“Phishing Detection System Through Hybrid Machine Learning Based on URL”

Submitted for the degree of

B. Tech

In

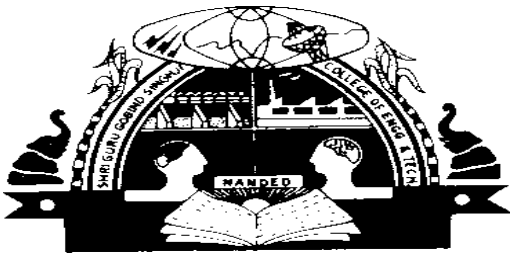
Computer Science and Engineering

By

Beedkar Chinmay Sanjay
Chaware Mahesh Shrihari
Galpelli Yash Chandrabhan
Khanzode Priyanshu Manoj

Under the Guidance of

Mrs. Jaishree Waghmare



SHRE GURU GOBIND SINGHAJI INSTITUTE OFSENGINEERING & TECHNOLOGY, NANDED (M.S.)

ACADEMIC YEAR

(2023-24)

CERTIFICATE

This is to certify that the project/ research work entitled "**Phishing Detection System Through Hybrid Machine Learning Based on URL**" being submitted by the group of ***Mr. Chinmay Beedkar (2020BCS052), Mr. Mahesh Chavare (2020BCS066), Mr. Yash Galpelli (2020BCS075) and Mr. Priyanshu Khanzode (2020BCS063).*** As a partial fulfillment of Cryptography and Network Security for the academic year 2023 – 2024. This project is a record of students work carried out by them under the supervision of Asst. Prof. Nisha H. Songire and guidance of Prof. Jayshree Waghmare

Dr. Jayshree Waghmare Project Guide	Dr. Suvarna Bansode Head Department of Computer Science	Dr. M. Kokre SGGSIE&T, Nanded

ABSTRACT

Currently, numerous types of cybercrime are organized through the internet. Hence, this study mainly focuses on phishing attacks. Although phishing was first used in 1996, it has become the most severe and dangerous cybercrime on the internet. Phishing utilizes email distortion as its underlying mechanism for tricky correspondences, followed by mock sites, to obtain the required data from people in question. Different studies have presented their work on the precaution, identification, and knowledge of phishing attacks; however, there is currently no complete and proper solution for frustrating them. Therefore, machine learning plays a vital role in defending against cybercrimes involving phishing attacks. The proposed study is based on the phishing URL-based dataset extracted from the famous dataset repository, which consists of phishing and legitimate URL attributes collected from 11000+ website datasets in vector form. After preprocessing, many machine learning algorithms have been applied and designed to prevent phishing URLs and provide protection to the user. This study uses machine learning models such as decision tree (DT), linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbours classifier (KNN), support vector classifier (SVC), and proposed hybrid LSD model, which is a combination of logistic regression, support vector machine, and decision tree (LR+SVC+DT) with soft and hard voting, to defend against phishing attacks with high accuracy and efficiency. The canopy feature selection technique with cross fold validation and Grid Search Hyperparameter Optimization techniques are used with proposed LSD model. Furthermore, to evaluate the proposed approach, different evaluation parameters were adopted, such as the precision, accuracy, recall, F1-score, and specificity, to illustrate the effects and efficiency of the
models. The results of the comparative analyses demonstrate that the proposed approach outperforms the other models and achieves the best results.

ACKNOWLEDGEMENT

It is a privilege for me to have been associated with **Mrs. Jaishri Waghmare Ma'am**, my guides during this project work. I have greatly benefited by her valuable suggestions and ideas. It is with great pleasure that I express my deep sense of gratitude to them for their able guidance, constant encouragement, and patience throughout the work.

I am also thankful to **Dr. Manesh Kokre Sir**, Director and **Mrs. Suvarna Bansode Ma'am**, Head of Computer Science and Engineering Department for their constant encouragement & cooperation.

I am also thankful to Asst. Prof. **Mrs. Nisha Songire Ma'am** for her kind help during this work. She has been a constant support during this project.

I take this opportunity to thank Chinmay, Yash, Mahesh for providing company during the work.

K. Priyanshu Manoj

Table of Contents

List of Tables

List of Figures

Chapter No. 1
Introduction

INTRODUCTION

The internet plays a crucial role in various aspects of human life. The Internet is a collection of computers connected through telecommunication links such as phone lines, fiber optic lines, and wireless and satellite connections. It is a global computer network. The internet is used to obtain information stored on computers, which are known as hosts and servers. For communication purposes, they used a protocol called Internet protocol/transmission control protocol (IP-TCP). The government is not recognized as an owner of the Internet; many organizations, research agencies, and universities participate in managing the Internet. This has led to many convenient experiences in our lives regarding entertainment, education, banking, industry, online freelancing, social media, medicine, and many other fields in daily life. The internet provides many advantages in different fields of life. In the field of information search, the Internet has become a perfect opportunity to search for data for educational and research purposes. Email is a messaging source in fast way on the Internet through which we can send files, videos, pictures, and any applications, or write a letter to another person around the world. E-commerce is also used on the internet. People can conduct business and financial deals with customers worldwide through e-commerce. Online results are helpful in displaying results online and have become a more useful source of the covid-19 pandemic in 2020. Many classes and business meetings are performed online, which requires time and is fulfilled through the internet.

Owing to the increase in data sharing, the chances of loss and cyber-attack also increase. Online shopping is the biggest Internet use that helps traders sell projects online worldwide. Amazon operates a large online sales system. Fast communication is performed through the Internet, which is currently used through Facebook, Instagram, WhatsApp, and other social networks, making communication fast and easily available. Therefore, it is necessary to maintain a privacy policy in which communication and its users cannot be defective. The Internet provides a great opportunity for attackers to engage in criminal activities such as online fraud, malicious software, computer viruses, ransomware, worms, intellectual property rights, denial of service attacks, money laundering, vandalism, electronic terrorism, and extortion. Hacking is a major destroyer of the Internet through which any person can hack computer information and use it in different ways to harm others. Immorality, which harms moral values, is a major issue for the younger generations. Detecting these websites rather than websites that appear simple and secure, will help people. Therefore, an awareness of these websites is necessary. Viruses can damage an entire computer network and confidential information by spreading to multiple computers. It is not suitable to use unauthorized websites on the internet. Phishing detection is required for all of these aspects to secure our computer system. Cyber security has become a major global issue. Over the last decade, several anti-phishing detection mechanisms have been proposed. These studies have mainly focused on the structure of a uniform resource locator (URL) based on feature-selection methods for machine learning.

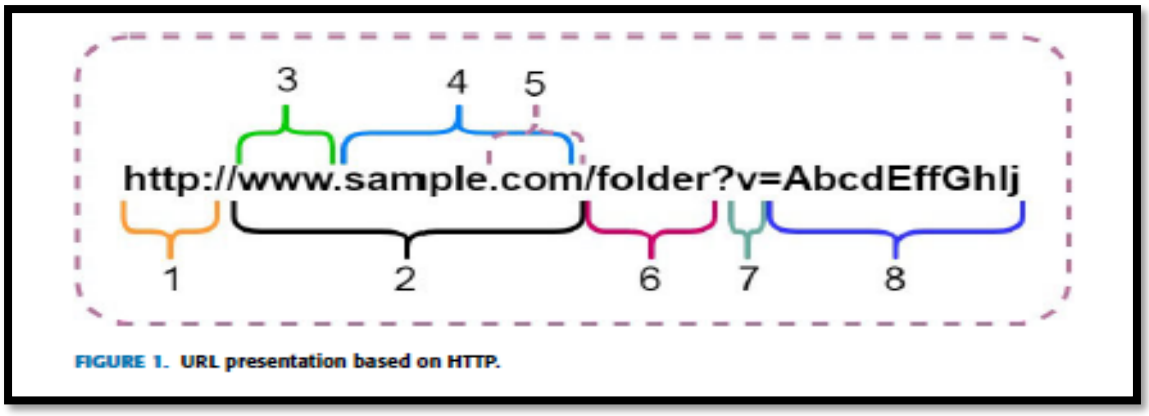


FIGURE 1. URL presentation based on HTTP.

Berners-Lee (1994) developed the URL. The format of the URL is defined by preexisting sources and protocols. Pre-existing systems, such as domain names with syntax of file paths, were created and proposed in 1985. Slashes were used to separate the filenames and directories from the path of a file. Double slashes were used to separate the server names and file paths. Berners-Lee then introduced dots to separate the domain names. HTTP URL consists of a syntax which is divided into five components which are in hierarchical sequence.

A uniform resource locator (URL) are the most significant category of uniform resource identifiers (URI). URI is characteristic strings used over networks to detect resources. Navigation of Internet URLS is important. The URL comprises a component of a non-empty scheme that is followed through the colon (:). It consists of a sequence of characters that begin with a letter and follow any combination of letters, digits, plus, hyphen, or minus. These schemes are case sensitive. Some of these schemes include ftp, data, file, HTTP, HTTPS, and IRC, which are registered by the Internet assigned numbers authority (ANA). Otherwise, in practice, mostly non-registered schemes are used.

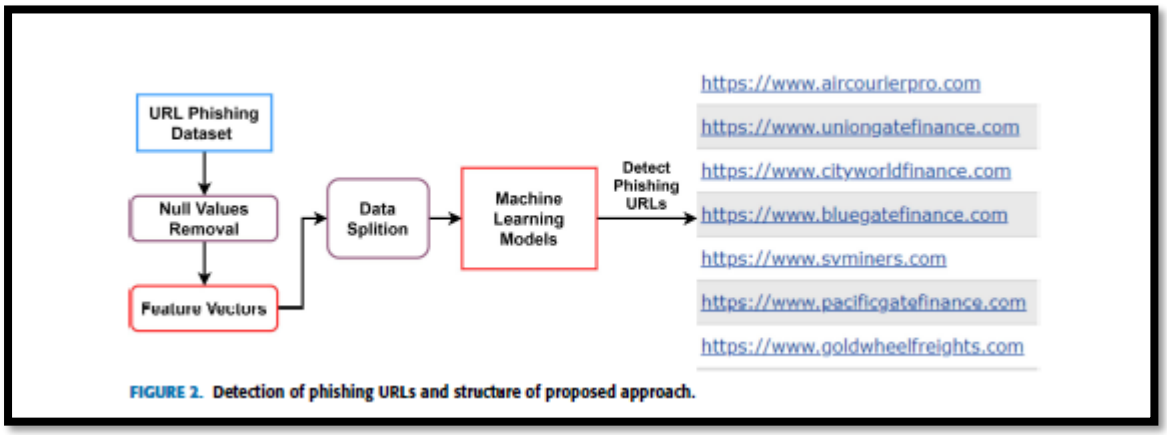


FIGURE 2. Detection of phishing URLs and structure of proposed approach.

HTTP or HTTPS Both are used in the process of data retrieval from the web server to view content in a browser HTTPS uses Secure Sockets Layer (SSL) which used to encrypt the connection between the server and end user. HTTPS used to vital the personal information such as passwords, Identification of data come from unauthorized and illegal access, and credit card numbers. HTTPS and HTTP used port numbers of TCP/IP as 433 and 80 Currently, numerous types of cybercrime are organized through the internet. Hence, this study mainly focuses on phishing attacks. Phishing is a type of cybercrime in which

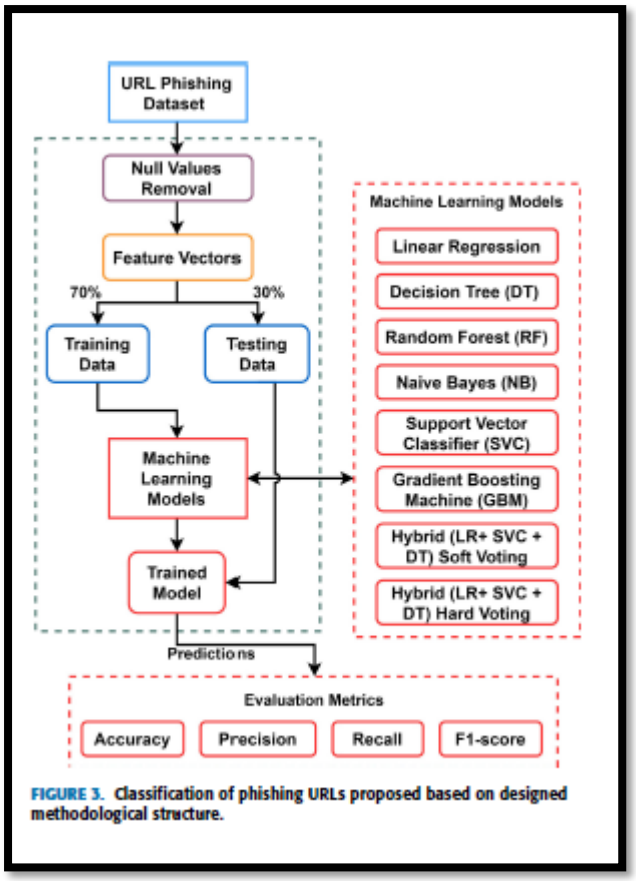


FIGURE 3. Classification of phishing URLs proposed based on designed methodological structure.

subjects are baited or fooled into surrendering delicate data; for example, social security numbers individually recognizable data and passwords. The acquisition of such data was performed deceitfully. Given that phishing is an exceptionally broad theme, this study ought to focus explicitly on phishing sites. This study divided a simple phishing attack into four types. First, it creates a phished website that resembles a legitimate site.

Second, they would send the uniform asset locator (URL) connection of the website for legitimate use by feigning it to be an authentic organization or association. Third, the individual endeavors to persuade the loss to visit a fraudulent website. Fourth, trustful casualties tap into the connection between counterfeit sites and acquire useful information. Finally, by utilizing the individual data of the person in question, the phisher will use the data to perform extortion exercises. Nonetheless, phishing assaults are not performed expertly to maintain strategic distance from clients or casualties. Phishing is a security risk to many people, particularly those who do not know about threats to online websites. FBI gives a report, lowest loss of 2.5 billion had become effected by phishing frauds between the periods of October 2013 to February 2016. Most people do not check or think about websites' URLs on their computer screens. Sometimes, phishing frauds become phishing websites, which can be discouraged by penetrating whether a URL belongs to a phishing or a legitimate website. Recently, several phishing attacks have been reported worldwide. A phishing attack is the scam of phishing in PayPal services for the user's login details. It arises from a normal email that contains phishing content, but the victims have lost control and access to personal or financial management, in extension to their login credentials. At the same time, another phishing attack came into being one of Australia's largest IVF providers hit by phishing scams. In this attack, attackers obtain the main information of the patient's name, details of the contact, date of birth, cast designation, financial information, information on medical insurance, driving license number, and the number of passports. Private information from the faculty of the Singapore Ministry of Defense was leaked after the employee received a bogus email containing a malicious file. An employee opens an email with bogus content and gives attackers access to a host of personal information. As a result of this attack, 2400 employees were exposed, including their NRIC (National Registration Identity Card) number, names, contact details, and addresses. Several systems and mechanisms have been designed for detecting phishing attacks. However, accurate results have not been obtained. The main purpose of this research is to create a phishing website detection system that performs better than previously designed mechanisms to enhance security and accuracy and obtain better results to avoid any loss. The web tool

PHISHTANK was proposed to detect phishing attacks. PHISHTANK is based on different features that determine whether a website is secure or malicious or not. A URL structure is defined to detect a phishing attack using the URL. In the proposed study, machine learning algorithms were used with the features of the URL to solve classification problems. Effective features for training purposes were selected based on an effective phishing detection mechanism

–

Chapter No. 2:
Literature Review

Literature Review

Phishing is the most significant issue in the field of networks and the Internet. Many researchers have attempted to provide facilities to protect users from cyber-attacks by preventing the phishing of URLs using machine learning, deep learning, black lists, and white lists. Two groups of phishing detection systems have been proposed and implemented in previous studies: list-based and machine-learning-based phishing identification systems. This section is divided into two parts: previous list-based and machine-learning-based studies. A. LIST BASED PHISHING IDENTIFICATION SYSTEM Phishing identification systems based on List use two different lists white lists and blacklists for the association and classification of authorized and phishing webpages. Whitlistbased Phishing identification systems produce protected and reliable websites to

S.No	Paper Tile	Author	Dataset Used	Dataset URL	Features Used	Performance metrics
1	Phishing Detection Using Machine Learning Techniques	author -> Vahid Shahrivari , Mohammad Mahdi Darabi, Mohammad Izadi	Phishing Dataset	https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning?resource=download	URL-based features: Domain, subdomain, path, length, entropy. Text-based features: Lexical analysis of email content.	Detection Rate
						False Positive Rate
						False Negative Rate
						Accuracy
2	An intelligent cyber security phishing detection system using deep learning techniques	Ping Yi , Yuxiang Guan,Futai Zou,Yao Yao,Wei Wang,Ting Zhu	Phishing Dataset	https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning?resource=download	original feature and interaction feature.	1) positive predictive value
						2) False Positive Rate
						3) False Negative Rate

produce the required data. A suspicious website just needs to match the website of the whitelists; if it is not in the whitelist, it means it is suspicious and threatened by the user. In [20]. To develop a whitelist-based system that generates a whitelist by monitoring and recording the IP address of every website that contains the login interface for the end-user used by the users to enter their details. When the user uses this login interface, the Windows 2008 system displays a warning for the incompatibility of registered information details. This is why this system mechanism suspects legitimate sites visited by users for the first time. Reference [21] developed a system that alerts users about a phishing website by periodically and automatically maintaining and updating the whitelist.

The performance of this system depends on two factors: the extraction of attributes hidden in the link between the source code and the module that matches the IP address of the domain. According to the preliminary conclusions, 86.02 the true positive rate was 1.48% falsenegative score was this study. Blacklists were collected based on the records of URLs known as phishing websites. Numerous sources, such as user notifications, detection of

spam systems, and third-party authorities, are used to collect record entries for list creation. The blacklist makes it possible for systems to prevent attackers from recording their IP addresses and URLs. Therefore, next time the attackers must use a new URL or IP address because the blacklist-based system detects their previous URLs or IPs. System security management can automatically update the blacklist periodically to prevent new attackers by identifying malicious URLs or IPs. Alternatively, users can download these lists to update their security system. Zeroday attacks mostly affect systems because blacklist-based systems are not able to detect a new or first-day attack. These intrusion detection systems exhibit a lower false-positive score than systems based on machine learning. The accuracy of the detection of intrusions or attacks of these systems based on the blacklist is very high, and with success rate of approximately 20%, according to [22] and [23]. Consequently, this shows that the identification systems of some companies based on blacklist mechanisms, such as Phish- Net [24] and Google Safe Browsing API [25], are reliable for detecting phishing attacks based on blacklists. Approximate matching algorithms are used by these security systems to match malicious URLs with URLs present in the blacklist. Frequent updates are required for blacklists that use these systems. In addition, the accelerated

Novelty	Research Gaps	Strengths	Weakness	limitations	REFERENCE (this paper) in IEEE fomat
Creating effective hybrid models that leverage the strengths of different approaches is a promising area for innovation.	2 years	1) Real-world Applicability	1) Computational Complexity	1) Dataset Limitations	1) FBI, "Ic3 annual report released."
		2) Large and Diverse Dataset	2) Overfitting	2) Limited Attack Patterns	2) A. K. Jain and B. B. Gupta, "Phishing detection: Analysis of visual similarity based approaches,"
			3) User Experience Consideration		
Combine information from multiple sources such as URLs, text content, and user behavior.	2 years 2 month	1) Feature Extraction	1) Data Scarcity and Quality	1) Data Privacy	1) https://en.wikipedia.org/wiki/Web service
		2) Learning from Data	2) Imbalanced Data:	2) Overfitting	2) H.-C. Huang, Z.-K. Zhang, H.-W. Cheng, and S. W. Shieh, "Web application security: Treats, countermeasures, and pitfalls," Te Computer Journal, vol. 50, no. 6, pp. 81–85, 2017.

increase in blacklists demands extravagant system support [26], [27]. This study [6] uses a browser extension approach for phishing and URL detection and has an 85% accuracy rate; however, in recently, several automatic phishing detection mechanisms have been proposed [7]. This study used shortened URL features for the detection process, has 92% accuracy. Delta Phish [8] is a phishing-detection mechanism. It uses several URL features to train supervised predictive models, and its accuracy rate is higher than 70%. This study [9] proposes a Phish-Safe detection mechanism to detect malicious websites. This study used SVM and naive Bayes as a supervised-based machine learning approaches for phishing detection and achieved 90% accuracy. In this study, [10] ensemble learning technique was used for phishing attack detection in the emails. There are replaced feature selection techniques that are used to move such features that are not associated with accuracy and achieve 99% accuracy using only 11 features. In another study [11], The Phi DMA approach was used in another study. This approach used five-layers URL feature layers, lexical layer, whitelist layer, and achieved an accuracy of 92%. In another study [12], the investigation of phishing was detected through SVM. In this study, six features were

Classification method	Algorithm used	Tools/platform used	Research findings	Results	Conclusion
1) K Near Neighbors	K Near Neighbors	PyTorch	A. P. Rosiello, E. Kirda, F. Ferrandi, et al., "A layout-similarity-based approach for detecting phishing pages," in 2007	Good result for ensembling classifiers namely, Random Forest, XGBoost both on computation duration and accuracy.	ensembling classifiers namely, Random Forest, XGBoost both on computation duration and accuracy.
2) logistic regression					
3) neural network					
4) ada booster					
5) random forest					
6) decision tree					
1) Convolutional Neural Networks (CNNs)	K step CD k	TensorFlow	we introduce DBN to detect phishing websites and discuss the detection model and algorithm for DBN. We train DBN and get the appropriate parameters for detection in the small data set. In the end, we use the big data set to test DBN and TPR is approximately 90%.	Researchers have outperform traditional rule-based or heuristic-based systems by leveraging the power of deep learning to identify subtle and evolving phishing patterns.	In this paper, we analyze the features of phishing websites and present two types of feature for web phishing detection: original feature and interaction feature
2) Fully Connected Networks					

obtained from the domain address, and the empirical results showed an accuracy of 95%. Another study [13] developed a phishing detection system using a typo squatting and phoneme-based approaches. Using these techniques, an accuracy of 99% is achieved. B. MACHINE LEARNING BASED IDENTIFICATION SYSTEM Machine learning is the most popular technique for identifying malicious and suspicious websites by using URLs. Classification of phishing URLs is an important domain in machine learning. A large number of data features are required to acquire machine-learning-based security systems and to train the model on features that are associated with legitimate and phishing website labels. The outstanding performance of machine learning algorithms allows them to easily detect hidden or first-time attacks that are not on a blacklist. The authors [28] developed a phishing detection system based on text classification named CANTINA. This technique extracts features as keywords using a feature extraction technique known as term frequency inverse document frequency (TFIDF). These extracted keywords were used to search the Google search engine, and if any of these websites were found, they were classified as legitimate websites. However, the achievements of this study are restricted because they are particularly sensitive to English vocabulary. Subsequently, another enhanced approach was proposed by [29], which was based on the attributes of 15 different HTMLs, named CANTINA+. The highest accuracy of 92% was achieved by this system, which produced a tremendous number of false-positive predictions. Reference [30] developed an anti-phishing-based security system called Phish- WHO, which consists of three levels to distinguish whether a website is legitimate. The first level consists of a procedure to extract keywords to identify malicious websites, and second-level keywords are used to identify possible associated domains using a search engine. The victim domain was distinguished by utilizing the features obtained from these websites. Finally, at the last level, the system determines whether the website with doubts at the last level is authorized.

Ref:

Literature	Summary	Pros.	Cons.
[41].	Email based Phishing Detects system using machine learning and NLP techniques.	The major advantage is that the NLP is used to detect the appropriate sentences.	It depends on The email text content analyses. ML is utilizing in the creation of blacklist based on pairs of malicious keywords.limited dataset of 5,009 from phishing and 5,000 from legitimate emails.
[42]	Proposed an entropy based collaborative mechanism for early detection of low rate and high rate DDOS attack and flash events. Packet Header, Time Window size, and other generalized parameters	CAIDA, MIT Lincoln, and FIFA	F measure, precision, False Positive rate and accuracy
[29].	The rich machine learning based system is implemented to detect the phishing websites and URLs based on contents	The main is to catch the novel phishing URLs based on frequently evolving attacks. They expands the number of features for URLs attributes from their previous work (Zhang,2007).	4883 legitimate and 8110 phishing website based limited dataset was used. use services of the third-party companies. use 100 site data collected belongs to only English language and location-specific.
	The machine learning based detection of phishing attack on the client-side through web pages.	It is not dependent on the services of third parties and provide detection in real time. The high accuracy achieved in detection.	but require to analyse the complete page for accessing the source code. The limited dataset of 19 features based on URLs and Source Code.
[40]	The Principal Component Analyses (PCA) used with the Random forest classifier to classify the combined image analyses and heuristic feature based analyses.	independence from language. achieve highest accuracy in detection. also check the web page is replaced with the image or not and detect phishing.	limited dataset was used with 2,119 and 1,407 phishing and legitimate. The legitimate dataset is produced only from the top Alexa's websites. dependent on features of third-party service. 16 features based hyperlink, third party and URL obfuscation based features.
[39]	The combined approach is proposed by utilizing the neural network and reinforcement learning techniques to detect the phishing in emails.	It fast in detecting phishing emails before the end user saw it. does not dependent on services of third party. provide detection of real time.	The limited number record used in the dataset such as 9,118 data and 50.0% are from phishing. Blacklist of PhishTank is used. Only 50 features are used and 12 are from URL based features.
[31]	The Identification of the phishing websites by categorizing them by using the URL attributes.	These systems are appropriate for the client side employment. THese are online classification based system. Resilient to noisy data training.	Use third-party services. The dataset is limited with the 8,155 Legitimate and 6,083 malicious URLs.
[38]	the classification based on neural network with a stable and simple Monte Carlo algorithm.	Its not dependent on the services of third parties. provide real-time detection. Enhance the rate of accuracy and the detection stability. able to detect novel phishing websites also known as zero-day attack.	this system needs to first download the complete page. also used services of third-party. using limited 11,055 data, 55.69% belong to phishing, 30 features used which are address bar, abnormal, HTML, javascript and domain based features.
[36]	Uses NLP for creating some features and with the use of these features classifies the URLs by using three different machine learning approach.	features based on the NLP. The 3 different machine learning algorithms are used and also used hybrid features. 7% increased performance in comparison of Buber, 2017a. 278 features which are consists of 40 NLP and 238 word features.	The dataset is limited and consists of 3,717 malicious and 3,640 legitimate URLs.
[37]	The artificial network proposed based on the particularly self structuring neural networks.	This system was implemented based on adaptive techniques in producing the network. Provide the language based independence.	The services of the third party are used like the domain age. The dataset is limited with the number of 1,400 data and 17 number of features.
[34]	The non linear regression on the bases of a meta-heuristic algorithm by using two methods of feature selection such as wrapper and decision tree.	The original repository of datset UCI is decreased from 30 to 20 number of features that will helps in achieving the better outcome with the methods of decision trees.	The limited dataset is used with 11,055 legitimate and phishing websites and dependent on third-party services with 20 features
[33]	Define some URL features, and with them, they generate some rules with apriori and predictive apriori rule generation algorithms.	fast detection with rules (especially with apriori rules)	classification for based on the rules. rules dependent that quality of the rules effects the work. Th dataset is limited to 1200 URLs of phishing and 200 are legitimate. There are 14 features are Heuristic, 9 priori and 9 predictive apriori rules.

Chapter No. 3:
Related work

RELATED WORK

The internet is a vast network-based industry full of hackers, attackers or cyber criminals. Civilians, businessmen, industries, and every market that consists of the Internet and networks need security to prevent phishing and provide protection to their customers, as well as to their own system safety. The methodology proposed in this study was successfully implemented as a prototype using a dataset comprising phishing and legitimate URLs. These experiments are carried out using many machine learning algorithms that are discussed separately in each heading to evaluate and illustrate the effects of the machine learning algorithms that are given below.

A. EXPERIMENTAL RESULTS OF DECISION TREE:

The decision tree algorithm depends on tree-based architecture, which consists of several internal nodes and leaves that carry data according to the patterns found in the dataset. The sklearn library was used to access the tools for implementing the decision tree algorithm. Table 2. presents the results of the proposed decision tree algorithm with the phishing dataset to classify URLs in binary classes of 0 and 1.

Decision tree algorithms consist of many parameters, but the most effective parameter that affects the training and prediction accuracy of the model is max_depth. This parameter defines the depth of the tree in terms of its level. The more the levels, the more complex the structure becomes with each level, but this makes it easier for the model to extract the patterns from the dataset for training. Table 2. shows the results of the decision tree with different numbers of max_depth such as 0, 5, 10, 20, and 30. An increase in the depth of the tree increases the accuracy and other results of the model. However, at a depth of 30 the model showed the highest accuracy of 95.41%, precision of 95.8%, recall of 96%, specificity of 94.66%, and recall 95.91%.

TABLE 2. Results for the performance of the decision tree model.

max depth	Accuracy	Precision	Recall	Specificity	F1-score
0	94.9	95.46	95.41	94.25	95.44
5	92.07	89.67	96.97	85.85	93.18
10	94.3	94.59	95.23	93.09	94.92
20	95.38	95.7	96.06	94.53	95.88
30	95.41	95.8	96	94.66	95.91

The model presented an accuracy of 95.41 %, which means that the model had an overall accuracy of 95.41%. The Figure 6 presents the results in the form of bar

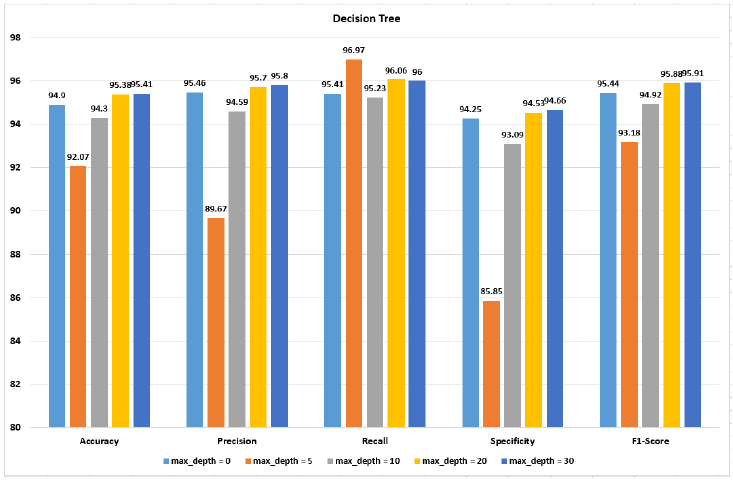


FIGURE 6. Experimental results of the decision tree model.

graph that illustrates the very clear difference visually in between each training depth level.

B. EXPERIMENTAL RESULTS OF NAIVE BAYES:

The naive Bayes algorithm consists of probability mechanisms that extract patterns from the dataset using the formula presented in Equation 6. The linear naive Bayes algorithm was used for this dataset because most datasets are presented in the form of discrete values, and the linear naive Bayes algorithm is appropriate according to the dataset. The naive Bayes algorithms showed highest results in Table 3, with accuracies of 88.39%, precision 94.92%, recall 83.71%, specificity 94.32%, and F1 score 88.96%, respectively. The accuracy shows the overall model accuracy prediction rate of the extent to which the model predicts or distinguishes between legitimate and phishing URLs. The precision illustrates the true positive rate of the model from all the true and false phishing predictions that the extent to which the model predicts the URLs phishing and, in reality, these URLs are also phishing. Recall presents the sensitivity of the model, which illustrates how many predictions are phishing URLs from all the true positive and false negative predictions. The F1 score is the harmonic mean of the precision and recall, which represents the balance between precision and recall results. Figure 7. presents the results in a visual form.

TABLE 3. Results for the performance of the naive bayes model.

Accuracy	Precision	Recall	Specificity	F1-score
88.39	94.92	83.71	94.32	88.96

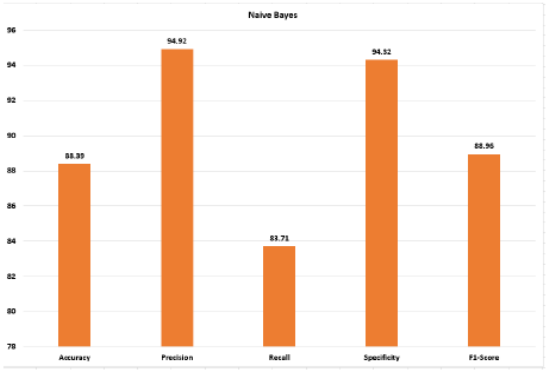


FIGURE 7. Experimental results of the naive bayes model.

TABLE 4. Results for the performance of the linear regression model.

normalization	Accuracy	Precision	Recall	Specificity	F1-score
False	58	99.6	27.11	99.7	41.74
True	58.83	100	26.37	100	41.74

C. EXPERIMENTAL RESULTS OF LINEAR REGRESSION:

Linear regression is a learning model that presents the best results with normalization=True. The linear regression algorithm reduces the residual sum of the square rate by observing the target, and the predictions are made using approximation methods. The highest results were achieved an accuracy of 58.83%, precision of 100%, recall of 26.37%, specificity of 100% and an F1-score of 41.74%. A visualization of the performance of the model is shown in Figure 9

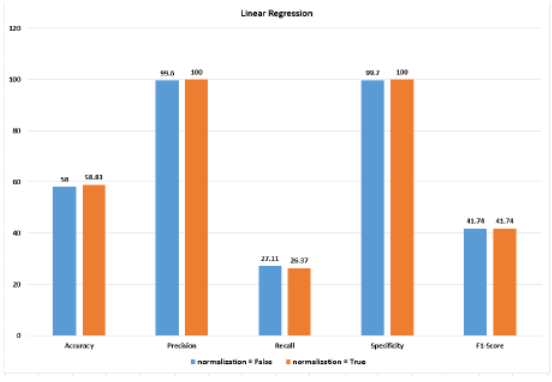


FIGURE 8. Experimental results of the linear regression model.

D. EXPERIMENTAL RESULTS OF K-NEIGHBORS CLASSIFIER:

TABLE 5. Results for the performance of the K-neighbors classifier model.

n neighbors	Accuracy	Precision	Recall	Specificity	F1-score
2	61.56	77.38	44.12	83.66	56.206
3	63.12	67.02	66.99	58.23	67.08
4	58.63	68.79	47.578	72.65	56.25

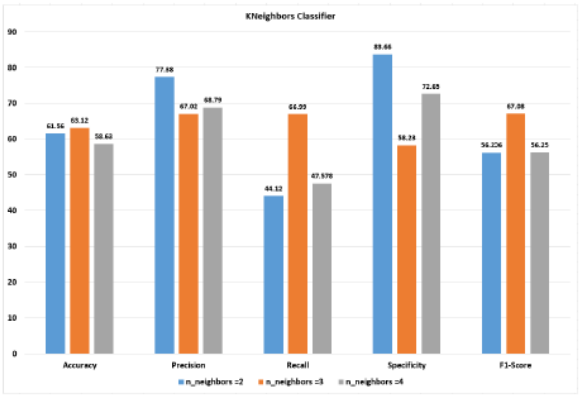


FIGURE 9. Experimental results of the K-neighbors classifier model.

The K-Neighbors classifier is dependent on K nearest neighbors and classifies the texting input by predicting the class. K-Neighbor was originally a clustering technique, but it is also effective with labelled datasets and in performing classification based on these true labels. The K-Neighbors classifier was selected for the experiments becasue of its functionality. It creates groups of dataset points that are named features based on the centroids selected according to the number of classes. N_neighbors is the hyperparameter used with the K-Neighbors classifier because it needs to know the number of groups it has to make. The experiments were performed with three different numbers of n_neighbors 2, 3, and 4. The results has been shown in Table 5. The highest results were obtained with no. 3 n_neighbors: accuracy achieved that are accuracy 63.12%, precision 67.02%, recall 66.99%, specificity 58.23%, and F1-score 67.08% as shown in Figure 9. These results are relatively lower than those of the other algorithms but higher than those of the linear regression algorithm.

E. EXPERIMENTAL RESULTS OF SUPPORT VECTOR MACHINE MODEL:

The support vector machine consists of the concept of a hyperplane that differentiates the data by using a plane, and by setting the hyperparameter the hyperplane sets its position that accurately differentiates between the phishing and legitimate data URLs. The highest accuracy is obtained with the maximum number of iteration parameters which is max_iter. Max_iter represents the number of

iterations performed by the SVM algorithm for training. In each iteration, it measures the distance between the hyperplane and the data points of the dataset. Subsequently, in each iteration, the data points were classified into their predicted classes. Then, according to the newly classified data points, the iteration was again performed to make it more accurate for prediction purposes and to obtain the highest accuracy results. Table 6. presents the results with max_iter values as 10, 20, 30, 40, and 50. The highest results achieved with 50 max_iter with an accuracy of 71.8%, precision of 96.34%, recall of 49.81%, specificity of 97.606%, and F1-score of 65.67%. Figure 10. presents a visual representation of the results and illustrates the clear differences between the results of each iteration.

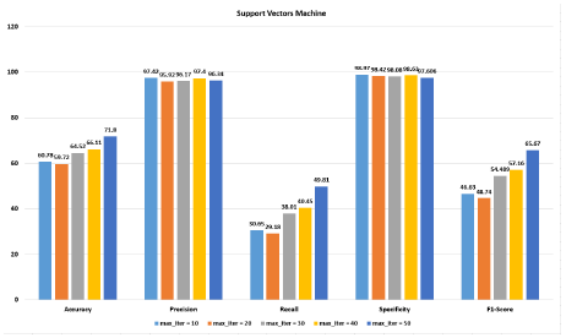


FIGURE 10. Experimental results of the support vector machine model.

TABLE 6. Results for the performance of the support vector machine model.

max iter	Accuracy	Precision	Recall	Specificity	F1-score
10	60.78	97.42	30.65	98.97	46.63
20	59.72	95.92	29.18	98.42	44.74
30	64.52	96.17	38.01	98.08	54.489
40	66.11	97.4	40.45	98.63	57.16
50	71.8	96.34	49.81	97.606	65.67

TABLE 7. Results for the performance of the random forest model.

max depth	Accuracy	Precision	Recall	Specificity	F1-score
10	95.32	94.36	97.46	92.61	95.88
20	96.8	96.68	97.62	95.76	97.15
30	96.77	96.73	97.51	95.83	97.12
40	96.77	96.73	97.51	95.83	97.12
50	96.77	96.73	97.51	95.83	97.12

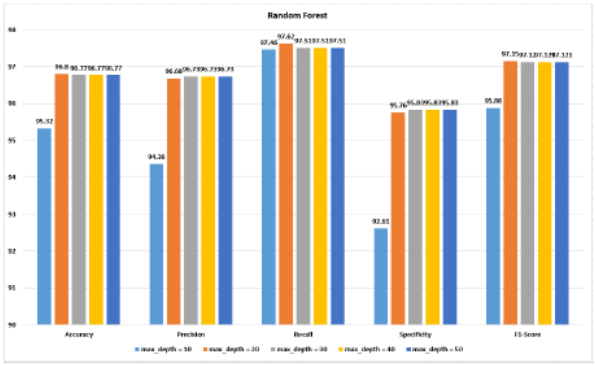


FIGURE 11. Experimental results of the random forest model.

F. EXPERIMENTAL RESULTS OF RANDOM FOREST MODEL:

Random forest is an ensemble technique that combines multiple decision tree algorithms. The random forest algorithm divides samples into different numbers and creates a decision tree for each sample. Then, each decision tree predicts its results, and finally, the averaging methods are used with the sum of every decision tree result. This technique helps the model extract effective prediction results with the phishing URLs dataset.The results has been shown in Table 7 and Figure 11. The highest results were achieved with the max_depth hyperparameter at different depth rates such as 10, 20, 30, 40, and 50. The highest results were achieved with a depth of 30, with an accuracy of 96.77%, precision of 96.73%, recall of 97.51%, specificity of 95.83%, and F1-score of 97.12%. Random forest outperformed all other algorithms and achieved the highest results for all the applied machine learning algorithms. Further, the comparative analyses section presents the comparative results of the applied and proposed ensemble model, which illustrates the difference in the results of the machine learning model. Figure 11. presents a visual presentation of the results of the random forest model at every

depth rate.

G. EXPERIMENTAL RESULTS OF GRADIENT BOOSTING MODEL:

Gradient boosting is an ensemble learning model that consisting of the architecture of multiple trees. However, the working mechanism makes it more efficient and effective for extracting deep patterns from the data. Gradient boosting comprises the boosting and bagging concepts. Gradient boosting selects the samples from the dataset, creates a tree according to the samples, and performs learning iterations on these data. The samples were selected randomly from the dataset records, and the remaining samples were placed in bagging which was used with the next upcoming iterations of the learning process. The gradient boosting model also performs better with hyperparameter tuning of the param eter max_depth, such as 2, 5, 8, 10, and 12. The highest results were achieved with an accuracy of 70.34%, precision of 99.65%, recall of 47.41%, specificity of 99.79%, and F1-score of 64.10%, with a depth of 10. Figure 12. presents the results in the visualized form of a bar graph that illustrates variations in the results.

TABLE 8. Results for the performance of the gradient boosting model.

max depth	Accuracy	Precision	Recall	Specificity	F1-score
2	62.37	100	32.68	100	49.26
5	67.62	99.87	42.17	99.93	59.3
8	68.17	99.38	43.41	99.65	60.43
10	70.34	99.65	47.24	99.79	64.1
12	68.17	99.38	43.41	99.65	60.43

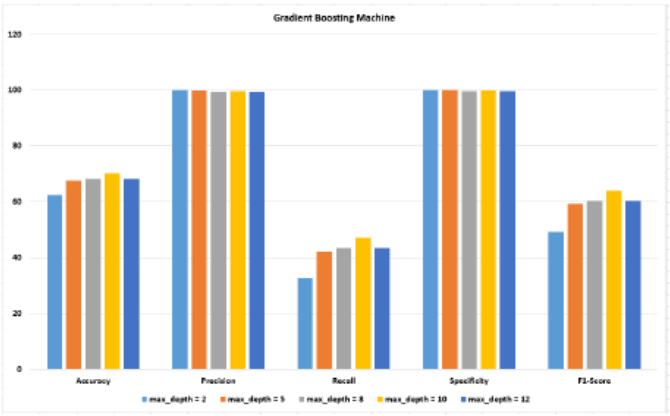


FIGURE 12. Experimental results of the gradient boosting model.

H. EXPERIMENTAL RESULTS OF PROPOSED APPROACH:

A hybrid approach was adopted to enhance the results and efficiency of the machine learning models. The linear regression (LR), support vector classifier (SVC), and decision tree (DT) are combined as (LR+SVC+DT) using two different voting techniques, soft and hard. Voting methods are used to combine multiple machinelearning models and perform averaging operations on the results of each combined model. The Canopy

TABLE 9. Results for the performance of the hybrid model (LR+SVC+DT).

Voting	Accuracy	Precision	Recall	Specificity	F1-score
Soft	95.23	95.15	96.38	93.77	95.77
Hard	94.09	93.31	96.33	91.25	94.79

based feature selection method with cross fold validation and grid search hyper parameter tuning technique is used with proposed ensemble LSD model. Although

this technique improved the results with much higher expectations, in this study, the hybrid model achieved results, with accuracy of 95.23%, precision of 95.15%, recall of 96.38%, specificity of 93.77%, and F1-score 95.77%, respectively. These results are much higher and better than those of the other applied machine learning algorithms but lower than those of the random forest model. Figure 13. illustrates the differences between the results of the hybrid (LR+SVC+DT)

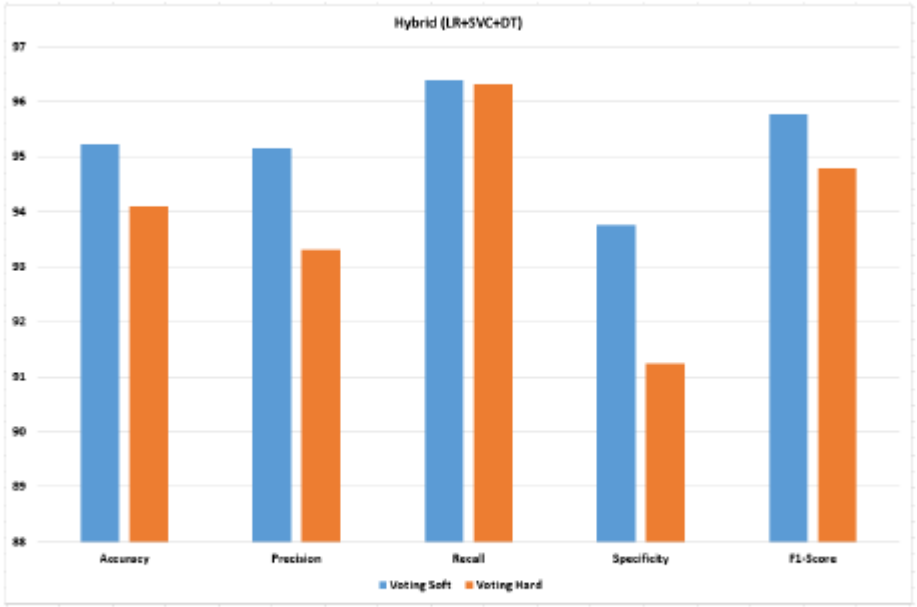


FIGURE 13. Experimental results of the hybrid (LR+SVC+DT) model.

Chapter No. 4:

Results and Discussion

DISCUSSION

Different machine learning models were used in this study and the previous sections presented the results and effects of the machine learning model on the classification process of phishing and legitimate URLs. Comparative analyses of all the multiple machine learning models are presented in this section. Table 11. and Figure 14. presented the clear and significant effects of machine learning models in this study. The highest results were achieved with proposed approach, with an accuracy of 98.12%, precision of 97.31%, recall of 96.33%, specificity of 96.55%, and F1-score of 95.89%, which outperformed the other utilized machine learning models.

TABLE 10. Results for the performance of the hybrid model (LR+SVC+DT).

Models	Accuracy	Precision	Recall	Specificity	F1-score
Linear Regression	58.83	100	26.37	100	41.74
Decision Tree	95.41	95.8	96	94.66	95.91
Random Forest	96.77	96.73	97.51	95.83	97.12
Naive Bayes	88.39	94.92	83.71	94.32	88.96
Support Vector Machine	71.8	96.34	49.81	97.606	65.67
Gradient Boosting Machine	70.34	99.65	47.24	99.79	64.1
Hybrid (LR+SVC+DT) soft	95.23	95.15	96.38	93.77	95.77
Hybrid (LR+SVC+DT) hard	94.09	93.31	96.33	91.25	94.79
Proposed approach	98.12	97.31	96.33	96.55	95.89

The comparative analyses illustrate that the machine learning model that consists of linear approaches or probabilistic approaches, such as linear regression and support vector machines, do not perform very well and show very low results. The ensemble and tree-based models presented highly effective and significant results in the classification of phishing URLs. The highest and most efficient results were achieved with the proposed approach, with an accuracy of 98.12%, precision of 97.31%, recall of 96.33%, specificity of 96.55%, and F1-score of 95.89%. These results illustrate that the random forest model outperforms all the other machine learning models. Comparative analyses of the machine learning algorithms showed that the ensemble tree architecture-based models presented better results than linear and probabilistic models. The hybrid model (LR+SVC+DT) performed better and yielded higher accuracy results than the other machine learning models, with an accuracy of 95.23%, precision of 95.15%, recall of 96.38%, specificity of 93.77%, and F1-score of 95.77%, but lower than that of the proposed approach.

Chapter No. 5:

Conclusion and Future Scope

CONCLUSION

The Internet consumes almost the whole world in the upcoming age, but it is still growing rapidly. With the growth of the Internet, cybercrimes are also increasing daily using suspicious and malicious URLs, which have a significant impact on the quality of services provided by the Internet and industrial companies. Currently, privacy and confidentiality are essential issues on the internet. To breach the security phases and interrupt strong networks, attackers use phishing emails or URLs that are very easy and effective for intrusion into private or confidential networks. Phishing URLs simply act as legitimate URLs. A machine-learning-based phishing system is proposed in this study. A dataset consisting of 32 URL attributes and more than 11054 URLs was extracted from 11000+websites. This datasetwas extracted from the Kaggle repository and used as a benchmark for research. This dataset has already been presented in the form of vectors used in machine learning models. Decision tree, linear regression, random forest, support vector machine, gradient boosting machine, K-Neighbor classifier, naive Bayes, and hybrid (LR+SVC+DT) with soft and hard voting were applied to perform the experiments and achieve the highest performance results. The canopy feature selection with cross fold validation and Grid search hyper parameter optimization techniques are used with LSD Ensemble model. The proposed approach is evaluated in this study by experimenting with a separate machine learning models, and then further evaluation of the study was carried out. The proposed approach successfully achieves its aim with effective efficiency. Future phishing detection systems should combine list-based machine learning-based systems to prevent and detect phishing URLs more efficiently.

REFERENCES

[1] N. Z. Harun, N. Jaffar, and P. S. J. Kassim, ‘‘Physical attributes significant in preserving the social sustainability of the traditional malay settlement,’’ in Reframing the Vernacular: Politics, Semiotics, and Representation. Springer, 2020, pp. 225–238.

[2] D. M. Divakaran and A. Oest, ‘‘Phishing detection leveraging machine learning and deep learning: A review,’’ 2022, arXiv:2205.07411.

[3] A. Akanchha, ‘‘Exploring a robust machine learning classifier for detecting phishing domains using SSL certificates,’’ Fac. Comput. Sci., Dalhousie Univ., Halifax, NS, Canada, Tech. Rep. 10222/78875, 2020.

[4] H. Shahriar and S. Nimmagadda, ‘‘Network intrusion detection for TCP/IP packets with machine learning techniques,’’ in Machine Intelligence and Big Data Analytics for Cybersecurity Applications. Cham, Switzerland: Springer, 2020, pp. 231–247.

[5] J. Kline, E. Oakes, and P. Barford, ‘‘A URL-based analysis of WWW structure and dynamics,’’ in Proc. Netw. Traffic Meas. Anal. Conf. (TMA), Jun. 2019, p. 800.

[6] A. K. Murthy and Suresha, ‘‘XML URL classification based on their semantic structure orientation for web mining applications,’’ Proc. Comput. Sci., vol. 46, pp. 143–150, Jan. 2015.

[7] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, ‘‘Phishing website detection: An improved accuracy through feature selection and ensemble learning,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252–257, 2019.

[8] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, ‘‘PhishAri: Automatic realtime phishing detection on Twitter,’’ in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.

[9] S. N. Foley, D. Gollmann, and E. Snekkenes, Computer Security— ESORICS 2017, vol. 10492. Oslo, Norway: Springer, Sep. 2017.

[10] P. George and P. Vinod, ‘‘Composite email features for spam identification,’’ in Cyber Security. Singapore: Springer, 2018, pp. 281–289.

[11] H. S. Hota, A. K. Shrivas, and R. Hota, ‘‘An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique,’’ Proc. Comput. Sci., vol. 132, pp. 900–907, Jan. 2018.

[12] G. Sonowal and K. S. Kuppusamy, ‘‘PhiDMA—A phishing detection model with multi-filter approach,’’ J. King Saud Univ., Comput. Inf. Sci., vol. 32, no. 1, pp. 99–112, Jan. 2020.

[13] M. Zouina and B. Outtaj, ‘‘A novel lightweight URL phishing detection system using SVM and similarity index,’’ Hum.-Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.