IEEE Transaction on Machine Learning,Volume:11,Issue Date:03.March.2023

Phishing Detection System Through Hybrid Machine Learning Based on URL

ABDUL KARIM[1], MOBEEN SHAHROZ[2], KHABIB MUSTOFA[1], SAMIR BRAHIM BELHAOUARI[3], AND S. RAMANA KUMAR JOGA[4]

[1]Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
[2]Department of Artificial Intelligence, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan
[3]Division of Information and Computer Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[4]Dadi Institute of Engineering and Technology, Anakapalle, Andhra Pradesh 531002, India

Corresponding author: Khabib Mustofa (khabib@ugm.ac.id)

This work was supported by Student's Final Project Recognition RTA (Rekognisi Tugas Akhir) Grant for the fiscal year 2022.

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-neighbors 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 valoidation 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.

INDEX TERMS Voting classifier, ensemble classifier, machine learning, uniform resource locator (URL), logistic regression, support vector machine, and decision tree (LSD), protocol, cyber security, social networks.

I. 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

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani.

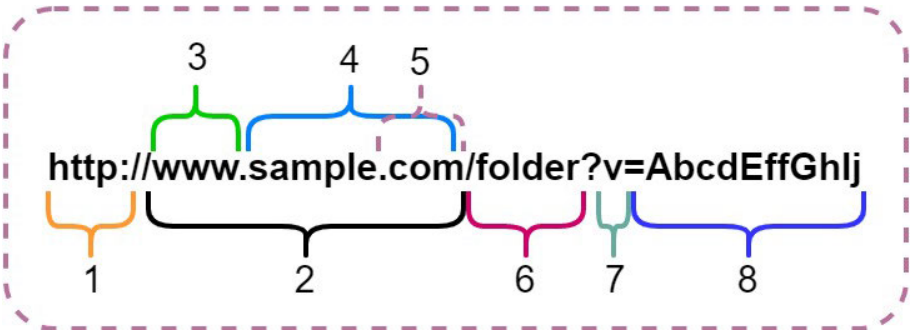

FIGURE 1. URL presentation based on HTTP.

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. 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.

In the Figure 1, label1 is representing HTTP (Hypertext transfer protocol) that is used for obtaining resource as per client request. Label 2 represents a hostname, the host came is further divided into three subdomains: top-level domain (also called web address), and domain labeled 6 refers to the directory of a web server. Label 7 ''v'' character holds a value ''AbcdEffGhIJ'' and a label 6 ''?'' initialize the parameter x in a URL. URL commonly represent website addresses [64], [5]. In, HTTP functions were used as the request protocol in the computing model of the client server. This defines the communication rules. Servers and web browsers use HTTP to exchange webpages. The web browser is a client and the computer is the host on which the app is running.

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. HTTP or HTTPS Both are used in the process of data retrieval from the web server to view content in a browser. HTTPS [1], [2] uses

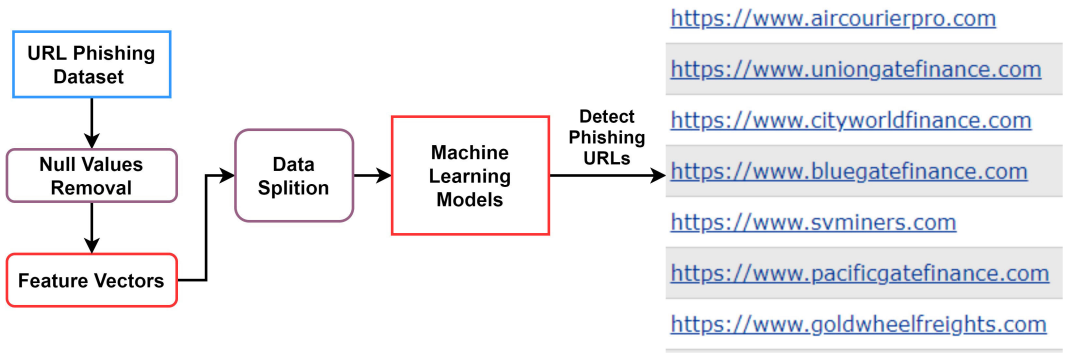

FIGURE 2. Detection of phishing URLs and structure of proposed approach.

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 [3] 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 [14] in which 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 [15] 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 [16] 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 [17] 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 [17] 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 [17] 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 [18], [19] 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. The general architecture of the proposed approach is shown in Figure 2. The major contributions of this study are as follows.

- Phishing URL-based cyberattack detection is proposed in this study to prevent crime and protect people's privacy.
- The dataset consists of 11000+ phishing URL attributes that help classify phishing URLs based on these attributes.
- Machine learning models have been applied, such as decision tree (DT), linear regression (LR), naive Bayes (NB), random forest (RF), gradient boosting machine (GBM), support vector classifier (SVC), K-Neighbors classifier (KNN), and the proposed hybrid model (LR+SVC+DT) LSD with soft and hard voting, which can accurately classify the threats of phishing URLs.

- Cross-fold validation with a grid search parameter based on the canopy feature selection technique was used with the proposed LSD hybrid model to improve prediction results.
- The proposed methodology must be evaluated using evaluation parameters, such as accuracy, precision, recall, specificity, and F1-score.

The remainder of this paper is organized as follows. Section II presents the previous work of researchers and authors who have contribute to the related domains. Section III presents the materials and methods used in this study in the experimental and implementation phases. Section IV presents the experimental and comparative results analyzed to evaluate this study in a scientific manner. Finally, section VI presents the conclusions of the study.

II. RELATED WORK

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. Whitlist-based Phishing identification systems produce protected and reliable websites to 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% false-negative 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. Zero-day 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 PhishNet [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 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 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 PhishWHO, 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.

In 2011, [31] proposed a system for the identification of phishing websites by classifying these websites by utilizing the number of attributes, such as directory, file name, domain name, counting the number of special characters, and length. By applying a support vector machine (SVM), security systems can classify phishing websites in offline mode. Other techniques and machine learning algorithms, such as weighted confidence, adaptive regularization of weights, and online perceptrons, are adopted for classification in online mode. In the analyses of the comparative results, the experiments show that the adaptive regularization of weights algorithm outperforms the other algorithms by achieving the highest accuracy rate and utilizing a minimum number of system resources. The ranking and message title based on the incoming message were ranked as described by Islam and Abawajys study [32]. These studies produced a classification system based on multiple layers to clarify the significance of messages. The experimental results revealed that the proposed method decreased the number of false positives. The discriminant features are extracted by [33] and associated with the security of the transport layer synchronically among features of attributes that are based on URLs, such as the total number of used slashes, length, positions of dots, and numbers in the subdomain and URL names. The Apriori algorithm was established on the basis of rule detection using rule mining. The experimental results revealed that 93% accurately detected the phishing URLs.

In modern research [34], a nonlinear regression approach is used to determine whether a website is authorized or phishing. This previous study preferred the use of metaheuristic algorithms, such as the harmony search approach and SVM for the training of the system. Accordingly to this, harmony search provides a more reliable accuracy rate of 94.13% for training and 92.80% for test methods, sequentially by employing it on approximately 11,000 webpages. The [35] presents the previous version of this same study that proposed the phishing identification system based on 209 word and 17 features based on Natural Language Processing (NLP). Previous studies have shown that NLP has a significant effect. This consists of features extracted using NLP. However, it is necessary to enhance the number of features based on NLP and word vectors. Accordingly, in continuing research, the focus of this study focused on this matter and attained more reliable outcomes with an accuracy rate of 7%. The [36] the representation of vector created using NLP was improved in the proposed system, and to evaluate the study three separate machine-learning models were examined based on the accomplished accuracy score. The [37] study in presents a phishing detection system implemented for classification using dynamic self-structure-based neural networks. There were 17 features used and mostly belonged to the services of the third party. Consequently, accomplishing a real scenario requires considerably more time; however, it can yield more dependable accuracy results. It utilizes an insufficient dataset with 1,400 records but confers high recognition for noisy data.

The [38] introduce a new neural network-based approach for the classification system for the identification of phishing websites by applying the risk minimization principle and Monte Carlo algorithm. Thirty features were used, which were classified into four major fields: abnormal features, address bar-based features, JavaScript-based features, HTML, and domain-based features. The identification mechanism achieves 97.71% an accuracy score of 1.7% and a positive rate in the experimental investigations. Thus, many researchers have focused on security mechanisms for the detection of phishing using URLs. Some researchers have proposed machine learning-based systems for email classification to detect phishing emails based on email packet data. A combination of reinforcement learning and neural networks was proposed in [39] for the classification of phishing and authorized URLs. This system consists of 50 features grouped into four different classifications or classes: a header of the mail, HTML content, contents of the URLs, and main text. The major focus of this system is on the email based on URL-extracted features that contain the similarity score with the approach proposed in that study. The dataset consists of 9118 emails data from which 50% are belong to legitimate and the rest of the emails are recorded as phishing. The highest result achieved by this approach was 1.8% of

TABLE 1. Comparison of the existing systems.

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.

false positive rate, with a 98.6% accuracy score. Different hybrid approaches and image-checking-based systems consist of machine learning and deep learning. Some have been proposed by other researchers [40].

One of the major dependencies was faced by these systems that were developed for the image dependent phishing detection was the dataset or database for the initial purposes or website history as prior knowledge of the web pages. However, the proposed approach was independent of these dependencies. Three classes of features were used: hyperlink, third-party, and URL obfuscation-based features. However, the accuracy rate increased to 99.55%, and the detection time is also increased by using third-party services.

A recent study used NLP [41] for the detection of phishing emails. It presents a system based on the semantic content analysis of emails, such as a simple text problem, for the identification of suspicious intent. To achieve this, NLP is utilized based on command and question-based sentences, and then a specific words-based blacklist is utilized for the detection of phishing contents. The dataset consisted of 5009 emails labelled as phishing and 5000 are labelled as legitimate emails for training and testing purposes. The highest precision rates were achieved at 95% of the experimental results.

III. MATERIAL AND METHODS

Phishing detection based on URLs proposed in this study. The classification of phishing URLs was implemented using machine learning algorithms. Cybercrimes are growing with the growth of Internet architecture worldwide, which needs to provide a security mechanism to prevent an attacker from getting confidential content by breaching the network through fake and malicious URLs. A phishing dataset was used to perform the experiments. The dataset is in the form of data vectors that require null-value removal to remove unnecessary empty values. Multiple machine learning algorithms, such as decision tree (DT), linear regression (LR), naive Bayes (NB), random forest (RF), gradient boosting machine (GBM), support vector classifier (SVC), K-neighbors classifier, and the proposed hybrid model (LR+SVC+DT) LSD with soft and hard voting were used based on functional features, as shown in Figure 3. To improve the prediction results, a cross-validation technique with grid search hyper-parameter tuning based on canopy feature selection was designed using the proposed LSD hybrid model. Finally, predictions were made to classify the phishing URLs and evaluate their performance in terms of accuracy, precision, recall, specificity, and F1-score.

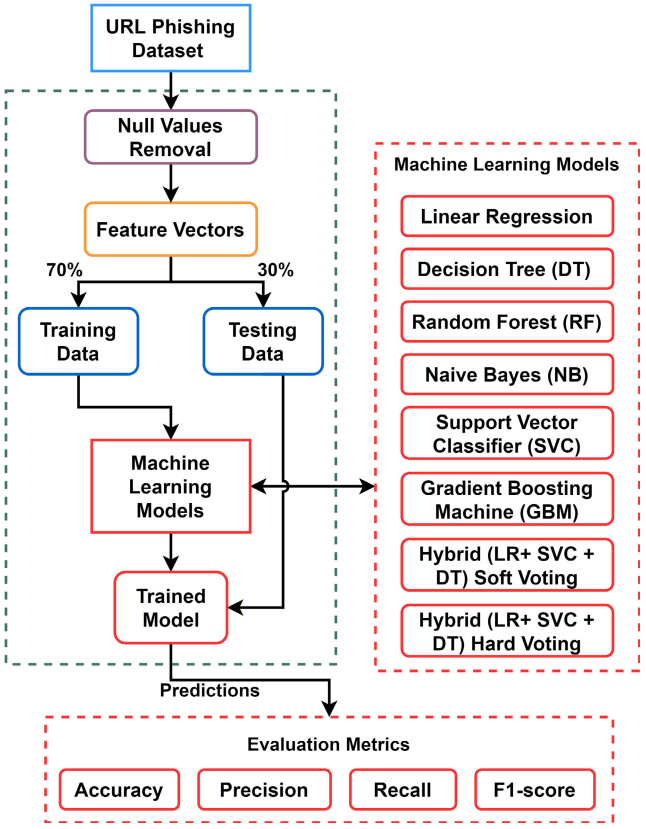

FIGURE 3. Classification of phishing URLs proposed based on designed methodological structure.

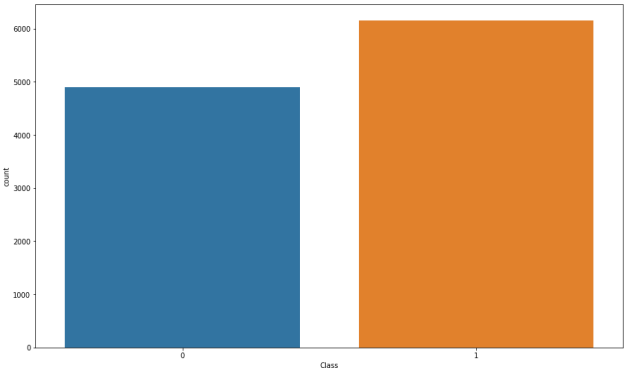

FIGURE 4. Dataset presentation according to number of classes phishing and legitimate, where (1) presents phishing and (0) legitimate URLs.

A. URL BASED PHISHING DATASET

The dataset was collected in a CSV file from the well-known dataset repository called Kaggle, which provides benchmark datasets for research purposes. The dataset consisted of 11054 number of records and 33 attributes extracted from 11000+ websites. The phishing and legitimate website URLs contain some common attributes such as UsingIP, LongURL, ShortURL, Symbol@, Redirecting//, PrefixSuffix-, SubDomains, HTTPS, DomainRegLen, Favicon, NonStdPort, HTTPSDomainURL, Reques- tURL, AnchorURL, LinksInScriptTags, and ServerFormHandler, which help to identify whether the URL is phishing. The dataset consisted of two classes: phishing and legitimate, as shown in Figure 4. The dataset was in the form of vectors that needed to be refined. The null values were removed from the dataset for preprocessing. After preprocessing, the complete dataset was converted into a single corpus and used for further processing. The complete corpus was divided into two partitions, 70% for training and 30% for testing the 70% training data were used to train the machine learning model, preserve 30% of the data for the predictions, and evaluate the performance of the proposed approach.

B. APPLIED MACHINE LEARNING ALGORITHMS

Machine learning algorithms are models that can be represented as mathematical models of real-life scenarios of world processes, known as algorithms. First, the algorithms are trained, and then, the trained model performs learning based on this training and extracts patterns from the dataset. After the training test split of the dataset, it was partitioned into training and testing data. The training data give the machine learning model as an input, and testing gives the trained model as an input that is ready to perform prediction on the testing data. Different machine-learning algorithms were used in this study, which provided different accuracy for different feature engineering techniques.

1) DECISION TREE

The decision tree classifier (DTC) is a non-parametric method used for classification and regression. The decision tree classifier recursively partitions the given dataset of rows by applying the depth-first greedy method [44] or the breadth-first approach [45] until all data parts relate to an appropriate class. A decision tree classifier structure was created for the root, internal, and leaf nodes. Tree construction was used to classify unknown data. At each inner node of the tree, the best separation decision is made using impurity measures [46]. The leaves of the tree were created from the class labels in which the data objects were gathered. The DT (decision tree) classification procedure is implemented in two stages: tree building and tree pruning [44]. It is very tasking and computationally fast because the training dataset is frequently traversed. For a single attribute, entropy is mathematically expressed as:

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{1}$$

The Entropy can be numerically expressed for various characteristics as follows:

$$E(T, X) = \sum_{c \epsilon X} p(c) E(c) \tag{2}$$

IG is defined mathematically by:

$$IG(T, X) = E(T) - E(T, X) \tag{3}$$

2) SUPPORT VECTOR MACHINE

A support vector machine (SVM) is a supervised machine learning algorithm defined by a separating hyperplane between different classes. In other words, given the labeled training data (supervised learning), the algorithm outputs an optimal hyperplane that categorizes new test data based on the training data. Support vector machine (SVM) can be used for both classification and regression. However, it is mostly used in classification problems, where it provides the best accuracy between two classes. In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features in the dataset), with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiates the two classes very well, and the classes in this dataset are zero and one that are easily separated by the hyperplane, as this algorithm performs the best for two classes [48], [49].

$$x.y = x_1 y_1 + x_2 y_2 = \sum_{i=1}^{2} (x_i y_i) \tag{4}$$

3) GRADIENT BOOSTING MACHINE

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models to create a strong predictive model to increase the accuracy of the model [50]. Decision trees are typically used to perform gradient boosting for data classification. Gradient boosting is a machine learning algorithm that is used for regression and classification problems, which produces a prediction model used for the classification of data in the form of an ensemble of weak prediction models, such as decision trees that decide to classify the data. Its tuning parameters, such as n_estimators = 100, max_depth = 12, and learning_rate = 0.01, are tuned and the algorithm performs well, where n_estimators is the number of boosting stages to perform well by the classifier and a large number usually results in a better performance of the algorithm, max_depth = 10, the maximum depth of a tree limits the number of nodes in the tree and tunes this parameter for the best performance; the best value depends on the input variables, which increases accuracy after tuning it; the learning_rate = 0.01 learning rate reduces the contribution of each tree by the learning rate parameter, and there is an adjustment between the learning rate and n estimators.

4) RANDOM FOREST

The random forest algorithm makes decision trees on the test data set, fines the prediction from each of them, and finally selects the best solution by implementing voting. This method is an ensemble method that is better than a single decision tree because it reduces over fitting by averaging the result. The random forest classifier [52] uses a decision tree as the base classifier. Random forest creates various decision trees; the randomization is present in two ways: first, random sampling of data for bootstrap samples as it is done in bagging [53], and second, randomly selecting the input features to create individual base decision trees [52]. Based on accuracy measures, the random forest algorithm is an existing ensemble technique that includes bagging and boosting.

$$F(x_t) = \frac{1}{B} \sum_{B}^{i=0} F_i(x_t) \tag{5}$$

5) NAIVE BAYES

The naive Bayes classifier is one of the simplest and most effective machine learning classification algorithms, which helps in building a fast machine learning classifier that can

make quick predictions from a given dataset. This is a probabilistic classifier, meaning that it is predicted based on the probability of an object. The naive Bayes algorithm is a probabilistic classifier built upon Bayes' theorem as follows:

$$P(A|B) = \frac{(P(B|A) * P(A))}{(P(B))} \tag{6}$$

A is the class and B is the feature vector represented in [55] and [56]. The naive Bayes algorithm is a probabilistic classifier built on Bayes' theorem. P (B|A), P(A), and P(B) are the probabilities measured from earlier known instances, such as training data [55], [56]. Classification errors are minimized by selecting a class that maximizes the probability P(A|B) for every occurrence [55], [57], [58].

6) K NEAREST NEIGHBORS(KNN) CLASSIFIER

The K-nearest neighbors (KNN) machine learning model is a supervised classifier utilized in machine learning for both classification and regression problems. The KNN model uses training data for the learning process and transforms them into data points according to the relationship measure, also known as the similarity or distance function-based Euclidean distance function, to classify the testing data points. KNN classifies the data points by voting on the results of K- nearest neighbors and calculating the similarity between them. K-nearest neighbors (KNN) are extensively utilized in text categorization because they are simple and efficient. However, KNN still experiences misfits in models whose outcomes of its hypotheses, such as the hypothesis that the training data are equally divided between classes [59], [60], [61].

7) HYBRID LR+SVC+DT USING SOFT VOTING AND HARD VOTING

The voting classifier is the simplest form of combining different classification algorithms, and selecting the combination rule is important for designing classifier ensembles. Voting combines the predictions of multiple machine learning algorithms [62]. The average voting scheme combines the predictions of three algorithms, namely, the random forest classifier, support vector machine classifier, and naive Bayes classifier. These algorithms provided the best accuracy compared to the others; therefore, three of these algorithms were used for averaging voting classification. It performed well and provided an accuracy of 95.75 in this voting scheme. Second, a voting classifier with the stacking method is used, and three classifiers are used for this voting purpose; support vector machine, random Forest, and naive Bayes, which perform well and provide the highest accuracy of approximately 97.24%. It has the best accuracy among all the algorithms and voting classifiers. In these methods, the driving policy is to build several estimators separately, and then calculate the average of their predictions. On average, the mixed estimator is generally better than any of the single-base estimators because its variation is reduced [62].

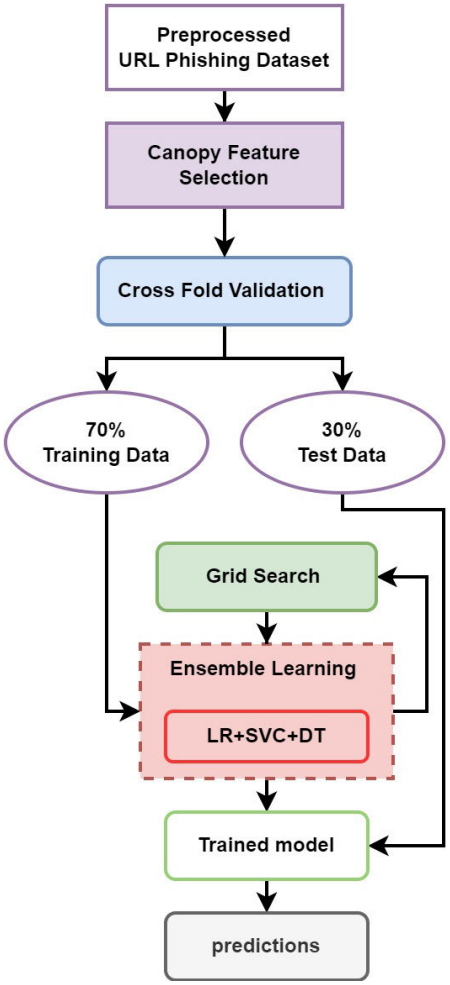

FIGURE 5. Proposed approach based on canopy feature selection with LR+SVC+DT (LSD) ensemble learning model using grid search hyperparameter tuning and cross fold validation.

a: PROPOSED METHODOLOGY ARCHITECTURE

The proposed approach presented in Figure 5. The canopy centroid selection method is used in clustering as a preprocessing step. Here, the canopy is used as feature selection method as a feature engineering step to select the most effective feature in the detection of phishing URLs. The Ensemble model is based on three different machine learning models such as linear Regression, Support vector Machine, and Decision Tree with Hyper parameter tuning technique to select the best parametric values for the training process of the model. The cross fold validation is used for the effective train test split which improves the training of the model.

C. EVALUATION PARAMETER

Machine learning performance must be evaluated using several evaluation parameters. The machine-learning algorithm provides results in the form of predictions. The evaluation parameters measure the number of true and false predictions made by the model in both legitimate and phishing classes.

Parameters such as accuracy, precision, recall, specificity and the F1-score were used.

Accuracy measures model performance in terms of the number of accurate predictions made by the model as shown in Equation 7.

$$Accuracy = \frac{((TP + TN))}{(TP + TN + FP + FN)} \tag{7}$$

Precision [63] is the evaluation parameter that is used for the analysis of the models in which precision identifies the frequency by which a classifier remains correct when we want to predict the positive class. Precision measures the positive rate of the model to the extent to which the model predicts the positive values and indicates the extent to which the model classifies the phishing URLs. The different classifiers performed well in terms of the precision.

$$Precision = \frac{TP}{(TP + FP)} \tag{8}$$

A metric used for the analysis of the classification models that answer out of all the likely positive labels, and how many times did the model accurately identify? To predict phishing and legitimate URLs, the classifier is correctly identified the classifier for both types of URLs [63].

$$Recall = \frac{TP}{(TP + FN)} \tag{9}$$

The F1 score is the harmonic mean of precision and recall, where the F1 score reaches its best value (perfect precision and recall). The general formula is as follows:

$$F1Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{10}$$

Therefore, an F1 score is required when inquiring about the balance between precision and recall.

IV. RESULTS AND DISCUSSION

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. 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

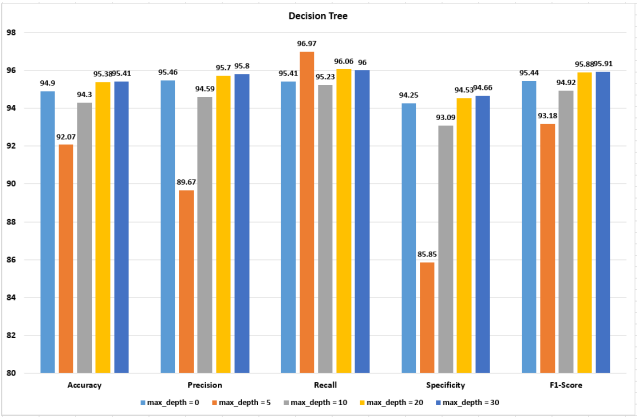

FIGURE 6. Experimental results of the decision tree model.

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%. 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 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

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

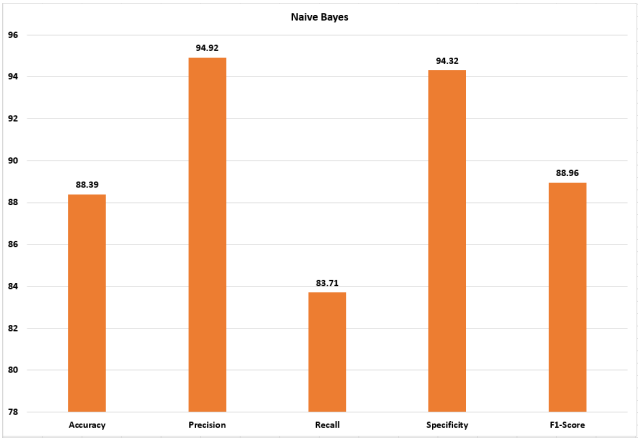

FIGURE 7. Experimental results of the naive bayes model.

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

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.

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

D. EXPERIMENTAL RESULTS OF K-NEIGHBORS CLASSIFIER

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.

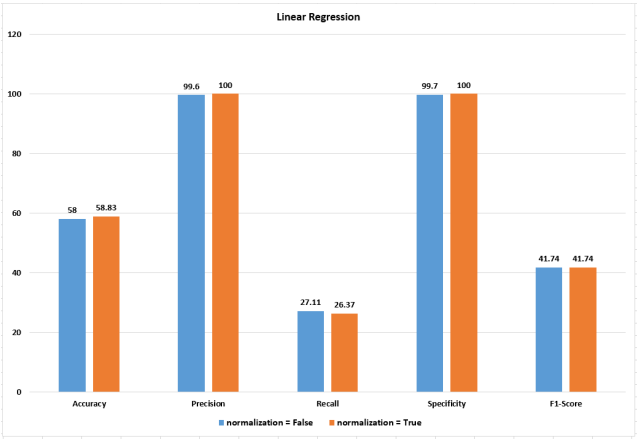

FIGURE 8. Experimental results of the linear regression model.

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

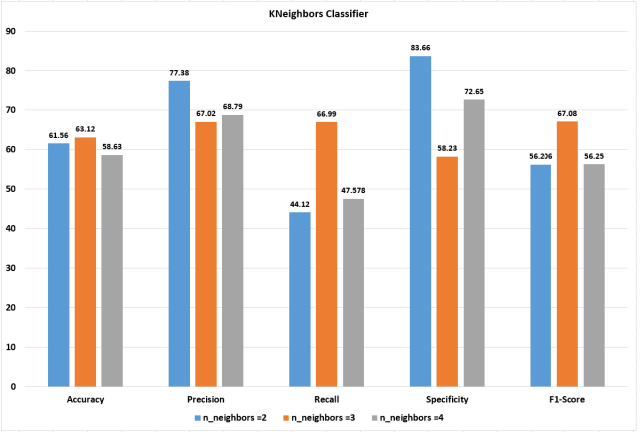

FIGURE 9. Experimental results of the K-neighbors classifier model.

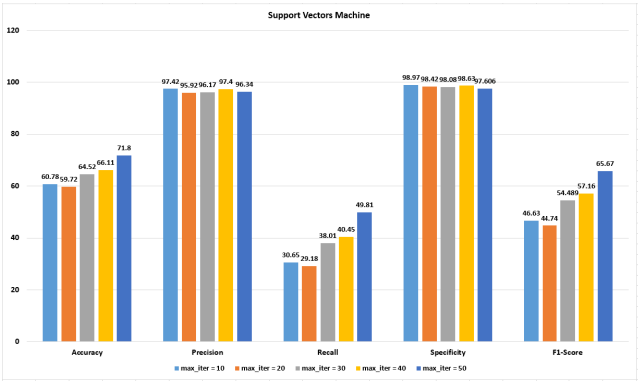

FIGURE 10. Experimental results of the support vector machine model.

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

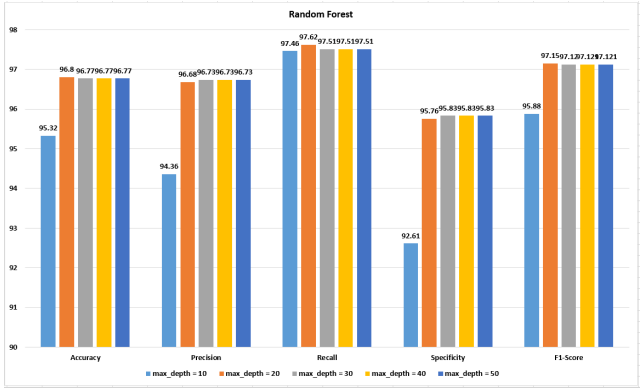

FIGURE 11. Experimental results of the random forest model.

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.

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-

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

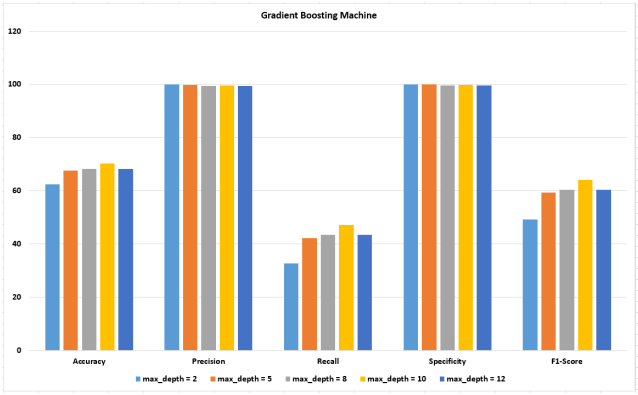

FIGURE 12. Experimental results of the gradient boosting model.

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

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.

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 machine-learning models and perform averaging operations on the results of each combined model. The Canopy 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.

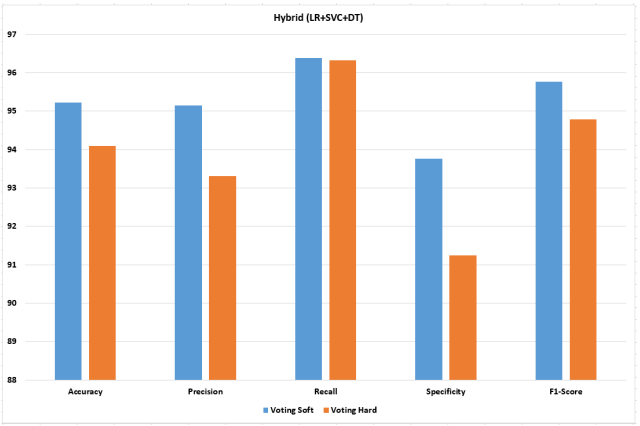

FIGURE 13. Experimental results of the hybrid (LR+SVC+DT) model.

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) model.

V. 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.

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.

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

TABLE 11. Results for the performance of the hybrid model LSD (LR+SVC+DT).

Literature	year	Techniques(Methods for detection)	performance
Proposed Approach	2022	Hybrid LSD model with Canopy feature selection	98.12% Base on the dataset consisted of 11054 number of records and 33 features extracted from 11000+ websites
[02]	2022	SVM and CNN-LSTM model used	Acquired best results with FPR 10^{-3}
[76]	2022	URL and HTML based feature using ANN model with MobileBERT	Shows highest 96% accuracy with 15 ANN model and with low feature shows 86% highest accuracy results
[73]	2022	Machine Learning	99.17%
[67]	2021	Extended support vector regression (X-SVR), Machine Learning	the accuracy, effectiveness, and computational efficiency of the proposed framework are fully verified. max. 86.01%
[74]	2021	Machine Learning	93.6%
[64]	2021	Concept of RNN, with ML technique	LURL has produced an average of 97.4% and 96.8% for Phishtank and Crawler datasets respectively. Reached an average of 93.8, 94.1, 96.7, and 93.6 for Phishtank and Crawler datasets.
[72]	2020	BLSTM classifiers	95.47%
[69]	2020	Machine Learning	6157 legitimate websites and 4898 phishing websites. Accuracy Min. max. 0.827 to 0.983
[66]	2019	Hybrid Ensemble Feature Selection (HEFS)	96.17%
[68]	2019	Random Forest algorithm with only NLP based features	97.98%
[65]	2019	recurrent convolutional neural networks (RCNN) model and THEMIS	accuracy of THEMIS reaches 99.848%
[70]	2018	Machine Learning	89.2%
[71]	2017	Machine Learning	80%

TABLE 12. Analyses in the form of proposed approach in terms of execution time, throughput and latency.

ML Models	Execution time	Throughput	Latency
Linear Regression	859.046	19.9884	0.0624
Decision Tree	599.3688	20.2412	0.04687
Random Forest	515.34198	24.44202	0.04687
Naive Bayes	379.8808	25.7395	0.0468
Support vector Classifier	949.0435	18.9884	0.06758
Gradient Boosting Classifier	657.8543	38.5674	0.076592
Ensemble Model LSD	456.3419	24.44202	0.04687
Proposed approach	370.9845	50.8565	0.0412

In addition, the conclusions of most studies are based on tiny datasets and cannot be applied to larger populations. Using large, this study presents a hybrid model for phishing detection prediction that overcomes these constraints and achieves a greater level of accuracy. The proposed method achieves 98.12% accuracy. Furthermore, the proposed LSD hybrid model using Grid search with cross fold validation and canopy feature selection to improve the accuracy of the voting techniques.

VI. 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

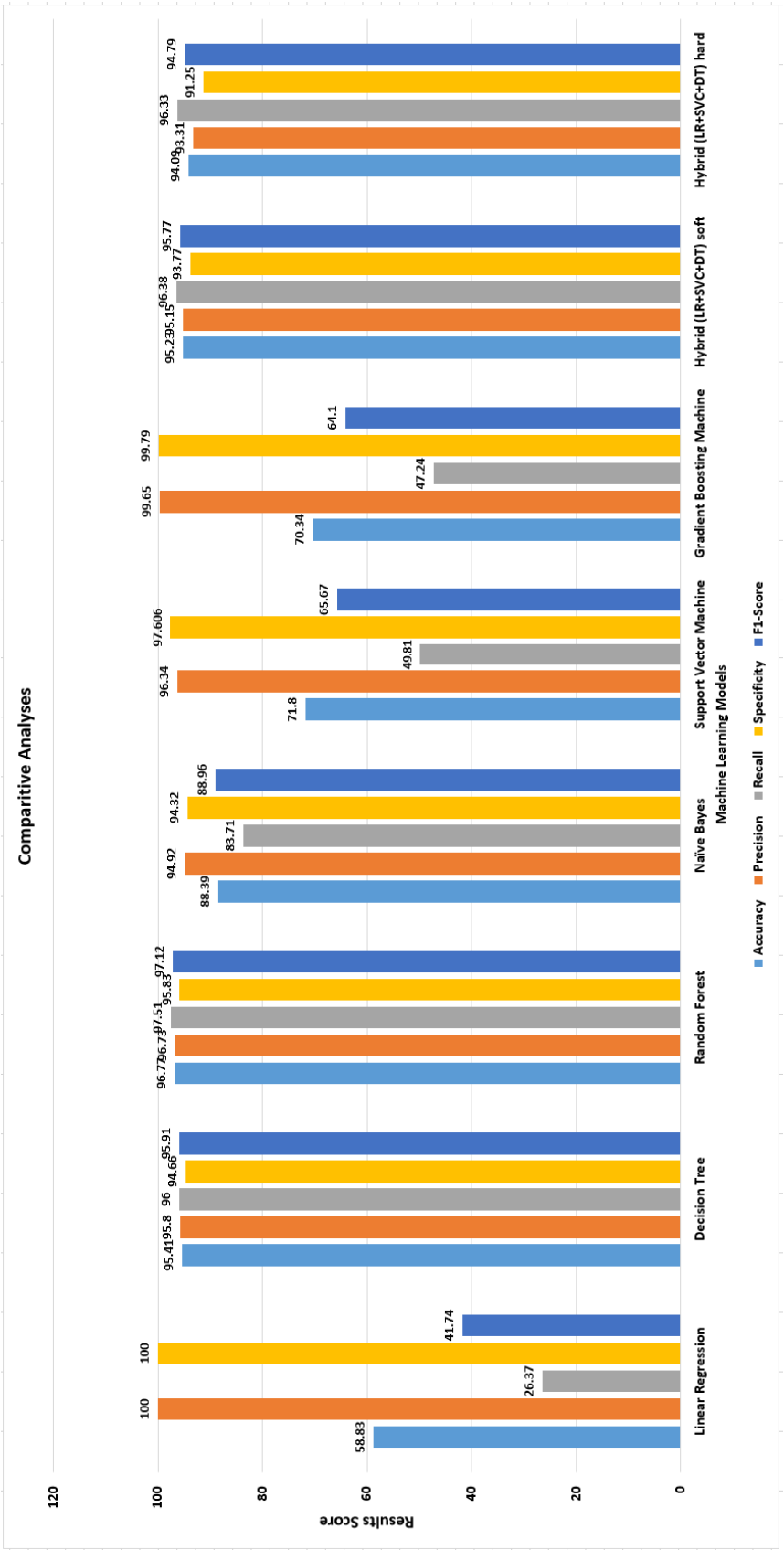

FIGURE 14. Comparative analyses of the experimental results of the proposed approach with applied machine models.

is proposed in this study. A dataset consisting of 32 URL attributes and more than 11054 URLs was extracted from 11000+ websites. This dataset was 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.

ACKNOWLEDGMENT

(Khabib Mustofa contributed equally to this work.)

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.

[14] R. Ø. Skotnes, "Management commitment and awareness creation—ICT safety and security in electric power supply network companies," *Inf. Comput. Secur.*, vol. 23, no. 3, pp. 302–316, Jul. 2015.

[15] R. Prasad and V. Rohokale, "Cyber threats and attack overview," in *Cyber Security: The Lifeline of Information and Communication Technology*. Cham, Switzerland: Springer, 2020, pp. 15–31.

[16] T. Nathezhtha, D. Sangeetha, and V. Vaidehi, "WC-PAD: Web crawling based phishing attack detection," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2019, pp. 1–6.

[17] R. Jenni and S. Shankar, "Review of various methods for phishing detection," *EAI Endorsed Trans. Energy Web*, vol. 5, no. 20, Sep. 2018, Art. no. 155746.

[18] (2020). Accessed: Jan. 2020. [Online]. Available: https://catches-of-the-month-phishing-scams-for-january-2020

[19] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank," in *Proc. Australas. Comput. Sci. Week Multiconf. (ACSW)*, Melbourne, VIC, Australia. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–11, Art. no. 3, doi: 10.1145/3373017.3373020.

[20] A. K. Jain and B. Gupta, "PHISH-SAFE: URL features-based phishing detection system using machine learning," in *Cyber Security*. Switzerland: Springer, 2018, pp. 467–474.

[21] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in *Proc. 4th ACM Workshop Digit. Identity Manage.*, Oct. 2008, pp. 51–60.

[22] G. Diksha and J. A. Kumar, "Mobile phishing attacks and defence mechanisms: State of art and open research challenges," *Comput. Secur.*, vol. 73, pp. 519–544, Mar. 2018.

[23] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2091–2121, 4th Quart, 2013.

[24] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2010, pp. 373–382.

[25] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "PhishNet: Predictive blacklisting to detect phishing attacks," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–5.

[26] P. K. Sandhu and S. Singla, "Google safe browsing-web security," in *Proc. IJCSET*, vol. 5, 2015, pp. 283–287.

[27] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, Mar. 2008, pp. 840–843.

[28] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in *Proc. 6th Conf. Email Anti-Spam (CEAS)*, Mountain View, CA, USA. Pittsburgh, PA, USA: Carnegie Mellon Univ., Engineering and Public Policy, Jul. 2009.

[29] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: A content-based approach to detecting phishing web sites," in *Proc. 16th Int. Conf. World Wide Web*, May 2007, pp. 639–648.

[30] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A feature-rich machine learning framework for detecting phishing web sites," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, pp. 1–28, Sep. 2011.

[31] C. L. Tan, K. L. Chiew, K. Wong, and S. N. Sze, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder," *Decis. Support Syst.*, vol. 88, pp. 18–27, Aug. 2016.

[32] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 191–195.

[33] R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 324–335, Jan. 2013.

[34] S. C. Jeeva and E. B. Rajsingh, "Intelligent phishing URL detection using association rule mining," *Hum.-Centric Comput. Inf. Sci.*, vol. 6, no. 10, pp. 1–19, 2016.

[35] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," *Soft Comput.*, vol. 23, no. 12, pp. 4315–4327, Jun. 2019.

[36] E. Buber, B. Diri, and O. K. Sahingoz, "Detecting phishing attacks from URL by using NLP techniques," in *Proc. Int. Conf. Comput. Sci. Eng. (UBMK)*, Oct. 2017, pp. 337–342.

[37] E. Buber, B. Diri, and O. K. Sahingoz, "NLP based phishing attack detection from URLs," in *Proc. Int. Conf. Intell. Syst. Design Appl.* Cham, Switzerland: Springer, 2017, pp. 608–618.

[38] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, Aug. 2014.

[39] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han, and J. Wang, "The application of a novel neural network in the detection of phishing websites," *J. Ambient Intell. Hum. Comput.*, vol. 14, pp. 1–15, Apr. 2018.

[40] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," *Decis. Support Syst.*, vol. 107, pp. 88–102, Mar. 2018.

[41] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Comput. Appl.*, vol. 31, no. 8, pp. 3851–3873, Aug. 2019.
[42] T. Peng, I. Harris, and Y. Sawa, "Detecting phishing attacks using natural language processing and machine learning," in *Proc. IEEE 12th Int. Conf. Semantic Comput. (ICSC)*, Jan. 2018, pp. 300–301.
[43] G. A. Jaafar, S. M. Abdullah, and S. Ismail, "Review of recent detection methods for HTTP DDoS attack," *J. Comput. Netw. Commun.*, vol. 2019, pp. 1–10, Jan. 2019.
[44] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," *Telecommun. Syst.*, vol. 68, no. 4, pp. 687–700, Aug. 2018.
[45] M. N. Anyanwu and S. G. Shiva, "Comparative analysis of serial decision tree classification algorithms," *Int. J. Comput. Sci. Secur.*, vol. 3, no. 3, pp. 230–240, 2009.
[46] L. Troiano and G. Scibelli, "A time-efficient breadth-first level-wise lattice-traversal algorithm to discover rare itemsets," *Data Mining Knowl. Discovery*, vol. 28, no. 3, pp. 773–807, May 2014.
[47] J. M. Byers, M. E. Flatté, and D. J. Scalapino, "Influence of gap extrema on the tunneling conductance near an impurity in an anisotropic superconductor," *Phys. Rev. Lett.*, vol. 71, no. 20, pp. 3363–3366, Nov. 1993.
[48] T. N. Phyu, "Survey of classification techniques in data mining," in *Proc. Int. Multiconf. Eng. Comput. Scientists*, vol. 1, 2009, pp. 1–5.
[49] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, no. 3 pp. 61–74, Mar. 1999.
[50] S. Archana and K. Elangovan, "Survey of classification techniques in data mining," *Int. Journal of Comput. Sci. Mobile Appl.*, vol. 2, no. 2, pp. 65–71, 2014.
[51] M. Ribalta, R. Bejar, C. Mateu, and E. Rubión, "Machine learning solutions in sewer systems: A bibliometric analysis," *Urban Water J.*, vol. 20, no. 1, pp. 1–14, 2023, doi: 10.1080/1573062X.2022.2138460.
[52] V. Y. Kulkarni and P. K. Sinha, "Pruning of random forest classifiers: A survey and future directions," in *Proc. Int. Conf. Data Sci. Eng. (ICDSE)*, Jul. 2012, pp. 64–68.
[53] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[54] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.
[55] M. S. Alam and S. T. Vuong, "Random forest classification for detecting Android malware," in *Proc. IEEE Int. Conf. Green Comput. Commun. IEEE Internet Things IEEE Cyber, Phys. Social Comput.*, Aug. 2013, pp. 663–669.
[56] P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classier," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 105–112.
[57] G. Boone, "Concept features in re: Agent, an intelligent email agent," in *Proc. 2nd Int. Conf. Auto. Agents*, 1998, pp. 141–148.
[58] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1998, pp. 4–15.
[59] G. I. Webb, "Naïve Bayes," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2011, doi: 10.1007/978-0-387-30164-8_576.
[60] S. Tan, "An effective refinement strategy for KNN text classifier," *Expert Syst. Appl.*, vol. 30, no. 2, pp. 290–298, Feb. 2006.
[61] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002.
[62] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (KNN) approach for predicting economic events: Theoretical background," *Int. J. Eng. Res. Appl.*, vol. 3, pp. 605–610, Sep. 2013.
[63] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Syst. Appl.*, vol. 62, pp. 1–16, Nov. 2016.
[64] P. Flach and M. Kull, "Precision-recall-gain curves: PR analysis done right," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 838–846.
[65] A. K. Dutta, "Detecting phishing websites using machine learning technique," *PLoS ONE*, vol. 16, no. 10, Oct. 2021, Art. no. e0258361.
[66] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism," *IEEE Access*, vol. 7, pp. 56329–56340, 2019.
[67] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci.*, vol. 484, pp. 153–166, May 2019.
[68] Y. Feng, Q. Wang, D. Wu, Z. Luo, X. Chen, T. Zhang, and W. Gao, "Machine learning aided phase field method for fracture mechanics," *Int. J. Eng. Sci.*, vol. 169, Dec. 2021, Art. no. 103587.
[69] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, Mar. 2019.
[70] V. Shahrivari, M. M. Darabi, and M. Izadi, "Phishing detection using machine learning techniques," 2020, *arXiv:2009.11116*.
[71] S. Kumar, A. Faizan, A. Viinikainen, and T. Hamalainen, "MLSPD—Machine learning based spam and phishing detection," in *Computational Data and Social Networks* (Lecture Notes in Computer Science), vol. 11280, X. Chen, A. Sen, W. Li, and M. Thai, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-04648-4_43.
[72] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," in *Proc. Australas. Comput. Sci. Week Multiconf.*, Jan. 2017, pp. 1–8.
[73] S. Wang, S. Khan, C. Xu, S. Nazir, and A. Hafeez, "Deep learning-based efficient model development for phishing detection using random forest and BLSTM classifiers," *Complexity*, vol. 2020, pp. 1–7, Sep. 2020.
[74] S. D. Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," *Ann. Data Sci.*, vol. 10, pp. 1–26, Mar. 2022.
[75] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in *Proc. 30th USENIX Secur. Symp. (USENIX Security)*, 2021, pp. 3793–3810.
[76] H. Shirazia, K. Haynesb, and I. Raya, "Towards performance of NLP transformers on URL-based phishing detection for mobile devices," Int. Assoc. Sharing Knowl. Sustainability (IASKS), Tech. Rep., 2022.

ABDUL KARIM received the Ph.D. degree from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2022. His current research interest includes natural language processing (NLP), artificial intelligence, machine learning, deep learning, and computer vision.

MOBEEN SHAHROZ received the M.C.S. and M.S. degrees in computer science from the Department of Computer Science, Khawaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree in computer science with The Islamia University of Bahawalpur. His current research interests include the Internet of Things (IoT), artificial intelligence, data mining, natural language processing, machine learning, deep learning, and image classification.

KHABIB MUSTOFA received the Ph.D. degree from the Vienna University of Technology, in 2007. He is currently a Lecturer with the Department of Computer Science and Electronics, Universitas Gadjah Mada. His research interests include semantic web and ontology management, information extraction, data analytics, and web technology.

SAMIR BRAHIM BELHAOUARI received the master's degree in telecommunications and network from Institut Nationale Polytechnique of Toulouse, France, in 2000, and the Ph.D. degree in mathematics from the Federal Polytechnic School of Lausanne, Switzerland, in 2006. He is currently an Associate Professor with the Division of Information and Communication Technologies, College of Science and Engineering, Qatar Foundation, Hamad Bin Khalifa University (HBKU). During last years, he also holds several research and teaching positions with Innopolis University, Russia; Alfaisal University, Saudi Arabia; the University of Sharjah, United Arab Emirates; University Technology PETRONAS, Malaysia; and EPFL Federal Swiss School, Switzerland. His research interests include applied mathematics, statistics, and data analysis, artificial intelligence, and image and signal processing (biomedical, bioinformatics, and forecasting), due to both mathematics and computer science backgrounds.

S. RAMANA KUMAR JOGA was born in Visakhapatnam, India, in 1988. He received the M.Tech. degree in power system and automation from GITAM University, Visakhapatnam. He is currently working as an Assistant Professor with Electrical and Electronic Engineering Department, Dadi Institute of Engineering and Technology, Anakapalle, AP. His research interests include power quality monitoring, power quality improvement, signal processing, power system protection, and machine learning.