pointed toward foreseeing whether a site is misled or real. Machine preparing is famous on the grounds that it can distinguish party time assaults and is greatat beating new kinds of phishing assaults. In our work, we had the option to precisely decide 98.4% by anticipating phishing or lawful area. (Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/ 2020) [4] Understanding Phishing Using Machine Learning Techniques Wahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/2020.The best method for recognizing these awful encounters is through AI. This is on the grounds that numerous phishing assaults are the most widely recognized types of AI. In this article, we think about the aftereffects of many AI strategies to foresee phishing. (Mohith Gowda HR, Adithya MV, Gunesh Prasad S & Vinay S/ 2020)[5]Inthisarticle, we need specialized ability to effortlessly recognize a phishing site on the client side that requirements to assemble a web crawler. In this framework, we utilize the erase rule to eliminate content or site highlights utilizing just the URL. The rundown comprises of 30 unique URLs and will then, at that point, be utilized to discover reality with regards to the site by irregular woods arrangement. (Banu, R., Anand, M., Kamath, A., Ashika, S., Ujwala, H. S., & Harshitha, S. N.2019) [6] The approach will also use Deep Learning frameworks with hierarchical long-short term memory networks (H-LSTMs) and attention mechanisms to model the emails simultaneously at the word and sentence level. Phishing attacks categorizes the emails based on certain properties which give more details about the source of phishing. Generally, most of the existing systems focus on email classification depending upon header part or body part. (Karabatak, M.,&Mustafa,T.2018)[7] This article inspects information assortment on the UCI site. Diminishing its size and looking at the presentation of positioning calculations is being contemplated in the news site of the phishing site. The portrayal of the phishing site is taken from the UCI information base of AI. The data set comprises of11055passages and 31 exercises. The presentation of the arranging calculation is currently contrasted with other data on the grouping calculations. At long last, contrasting the requesting elements of the informational indexes utilizing the overall calculations gave.(Shima, K., Miyamoto, D., Abe, H., Ishihara, T., Okada, K., Sekiya, Y., ... & Doi, Y. 2018) [8] In this review, we tested with realistic URL access history data taken from a research organization and data from the famous archive site of phishing site information, PhishTank.com. Our approach achieved 2~3% better accuracy compared to the existing DL- based approach.

PROPOSED METHODOLOGY ARCHITECTURE

There are innumerable spaces that can be defrauded, like web- based installments, webmail, monetary foundations, document stock piling or distributed storage. We band online installments are remembered for the rundown of best practices. Since phishing should be possible by email or lance phishing, the client ought to know about the effect and not be 100% certain about the general security activity. Machine preparing is probably the most ideal way to master phishing procedures as it dispenses with the risks of existing strategies. Endeavors togather individual data deceitfully a returning out to be more normal today. To assist clients with knowing how to access such a site, a framework has been executed that tells clients by means of email and a spring up window when they attempt to get to the site.

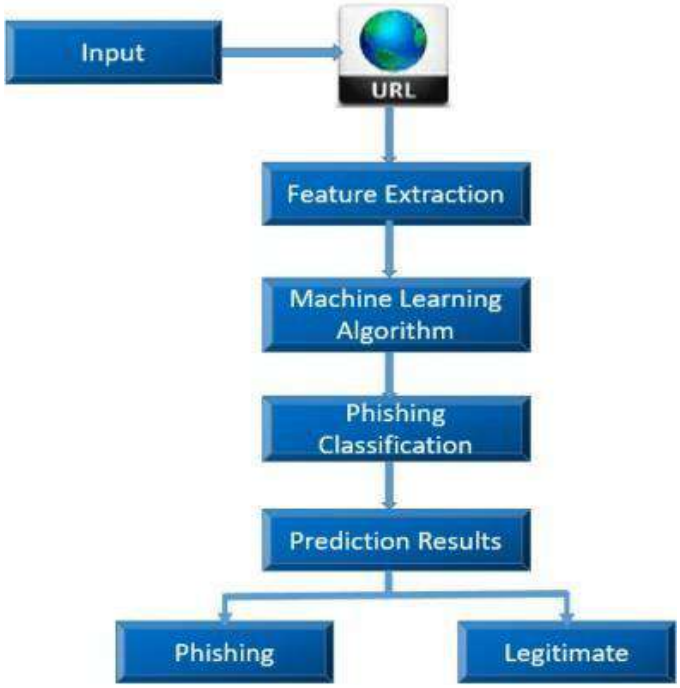


Fig.1.ArchitectureDiagram

This page gives a boycott identification framework known as a phishing site with the goal that the site can be told when looking or signing in.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

Along these lines, it tends to be utilized as a genuine apparatus to distinguish, convince, and forestall misrepresentation. A choice tree that makes a class or retreat structures a tree. Little information is erased as the related tree develops. The choice takes at least two branches, and the leaves show the arrangement or choice. The most noteworthy exactness in the tree grouping compares to the anticipated root. The authentication tree can be utilized both by classification and number. A choice tree that makes a class or retreat structure as a tree. It is designed together and utilizes the on the off chance that standard, which is characterized exhaustively by class. Methods are concentrated successively utilizing concurrent preparation data. However long you gain proficiency with the law, you can kill the twists of the law. This cycle will proceed until the preparation bundle meets the prerequisites. It depends on the "offer and win" circle through and through. All properties should be characterized. In any case, they need to think that it is first.

TECHNOLOGIES USED

A. DETECTION TECHNIQUE

Detection of phishing websites has received a lot of attention recently due to their impact on users' security. Therefore, many techniques have been developed to detect phishing websites varying from communication-oriented techniques, such as authentication protocols, black listing, and white-listing, to content- based filtering techniques. The blacklisting and white-listing techniques have not proven though to be sufficiently efficient when used in different domains, and thus they are not commonly used. Mean while, the content- based phishing filters have been widely used and have proven to be of high efficiency. In light of this, researches have focused on content-based mechanism and on developing machine learning and data mining techniques based on the header and body of emails.

Address Bar based Features

If an IP address is used as an alternative of the domain name in the URL, such as " http://125.98.3.123/fake.html",user scan besure that someone is trying to stealtheir personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link " http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Rule: IF $\{If\,T\square eDomain\ Part\square asanIPAddress \rightarrow P\square is\square ing\ Ot\square erwise \rightarrow Legitimate$

Long URL to Hide the Suspicious Part

Phishers can use long URL to hide the doubtful part in the address bar. For example:
http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.

Rule: IF$\{URLlengt\square$<54$\rightarrow feature=Legitimateelseif\ URLlength$≥54$and$≤75$\rightarrow feature=Suspicious$ Using URL Shortening Services" Tiny URL"

URL shortening is a method on the "World Wide Web"in which a URL maybe made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an "HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL. For example, the URL " http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4".

Rule:IF$\{TinyURL \rightarrow P\square is\square ingOt\square erwise \rightarrow Legitimate$ URL'shaving"@" Symbol

Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

Rule: IF$\{UrlHaving@Symbol \rightarrow PhishingOtherwise \rightarrow Legitimate$

Sub Domain and MultiSub Domains

Let us assume we have the following link: http://www.hud.ac.uk/students/A domain name might include the country-code top-level domains(ccTLD), which in our exampleis "uk". The "ac" part is shorthand for "academic", the combined "ac.uk" is called a second-level domain (SLD) and "hud" is the actual name of the domain. To produce a rule for extracting this feature, we firstly have toom it the (www.) from the URL which is infact a subdomain in itself.

PORT	Service	Meaning	Preferred Status
21	FTP	Transfer files from one host to another	Close
22	SSH	Secure File Transfer Protocol	Close
23	Telnet	Provide a bidirectional interactive text-oriented communication	Close
80	HTTP	Hyper test transfer protocol	Open
443	HTTPS	Hyper text transfer protocol secured	Open
445	SMB	Providing shared access to files, printers, serial ports	Close
1433	MSSQL	Store an retrieve data as requested by other software Applications	Close
1521	ORACLE	Access oracle database from web.	Close
3306	MySQL	Access MySQL database from web.	Close
3389	Remote Desktop	Allow remote access and remote collaboration	Close

Then, we have to remove the (ccTLD) if itexists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as "Suspicious" since it has one sub domain. However, if the dots are greater than two, it is classified as "Phishing" since it will have multiple sub domains. Otherwise, if the URL has no subdomains, we will assign "Legitimate" to the feature.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

Rule: IF {$DotsInDomainPart$=1→$LegitimateDotsIn$
$DomainPart$ =2→$SuspiciousOt\square erwise$→$P\square is$
HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough. The authors in (Mohammad, Thabtah and McCluskey 2012)(Mohammad, Thabtah and McCluskey 2013) suggest checking the certificate assigned with HTTPS including the extent of the trust certificate issuer, and the certificate age. Certificate Authorities that are consistently listed among the top trustworthy names include: "GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, DosterandVeriSign".Furthermore,bytestingoutourdatasets,we find that the minimum age of a reputable certificate is two years. Rule: IF {$Use\square ttpsandIssuer\ IsTrusted\ andAgeofCertificate$≥1 $Years$→$LegitimateUsing\square ttpsandI$ Domain Registration Length Based on the fac that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.Rule: IF {$DomainsExpireson$≤1$years$→$Phishing\ Otherwise$→$Legitimate$ Request URL

ABNORMAL BASED FEATURES

A favicon is a graphic image (icon) associated with a specific webpage. Many existing user agents such as graphical browsers and news readers show favicon as a visual reminder of the website identity in the address bar. If the favicon is loaded from a domain other than that shown in the addressbar, then the web page is likely to be considered a Phishing attempt.
Rule:IF{$FaviconLoadedFromExternalDomain$→ Table 1 Common ports to be checked Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate web pages, the webpage address and most of objects embedded within the web page are sharing the same domain.
Rule: IF{$\%ofRequestURL$<22%→$Legitimate\%of$
$RequestURL$≥22%and61%→$SuspiciousOt\square erwise$
Rule:
IF{$\%ofURLOfAnc\square or$<31%→ $Legitimate\%ofURLOf$
$Anc\square or$≥31%And≤67%→$Suspicious$
Linksin<Meta>,<Script>and<Link>tags Server Form Handler (SFH)
SFHs that contain an empty string or"about: blank" are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.
Rule: IF {$SFHis"about:blank"Or\ IsEmpty$→$P\square is\square ingSFH$
$Refers\ To\ ADifferentDomain$→$Suspicious$ HTML and Java Script based Features

Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least4times.Rule:IF{$\#ofRedirect$
$Page$≤1→$Legitimate\#ofRedirectPage$≥2$And$<4→ $SuspiciousOt\square erwise$→$P\square is$
Status Bar Customization Phishers mayuse JavaScript to show a fake URL in thes tatus bar to users. To extract this feature, we must dig-out the webpage source code, particularly the "on Mouse Over" event, and check if it makes any changes on the status bar. Rule:IF{$onMouseOverC\square angesStatusBar$→$P\square is\square ingIt\ Does't\ C\square angeStatus$
Bar→$Legitimate$

B. DOMAIN BASED FEATURES

Age of Domain

This feature can be extracted from WHO IS database(Whois2005). Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6months.
Rule: IF{$AgeOfDomain$≥6 $mont\square s$→$LegitimateOt\square erwise$→$Phishing$

Statistical-Reports Based Feature

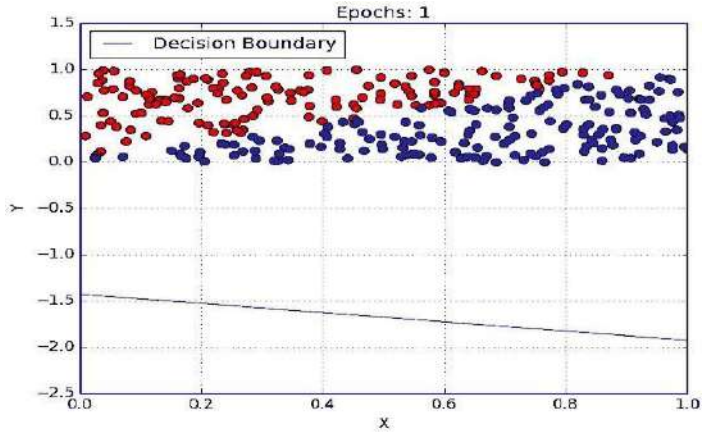
Several parties such as PhishTank (PhishTank Stats, 2010-2012), and Stop Badware(StopBadware,2010-2012) formulate numerous statistical reports on phishing websites at every given period of time; some are monthly and others are quarterly. In our research, we used 2 forms of the top ten statistics from PhishTank: "Top Domains" and "Top 10 IPs" according to statistical reports published in the last three years, starting in January2010 to November 2012. Whereas for "Stop Badware", we used "Top 50" IP addresses. Rule:IF{$HostBelongstoTopP\square is\square ingIPsorTopP\square is\square ing$
$Domains$→$P\square is\square ing\ Ot\square erwise$→$Legitima$

C. PHISHING WEBSITES FEATURES

One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been is seminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features. In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.



K-NearestNeighborNaive Bayes

Decision Trees/Random Forest Support Vector Machine Logistic Regression Regression: While a Regression problem is when the target variable is continuous (i.e. the output is numeric).

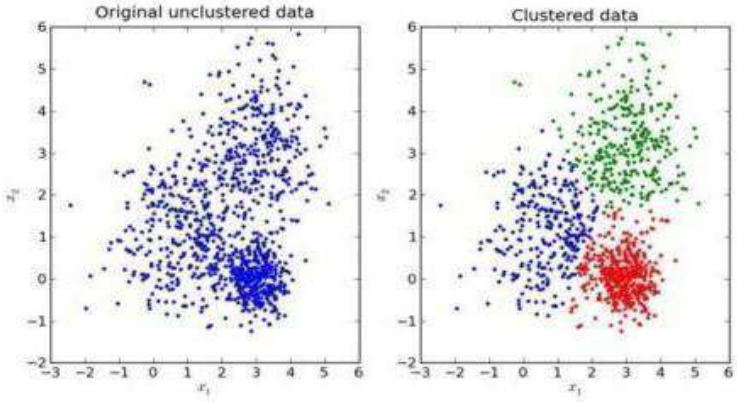


Fig.2 Clustering

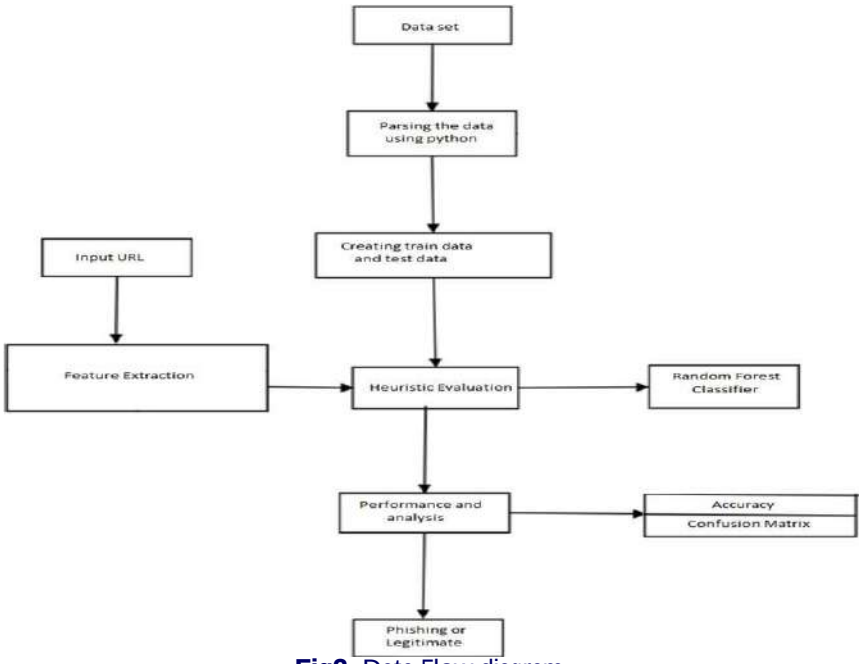


Fig2: Data Flow diagram

As shown in the above representation, we can imagine that the graph‘s X-axis is the _Test scores‘ and the Y-axis represents _IQ‘. So we try to create the best fit line in the given graph so that we can use that line to predict any approximate IQ that isn‘t present in the given data. These some most used regression algorithms: Support Vector Regression Decision Tress/Random Forest Gaussian Progresses Regression Ensemble Methods

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

IMPLEMENTATIONS AND RESULTS

The input design is the link between the information system and the user. It comprises the developing specifications and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read at a from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

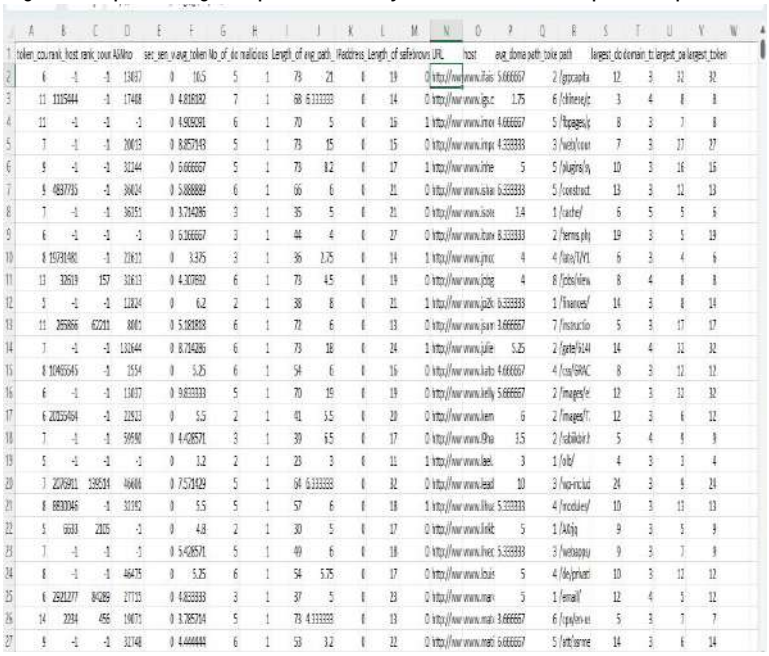


Fig 3. Data Set Output

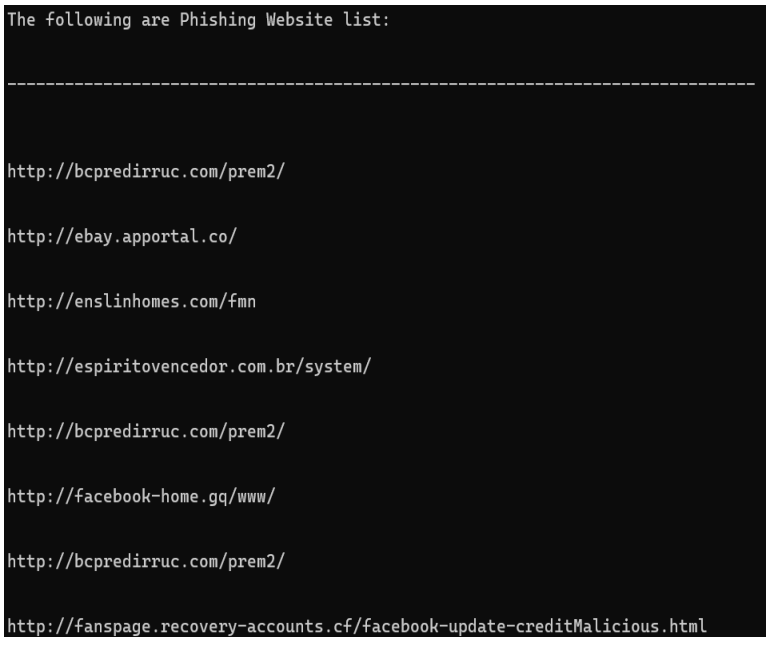


Fig: 4: Output page

Convey information about past activities, current status or projections of the Future. Signal important events, opportunities, problems, or warnings. Trigger an action. Confirm an action

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

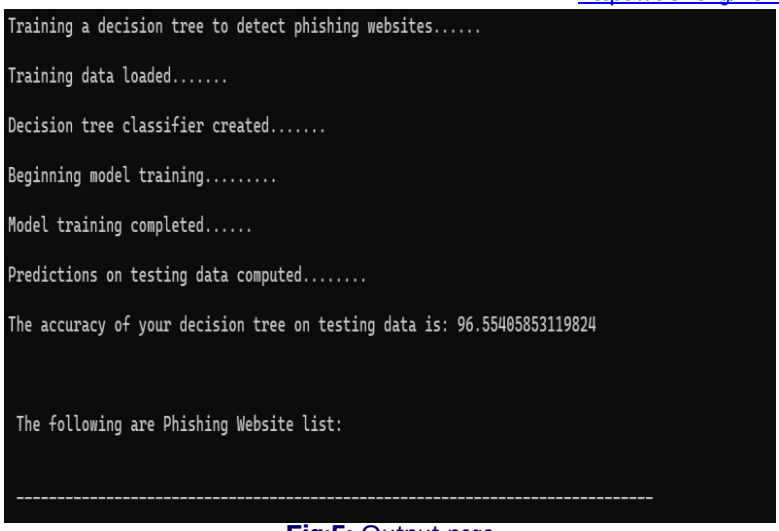


Fig:5: Output page

Several parties such as PhishTank (PhishTank Stats, 2010-2012), and StopBad ware(StopBadware,2010-2012) formulate numerous statistical reports on phishing websites at every given period of time; some are monthly and others are quarterly. In our research, we used 2 forms of the top ten statistics from PhishTank: "Top Domains" and "Top 10 IPs" according to statistical reports published in the last three years, starting in January2010 to November 2012. Whereas for "Stop Badware", we used "Top 50" IP addresses.

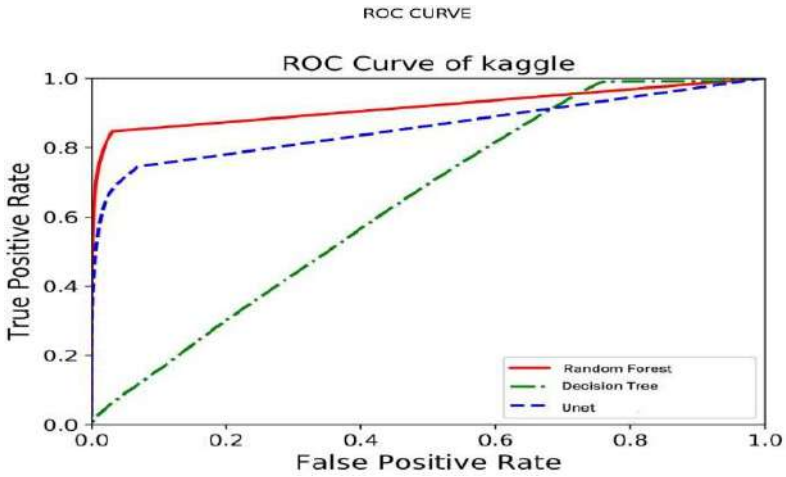


Fig: 6: ROC graph

CONCLUSION

Phishing is a cybercrime procedure utilizing both social building and specialized deception to take individual sensitive data. Besides, Phishing is considered as another extensive type of fraud. Experimentations against recent dependable phishing data sets utilizing different classification algorithms have been performed which received different learning methods. The base of the experiments is accurate measure. The aim of this research work is to predict whether a given URL is phishing website or not. It turns out in the given experiment that Random Forest based classifiers are the best classifier with great classification accuracy of 82.644% for the given dataset of phishing site. As future work we might use this model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm's in terms of classification accuracy.

REFERENCES

1. Mao,J.,Bian,J.,Tian,W.,Zhu,S.,Wei,T.,Li,A.,& Liang, Z. (2018). Detecting phishing websites via aggregation analysis of page layouts. Procedia Computer Science, 129, 224- 230.
2. Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary,SwapnaBorde "Detection of Phishing Websites using Machine Learning",International Journal of Engineering Research & Technology. 10(5), 2527-2531, 2021
3. Tyagi, I., Shad, J., Sharma, S., Gaur,S., & Kaur,G. (2018, February). A novel machine learning approach to detect phishing websites. In 20185th International conference on signal processing and integrated networks (SPIN) (pp. 425-430). IEEE..
4. Shahrivari, V.,Darabi, M. M., & Izadi, M. (2020). Phishing Detection Using Machine Learning Techniques. arXiv preprint arXiv:2009.11116.
5. HR,M.G.,Adithya,M.V.,&Vinay,S.(2020). Developmen to fanti-phishing browser based on random fore stand rule of extraction framework. Cybersecurity, 3(1), 1-14.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.39

6. Banu, R., Anand, M., Kamath, A., Ashika, S., Ujwala,H.S.,&Harshitha,S. N.(2019,May). Detecting phishing attacks using natural language processing and machine learning. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1210-1214). IEEE.
7. Karabatan, M., & Mustafa, T.(2018, March). Performance comparison of classifiers on reduced phishing website dataset. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-5). IEEE.
8. Shima, K., Miyamoto, D.,Abe,H.,Ishihara, T.,Okada, K.,Sekiya, Y.,&Doi,Y.(2018, February). Classification of URL bit streams using bag of bytes. In 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)(pp.1-5).IEEE.