Anonymity and Confidentiality in Website using Machine Learning

Anshuman Soni^{#1}, Aditya Mishra^{#2}, Anuj Mishra^{#3}, Shivam Singh^{#4}, Mr. UmaShanker^{#5}

¹⁻⁴ GNIOT(Engg. Institute), Department of Computer Science (AI&ML), ⁵Guide

Dr. A.P.J. Abdul Kalam Technical University.

Uttar Pradesh, India

¹anshumansoni4546@gmail.com, ²adityamishra.am71@gmail.com

cybersecurity experts are seeking reliable and stable detection techniques for identifying phishing websites.

This project utilizes machine learning technology to detect phishing URLs by analysing various features of both legitimate and malicious URLs.

Decision Tree, Random Forest, and Support Vector Machine algorithms are employed, with the objective of identifying the best-performing model by comparing accuracy rates, as well as false positive and false negative rates of each algorithm.

Keywords— Phishing, Cyber Security, URL, Machine Learning, Decision Tree, Random Forest, Support Vector Machine.

I. INTRODUCTION

Phishing websites are increasingly sophisticated, targeting unsuspecting users by mimicking legitimate sites to steal sensitive information like login credentials and financial data. Today, phishing has become a major cybersecurity concern due to how easy it is to create realistic-looking fake sites. While experts may identify fake sites, many regular users fall victim to these tactics, particularly in cases where attackers aim to acquire bank account credentials. In the United States, businesses lose around \$2 billion annually to phishing as clients become victims of these scams. Furthermore, the 2014 Microsoft Computing Safer Index Report estimated the global financial impact of phishing to be as high as \$5 billion, highlighting a widespread issue stemming from low user awareness. Phishing attacks exploit user weaknesses, making it challenging to fully prevent them while underscoring the need for enhanced detection techniques.

A common method to identify phishing sites involves using blacklisted URLs or IP addresses stored in antivirus databases-known as the "blacklist"

method. However, attackers can evade these blacklists Abstract— Phishing attacks are the simplest way to through URL manipulation and techniques like fast flux, obtain sensitive information from unsuspecting users. which uses automatically generated proxies to host Phishers aim to acquire critical data such as usernames, malicious pages, and algorithms that produce new URLs. passwords, and bank account details. In response, The primary downside to this approach is its inability to detect zero hour phishing attacks, which emerge too quickly for blacklist updates. Heuristic-based detection methods can identify some zero hour phishing attempts by focusing on typical characteristics of phishing attacks, but these features are not always present, leading to high falsepositive rates.

> To address the limitations of blacklist and heuristicbased methods, many researchers are now turning to machine learning techniques. Machine learning models require extensive data to predict phishing threats by analyzing past examples of legitimate and phishing URLs. This approach allows algorithms to assess a range of URL features and accurately detect phishing websites.

II. LITERATURE SURVEY

SR	PAPER	AUTHOR	PROPOSED	DISCUSSION
NO.	TITLE	NAME	METHODOLOGY	
1.	Phishing Website Detectio n Using Machine Learning : A Review	Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh	Subdomain and Principal Domain: Identifying phishing patterns. PageRank: 92% of phishing sites have low or no PageRank, indicating risk. Alexa Rank: Sites above 100,000 are often suspicious; those below are typically legitimate. Path Domain and Alexa Reputation: These elements help assess the URL's trustworthiness and overall site reputation.	The authors propose a phishing detection method using six URL-based criteria, including subdomain, main domain, and ranking factors, achieving 97% accuracy on PhishTank an DMOZ data.

³anujmishra7069@gmail.com,⁴shivamsinghcse19@gmail.com,

⁵umashanker.aiml@gniot.net.in

No. Title NAME	CD	DADED	AUTHO	BDODOCED	DISCUSSION	CD	DADED	AUTHOR	DDODOCED	DISCUSSION
2. A novel approach a protect a spread and are accuracy of examining the learning and states at white-list white-list white-list white-list white-list and process of the legitimacy of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated with the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within th	SR NO.	PAPER TITLE	AUTHO R NAME		DISCUSSION	SR NO.	PAPER TITLE			DISCUSSION
to protect any phishing phishing attacks at client side using auto- updated white-list All minimic personal auto- upda				Y					OGY	
by protect dam & B. (dentification of purphishing attacks at client side using auto- quoto- qualities white-list white-li	2.									
gasjinst phishing attacks at chent side using auto- updated white-list He have been added to a user specific white-list lengthmay be been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website been added to a user specific white-list lengthmay of the website website. Alexa and lengthmay of the website website, a caphibility of identifying novel machine learning fusion approach 3. Phishing website learning fusion approach approach approach which are applied in this research including the features such as web traffic, URL size and IP address of the larget website. There is proposed a novel fusion classifier with two priority algorithms (PA1) and PA2) resulting in phishing website detection in addressing phishing website detection in phishing website detection in phishing website detection in phishing website detection with a decision research in claims and neutral resulting in phishing website detection in addressing phishing website detection with a decision research proposed as novel flusion classifi		to protect		identification of		4.	Phishing	Α.	Bv	While the system
disclose any personal information on websites other than what has been added to a user specific whitelbist size It checks the legitimacy of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the website by examining the links incorporated within the HTML codes of the websage. Performance accuracy of 89.3 8% which proved the system's achieved an accuracy of 39.3 8% which proved the system's website detection and accuracy of identifying phishing website. Learning the features such as proposed a povel fusion clearning fusion approach approach and provided and from sources and performance in real-world applications. White a user is a matching with a user is provising and alerts for understand alerts for the legitimacy of the website website. Six It also website detection and accuracy of 39.3 8% which proved the system's achieved an accuracy of identifying phishing website detection approach approach approach approach and accuracy of identifying approach approach and accuracy of a cacuracy in a proposed an ovel fusion clearning the features such a such as decision trees, such as a surface and alerts for current and alerts for or understand and surface and provided with data from sources like surface and provided and accuracy or analysis. Alaboant, surface and provided and accuracy or analysis and accuracy or a			B. Gupta				websites	<u>Kulkarni</u>	employing	demonstrates
side using auto- updated white-list white-list should be a specific white-list should a suser-specific white-list should be a suser-specific white-list should a suser-specific white-list should be a suser-specific white-list should a suser-specific white-list s		attacks at		the users not to	DNS					classification, it
using auto- websites of the websites white-list white-list be been added to a user specific white-list beington, and a course websites. It also werifics walidity through thro				-						
updated white-list been added to a user specific whitelisted site. It checks the legitimacy of the website by examining the links incorporated within the HTML codes of the webpage. Performance evaluation using data from sources such as Stuffgate, Alexa and Phishiral achieved an accuracy of 89.38% which proved the system's capability of identifying phishing from sources machine learning flasion approach approach approach approach approach approach approach and activate of an over flasion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with growing white-learning flasion approach and policy and phishing website detection with a language in the features selection in phishing website detection with and from sources of advanced an over flasion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with and flature selection in phishing website detection with and flature selection in phishing website detection with the flature is proposed an over flasion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with the flature is proposed an over flasion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with and flature is phishing website detection with and flature is phishing website detection with and flature is phishing website detection with an analysis, achieving an advanced an accuracy of agorithms (PA1 and PA2) resulting in phishing website detection with an analysis, achieving an accuracy of accura		_					l remaining		system based	dataset and the use
user specific whitelisted site. It checks the legitimacy of the website by examining the links incorporated within the HTM. codes of the website website and far from sources such as Stuffgate, Alexa and Phish Tank accuracy of 89.38% which expandity of detection using website detection using nachine learning flusion approach a		updated		than what has	based on					
the checks the chinks incorporated within the HTML codes of the unitused detection and neural neture and neural neture and neural neture and they reached over 90 levels of accuracy of every 60 levels of accuracy in identifying genuine and malicious websites. A a novel with data and phishing websites the checkion in reachine charming fusion and they reached over 90 levels of accuracy in identifying genuine and malicious websites. A a novel checks the chinks in allow through the checks in the chinks in any over 90 levels of accuracy in identifying genuine and malicious websites. A a novel checks the chinks in any over 90 levels of accuracy of the checks the check the model's and they reached over 90 levels of accuracy in identifying accuracy of Eng		white-list								
legitimacy of the website by examining the links incorporated within the HTML codes of the webpage. Performance evaluation using data from sources such as Stuffgate, Alexa and encuracy of PhishTank achieved an accuracy of 89.38% which proved the system's capability of identifying hishing. 3. Phishing website detection using naron, P. Sturfgate, capability of identifying fusion approach approach Alexa and encuracy of aditing through hyperlink adalotes and accuracy of 89.38% which growed an accuracy of identifying PhishTank. PhishTank achieved an accuracy of identifying phishing. 3. Phishing website detection applied in this approach approach a machine learning fusion approach I sing the links incorporated with data proved the like system's capability of identifying phishing. A Lakshma decision tree, naive Bayesian, and neural enetwork were the four classifiers applied, and hyperlink and neural phyperlink and neural phyperlink achieved an accuracy of 89.38% which growed an accuracy of identifying achieving an adhesion tree, significantly achieving an achieve and phyperlink achieved an accuracy of identifying phishing websites. 3. Phishing accuracy of Systy, and random proved the larget website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website effectiveness of advanced a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing websites effectiveness of advanced and ecision tree, naive made network were the four classifiers and neural and neural and peral and every applicat, and heread over 90 levels of accuracy in identifying genuine and an envery applications. Surfigate, Alexa and Eachieved and accuracy of Eachieve and accuracy in identifying genuine and malicious websites. Surfigate, Alexa and Eachieve and Eachiev									websites.	These factors
accuracy of PhishTank acted teleption using determine learning fusion approach website approach as Sturya, Machine learning fusion approach approach a feature such as web site. There is proposed a novel fusion classifier under the target website. There is proposed a novel fusion classifier under the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing websites. A mark A mar				legitimacy of the	detection and				· ·	
sites. It also within the HTML codes of the websige. Performance evaluation using data from sources such as Stuffgate, Alexa and PhishTank accuracy of PhishTank accuracy of 89.38% which proved the system's capability of identifying phishing. 3. Phishing website detection using approach learning fusion approach 3. Phishing website detection approach accuracy of succiracy in identifying phishing and random forests have been applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority models and feature such as machine learning algorithms (PAI and PAZ) resulting in phishing website detection with dated sand feature such as machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect phishing websites of a dayloacy of machine learning approach to detect of the				examining the	untrusted					
codes of the websage. Performance evaluation using data from sources such as Stuffgate, Alexa and PhishTank achieved an accuracy of 89.38% which proved the system's capability of identifying website machine learning fusion approach approach approach approach approach with the starget website. There is proposed a novel fusion classifier with two priority algorithms (PAZ) resulting in phishing website detection with data and PAZ) resulting in phishing website detection with data and detection approach and approach and they reached over 90 levels of accuracy in identifying applied, and they reached over 90 levels of accuracy in identifying genuine and malicious websites. Stuffgate, Alexa, and phish Tank. Phish Tank. Phish Tank. Phish Tank achieving an accuracy of 89.38% when evaluated vealuated v									and neural	in real-world
Performance evaluation using data from sources such as Stuffgate, Alexa and PhishTank achieved an accuracy of 89.38% which proved the system's capability of identifying phishing. 3. Phishing website detection using approach approach significantly improves a marno, P. Surva, M. AdaBoost, SVM, and random fusion approach approach a register with two priority algorithms (PA1) and PA2) resulting in phishing website of addressing phishing the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1) and PA2) resulting in phishing website detection with data from sources such as Stuffgate, achieving an accuracy of 89.38% when evaluated with data from sources like and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and malicious websites. This model machine learning approach accuracy of significantly improves phishing and random forests have been applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1) and PA2) resulting in phishing website detection in addressing phishing the feature such as accuracy of fering significant indentifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and they reached over 90 levels of accuracy in identifying genuine and				codes of the	website				were the four	applications.
evaluation using data from sources such as Stuffgate, Alexa and accuracy of PhishTank achieved an accuracy of 89.38% when achieved an accuracy of identifying proved the system's capability of identifying phishing. 3. Phishing website detection approach learning fusion approach approach approach approach learning fusion approach approach approach learning fusion approach approach and anovel fusion classifier with two priority algorithms (PAI and PA2) resulting in phishing website detection with applishing website and and PA2) resulting in phishing website detection with accuracy of accuracy of accuracy of accuracy in identifying genuine and malicious websites. There is proposed a novel fusion classifier with two priority algorithms (PAI and PA2) resulting in phishing website detection with data from sources like susting an accuracy of phishing website such as theiroing an accuracy of accuracy of accuracy in identifying genuine and malicious websites. This model over 90 levels of accuracy in identifying genuine and malicious websites. That is preached over 90 levels of accuracy in identifying genuine and malicious websites. There is proposed a nore fusion classifier with two priority algorithms (PAI and PA2) resulting in phishing website detection with data from sources like surface, Alexa, and identifying genuine and malicious websites. That is model classifier accuracy of accuracy				Performance	through					
such as Stuffgate, Alexa and PhishTank achieved an accuracy of 89.38% when achieved an accuracy of 89.38% when achieved an accuracy of 89.38% when achieved an accuracy of system's capability of identifying PhishTank approach actection using Surya. M novel machine learning fusion approach anovel factures such approach approach anovel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with achieved an accuracy of sy 389.38% when evaluated with data from sources like system's Alexa and phishfank achieved an accuracy of sy 389.38% when evaluated with data from sources like system's Alexa, and identifying genuine and malicious websites. This model integrates two approach solarity accuracy printip address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with accuracy of accuracy in identifying genuine and malicious websites. S. A novel machine learning papproach to detect phishing websites S. Gaur Gigandee phishing websites This model integrates two approaches corracy, offering significant indentifying genuine and malicious websites. Forearity adaction accuracy in identifying genuine and malicious websites. S. A novel machine learning papproach to detect phishing websites This model integrates two approaches corracy phishing websites This model integrates two approaches with Random Forest This model integrates two approaches with Random with Random with Random with Random Forest Machine learning models and accuracy proaches. Scauracy ormachine learning models and accuracy proaches. Scauracy ormachine learning models and accuracy phishing websites This model integrates two approaches websites.									they reached	
PhishTank achieved an accuracy of 89.38% which grows with data from sources like system's capability of identifying phishing. 3. Phishing website detection using novel machine learning fusion approach approach approach A. Machine learning fusion approach A. Machine learning fusion approach Bala forests have been applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with achieved an accuracy of with data from sources like system's Stuffgate, capability of identifying genuine and malicious websites. This model machine learning approach significantly improves phishing websites 5. A novel machine learning acturacy identifying genuine and malicious websites. This model integrated beneath to detect phishing websites Sharma: Shard; S. Gaur Gidl, Model S. Gau				such as Stuffgate,	achieving an				levels of	
achieved an accuracy of with data from sources like system's capability of identifying phishing. 3. Phishing website detection using approach appr				PhishTank	89.38% when					
89.38% which proved the system's capability of identifying phishing. 3. Phishing website detection using Surya. M novel machine learning fusion approach Ababoost, SVM, and random forests have been applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing website detection with algorithms (PA1 and PA2) resulting in phishing detection with algorithms (PA1 and PA2) resulting in phishing detection algorithms (PA1 and PA2) resulting in phishing detection algorithms (PA1 and PA2) resulting in phishing algorithms (PA1 and PA2) resulting in phishing detection algorithms (PA1 and PA2) res									genuine and	
3. Phishing website detection using approach approach approach approach approach I al akshma detection using fusion approach approach approach approach approach approach approach approach are including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with the detection with the detection with the detection in addressing phishing website detection with the detection with the detection in addressing phishing website and the detection with the detection in addressing phishing the detection with the detection with the detection in addressing phishing website and the detection with the detection with the detection in addressing phishing website and the detection with the detection with the detection with the detection in addressing phishing website and the detection phishing website and the detection in addressing phishing website and the detection in addressing phishing website and the detection in addressing phishing website and the detection with the detection in addressing phishing website and the detection with the detection in addressing phishing website and the detection with the detection with the detection in addressing phishing website and the detection with the detection and decision trees, alara, and it algorithms (PA1 and brace) and addressing phishing website and the decision trees, alara, and addressing approach to detect machine learning approach to detect to detect on addressing approach to detect to detect on addressing approach to detect to				89.38% which						
3. Phishing website detection using narao, P. using narao, P. using fusion approach approach A. Machine learning methods such as decision trees, AdaBoost, SVM, novel machine learning fusion approach approach AdaBoost, SVM, AdaBoost, SVM, improves phishing detection applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with A. Machine learning methods such as decision trees, AdaBoost, SVM, improves phishing detection approach significantly improves phishing detection accuracy compared to earlier methods, showcasing the effectiveness of advanced machine learning models and feature selection in addressing phishing website detection with				system's	Stuffgate,					
3. Phishing website detection using novel machine learning fusion approach ABala Krishna Machine learning methods such as decision trees, significantly improves phishing detection accuracy compared to earlier methods, showcasing the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with Machine learning methods such as decision trees, significantly improves phishing detection accuracy compared to earlier methods, showcasing the effectiveness of advanced machine learning models and feature selection in phishing website detection with Machine learning methods such as decision trees, significantly improves phishing detection accuracy compared to earlier methods, showcasing the effectiveness of advanced machine learning models and feature selection in phishing website detection with Machine learning methods such as decision trees, significantly improves phishing detection accuracy compared to earlier methods, showcasing the effectiveness of advanced machine learning methods significantly improves phishing detection accuracy compared to detect phishing websites S. Gaur Gagandee phishing websites Model Linear Sharma: Gagandee phishing websites With Random Forest Model Cinear Sharma: Gagandee phishing websites Elearning methods, showcasing the effectiveness of advanced machine learning methods, significantly improves approach significantly improves phishing detection accuracy compared to detection accuracy compared to detect phishing websites Model Cinear Sharta. Sharma: S. Gaur Gagandee phishing websites Forest										
website detection using novel machine learning fusion approach approach There is proposed a novel target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with the detection with significantly improves phishing the phishing the phishing the detection agaptroach significantly improves phishing the phishing the detection agaptroach significantly improves phishing the detection agaptroach significantly improves phishing the detection accuracy compared to earlier machine learning approach significantly improves phishing the detection accuracy compared to earlier machine learning approach to detect the machine learning approach significantly improves the surfact the detection accuracy compared to earlier machine learning approach to detect the machine learning approach significant to detect the mac	3	Phiching	Λ		This					
using novel machine learning fusion approach Surya, M Bala AdaBoost, SVM, and random forests have been applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with Surya, M Bala AdaBoost, SVM, and random forests have been applied in this research including accuracy compared to earlier methods, showcasing the earlier methods, showcasing the effectiveness of advanced machine learning models and feature selection in addressing phishing website detection with Shad; S. Gaur (GLM) approach (GLM) accuracy, offering significant improvements in detecting phishing websites Sharma; S. Gaur (GLM) accuracy, offering significant improvements in detecting phishing websites Sharma; S. Gaur (GLM) accuracy, offering significant improvements in detecting phishing websites Sharma; S. Gaur (GLM) accuracy, offering significant improvements in detection in addressing phishing threats.	<i>J</i> .	website	<u>Lakshma</u>	methods such as	approach					
machine learning fusion approach Tyagi; J. Shad; S. Sharma; S			l							
learning fusion approach applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority and PA2) resulting in phishing website detection with applied in this research including compared to earlier methods, showcasing the learning approach to detect phishing websites There is proposed a novel fusion machine learning approach to detect phishing websites There is proposed a novel fusion and particular improvements in detecting phishing websites. There is proposed a novel fusion and particular improvements in detecting phishing websites.						5.	A novel			
the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with threats.		learning	rensina	applied in this	accuracy					
as web traine, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority and PA2) resulting in phishing website and PA2) resulting in phishing website phishing phishing phishing phishing phishing websites phishing websites phishing websites phishing websites phishing websites with Random Forest with Random Forest improvements in detecting phishing websites.					earlier		approach	S. Gaur	(GLM)	achieving 98.4%
address of the target website. There is proposed a novel fusion classifier with two priority models and algorithms (PA1 and PA2) selection in phishing website phishing website detection with threats. There is proposed of advanced machine learning models and algorithms (PA1 feature and PA2) selection in phishing website detection with threats.					· · · · · · · · · · · · · · · · · · ·		phishing		with Random	significant
There is proposed a novel fusion classifier with two priority models and algorithms (PA1 feature and PA2) selection in resulting in phishing website phishing detection with threats.				address of the	the		websites		Forest	
classifier with two priority models and algorithms (PA1 feature and PA2) selection in resulting in addressing phishing website phishing detection with threats.				There is proposed	of advanced					
priority models and algorithms (PA1 feature and PA2) selection in resulting in addressing phishing website phishing detection with threats.										
and PA2) selection in resulting in addressing phishing website phishing detection with threats.				priority	models and					
phishing website phishing detection with threats.				and PA2)	selection in					
detection with threats.										
				detection with						
				2770 accuracy.						

II. METHODOLOGY

The methodology of creating a phishing website detection system includes several basic steps. This section describes the survey design, data collection methods, and analytical methods for effective implementation of the system. Emphasis is placed on using machine learning algorithms to analyze a variety of resources, including URL structures and domain names. Through systematic training and analysis of multiple models, the method aims to develop a robust tool that can accurately identify phishing websites in real-time, and for providing cybersecurity measures users and organizations have improved.

• RESEARCH DESIGN:-

The goal of this project is to design a machine learning model for classifying a website as phishing or not, by performing an analysis of specific attributes such as the URL, the age of the domain, and possession of security certificates. The scope of the research will deal with data collecting, for example from PhishTank and Kaggle, feature extraction and deployment of several machine learning algorithms: Logistic Regression, Decision Trees, Random Forest, etc. Each of the model's performances will be addressed in terms of accuracy, precision, recall, and f-measure score. The proposed outcome is a proficient solid tool that classifies legitimate and phishing sites and improves overtime through continuous learning.

• DATA COLLECTION:-

Phishing Website Data: Gather datasets from publicly available sources such as PhishTank, UCI repository, or Kaggle, which contain records of both phishing and legitimate websites.Legitimate Website Data: Collect data from trustworthy sources for a balanced comparison with phishing websites.

DATA SETS:- A set comprising phishing URLs is retrieved from the PhishTank open-source service. The service provides an hourly-updated set of phishing URLs with multiple formats such as 'csv' and 'json'. The data can be downloaded from: https://www.phishtank.com/developer info.php. Out of which, 5000 phishing URLs are sampled randomly from the phishing database for training the ML models. The legitimate URLs are reaped from the open datasets provided by the University of New Brunswick, available at https://www.unb.ca/cic/datasets/url-2016.html. In contrast with the other mentioned types, the benign URL dataset would be used for this project. From that dataset, 5000 legitimate URLs will be randomly considered for training the ML models.

• FEATURE EXTRACTION:-

For address-based features based on the address bar, we analyse the domain of the URL, use of redirection syntax (for instance, "//"), presence of an IP address, http and https within the domain name, use of '@' sign, usage of URL-shortening services, URL length, and presence of "-" as a prefix or suffix to the domain. Domain-based features include DNS record, traffic of the website, the age of the domain name, and the end period of the domain.

Finally, HTML and JavaScript-based features consider interface redirection with iframes, status bar manipulation, blocking entries by right mouse clicking, and forwarding of a page on the web.

• TECHNOLOGIES USED:-

Detecting phishing websites using machine learning algorithms typically involves using supervised learning techniques. Some common machine learning algorithms and methods used for phishing website detection include:

- Logistic Regression: This algorithm is commonly used for binary classification tasks like phishing detection, where the output is either phishing or legitimate.
- Decision Trees: Decision trees can be effective for phishing detection as they can capture complex relationships between features. Ensemble methods like Random Forests and Gradient Boosting Machines (GBM) can also be used for improved performance.
- 3. Support Vector Machines (SVM): SVMs are good for binary classification tasks and can be trained to distinguish between phishing and legitimate websites based on various features.
- 4. K-Nearest Neighborss (KNN): The K-Nearest Neighbour is an easy-to-use and straightforward method which can also be utilized when it comes to identifying phishing websites. This is achieved using web features similarities.
- 5. Random Forest: Random Forest (RF) is an algorithm in machine learning whereby several decisional trees' outcomes are merged to get an answer.
- EVALUATION:- The evaluation of the Phishing Website Detection system measures its effectiveness in distinguishing phishing and non-phishing sites. Key performance metrics such as accuracy, precision, recall, and F1-score are used to assess the model. By comparing predictions against known data, the system's ability to anticipate phishing attacks is validated.

Results indicate that the system performs well in identifying phishing threats, contributing to enhanced online security for individuals and businesses. The methodology follows a systematic approach, incorporating well-structured research design, data collection, and machine learning-based analysis. The model is trained to recognize patterns in phishing websites, improving detection accuracy over time.

The evaluation results demonstrate that the system effectively differentiates between legitimate and fraudulent sites. By continuously refining its detection capabilities, the model strengthens cybersecurity efforts, making the internet a safer space for users.

III.ANALYSIS AND MODEL TESTING

Phishing website detection leverages advanced machine learning and data science techniques to accurately identify fraudulent websites. The process begins by establishing a comprehensive model based on characteristics of both legitimate and phishing websites, such as URL patterns, domain age, and website content.

The model is trained using vast datasets of phishing and nonphishing websites, enabling it to identify common patterns and anomalies indicative of phishing attempts. The model undergoes rigorous testing to assess its accuracy, precision, and recall in detecting phishing threats.

Continuous data integration during training strengthens the model, allowing it to adapt to new phishing tactics. As phishing techniques evolve, this iterative approach ensures that the system issues increasingly accurate alerts. Ultimately, modeling and analysis play a critical role in maintaining an effective phishing detection system, making it a key tool in combating online threats.

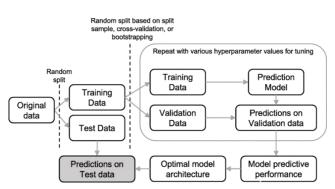


Fig.3.1:- Model analysis of proposed system

IV. WORKFLOW OF SYSTEM

The workflow of the phishing website detection system using machine learning can be described as follows:

- 1. Data Collection: The system begins by gathering a dataset of URLs, both legitimate and phishing, from sources like PhishTank, Alexa, and other repositories.
- 2. Feature Extraction: Key features from the URLs are extracted, including domain information, URL length, HTTPS usage, the presence of suspicious characters (like '@', '//'), and ranking factors such as Alexa Rank and PageRank.
- 3. Preprocessing: Data is cleaned and pre-processed for machine learning models, ensuring consistency, and removing unnecessary noise.
- 4. Model Selection and Training: Machine learning methodologies such as Random Forest, SVMs, or GLM were explored. The model was trained with the extracted features to distinguish between legitimate and phishing URLs.
- 5. Model Testing and Evaluation: The other methods which applicable include machine learning methods: Random Forest, Support Vector Machine (SVM), Generalized Linear model (GLM). The model then trained on the features extracted to categorize the URLs as either legitimate or phishing.6. Classification: Based on the trained model, the system predicts whether a given URL is a phishing site or legitimate.
- 7. Output: The system provides the result (legitimate or phishing) along with confidence scores for the prediction, aiding in real-time detection. This workflow ensures a comprehensive approach to phishing website detection by leveraging key URL features and machine learning for high accuracy and performance.

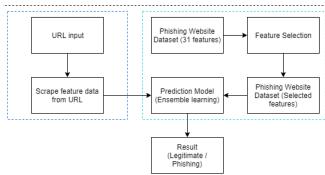


Fig.4.1:- Workflow of the system

V. RESULT AND CONCLUSION

identifies potential phishing threats.



Fig. 5.1:- System User Interface (Legitimate)



Fig. 5.2:- System User Interface (Phishing)

security. The results from model testing demonstrate its high work. accuracy in distinguishing phishing websites, providing Lastly, we appreciate everyone who encouraged and assisted detection techniques.

In conclusion, the system effectively mitigates phishing risks by leveraging sophisticated machine learning and analytical approaches.

Algorith m	Accur acy (%)	Prec ision (%)	all	F1- Scor e (%)	Remarks
Decision Tree	85.2	83.5	84.1	83.8	Easy to interpret but prone to overfitting
Support Vector Machin e (SVM)	89.7	88.2	87.5	87.8	High accuracy but computationa lly expensive
Random Forest	92.5	91.0	90.8	90.9	Best overall performance with strong generalizatio n

Table. 5.1:- Classifier's Performance

V.i. CONCLUSION

The phishing website detection system uses machine learning After comparing different machine learning models for and analytical models to assess and classify websites based phishing website detection, Random Forest clearly stands out on several parameters. Through advanced analysis of URL as the best performer. With an accuracy of 92.5%, it not only structures, domain attributes, and HTML features, the system delivers strong results but also generalizes well, making it a reliable choice for real-world applications.

> Support Vector Machine (SVM) also performs well with 89.7% accuracy, but its high computational cost can make it challenging to implement in large-scale systems. On the other hand, Decision Tree is easy to understand and implement, but with an accuracy of 85.2%, it falls short compared to the other two models and is more prone to overfitting.

> Overall, Random Forest is the most effective choice for detecting phishing websites due to its balance of accuracy, precision, and recall. Moving forward, further improvements can focus on refining feature selection and experimenting with deep learning techniques to push detection accuracy even higher.

VI. ACKNOWLEDGEMENT

We sincerely thank our mentors for their guidance and support in making Anonymity and Confidentiality in Website using ML system a reality. Our gratitude also goes to the Greater Noida Institute of Technology for providing the resources that helped us complete this project.

Special thanks to our project guide, Mr. Umashanker By utilizing an ensemble of machine learning algorithms, the **Sharma**, for his constant support and valuable insights into system provides accurate detection results, enhancing online machine learning and NLP, which played a key role in our

insights for refining security protocols and improving future us throughout this journey, your support made all the difference.

VII. REFERENCES

- [1] Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh, "Phishing Website Detection Using Machine Learning: A Review" June 2023Wasit Journal of Pure sciences 2(2):270-2812(2):270-281.
- [2] A K Jain & B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," EURASIP Journal on Information Security, vol. 2016, no. 1, pp. 1-11, 2016.
- [3] A. Lakshmanarao, P. Surya, M Bala Krishna, "Phishing website detection using novel machine learning fusion approach," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021: IEEE, pp. 1164-1169.
- [4] A. Kulkarni & L. Brown, "Phishing websites detection using machine learning," 2019.
- [5] I. Tyagi, J. Shad, S. Sharma, S. Gaur, and G. Kaur, "A novel machine learning approach to detect phishing websites," in 2018 5th International conference on signal pro-cessing and integrated networks (SPIN), 2018: IEEE, pp. 425-430.