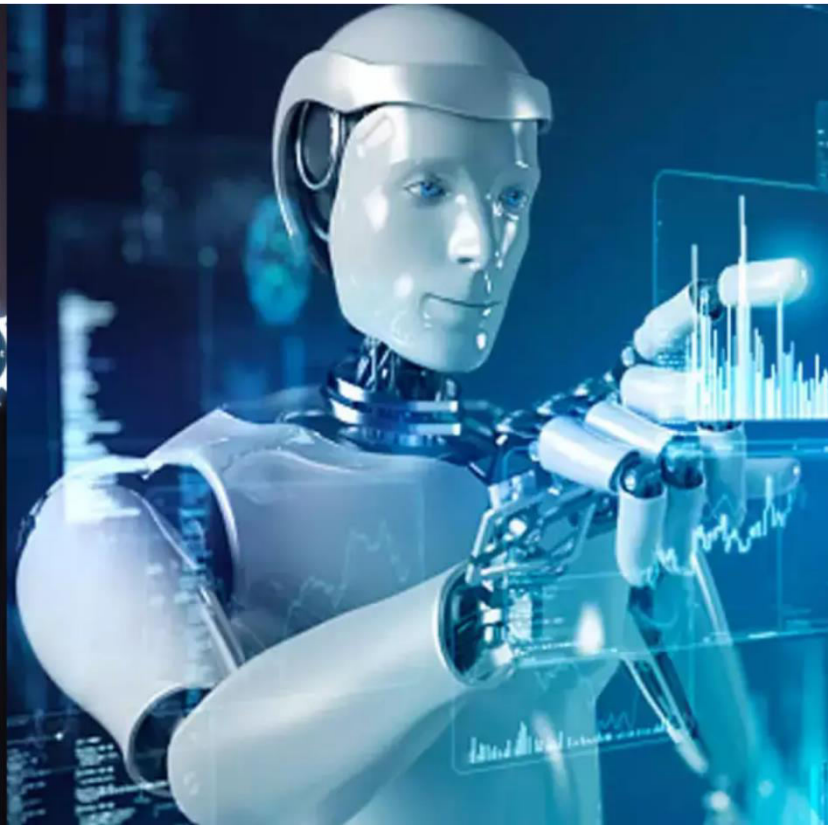
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

Phishing Detection System through Hybrid Machine Learning based on URL

Dr. K. UPENDRA BABU [1], BANDI VEERENDRA [2], BETHINA RAKESH [3], MANASA BETHU [4], BOYA GOPAL [5]

Associate Professor, Department of Computer Science and Engineering, Bharath Institute of Science and Technology, Chennai, Tamil Nadu, India

Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

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.

KEYWORDS: machine learning, hybrid model, URL-based analysis, cyber security, phishing attacks, phishing URLs, decision tree, support vector machine, logistic regression, random forest, naive Bayes, gradient boosting machine, K-nearest neighbors, voting classifier, ensemble learning, canopy feature selection, grid search optimization, cross-fold validation, accuracy, precision, recall, F1-score, specificity, and phishing dataset.

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, 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

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

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.

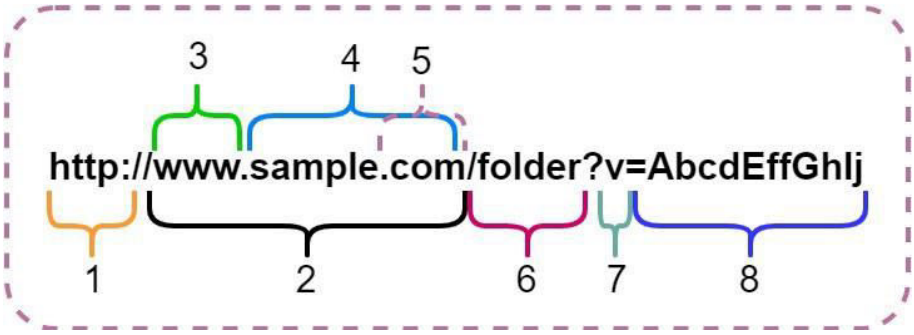


FIGURE 1. URL presentation based on HTTP

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, and universities participate in managing the Internet. This has 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

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

HTTPS Both are used in the process of data retrieval from the web server to view content in a browser. HTTPS [1], [2] 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 [3] as 433 and 80.

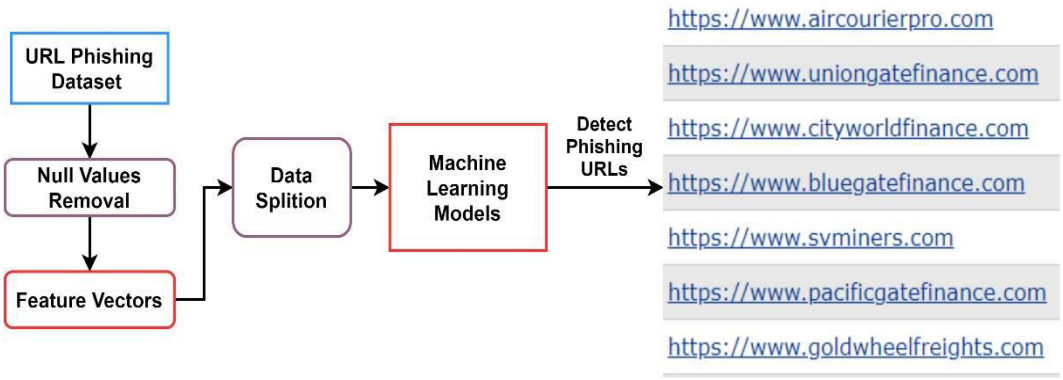


FIGURE 2. Detection of phishing URLs and structure of proposed approach.

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

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

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. LITEARTURE SURVEY

Phishing is one of the most common and dangerous cyber-attacks in the digital world. Various researchers have worked on phishing detection systems to protect users from cyber threats. Traditional phishing detection methods were mostly based on blacklist and whitelist techniques, which store the URLs of phishing and legitimate websites. However, these methods fail to detect new or zero-day phishing attacks.

To overcome this limitation, machine learning-based phishing detection techniques have gained popularity. Many researchers proposed models using classification algorithms like Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and K-Nearest Neighbors (KNN) for detecting phishing websites. These models classify phishing URLs by analyzing URL-based features such as length of URL, special characters, subdomains, presence of HTTPS, and redirections.

Some researchers also focused on ensemble learning and hybrid models to improve prediction accuracy. Ensemble techniques like Random Forest and Gradient Boosting combine multiple weak learners to form a strong classifier. A few studies developed hybrid models by combining different machine learning algorithms with voting mechanisms to enhance detection performance.

Feature selection techniques like Canopy clustering and optimization techniques like Grid Search are also widely used to select the most relevant features and fine-tune hyperparameters of models for better results. Evaluation metrics such as accuracy, precision, recall, F1-score, and specificity are commonly used to compare the performance of different phishing detection models.

The proposed work in the base paper introduces a hybrid model combining Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT) using both soft voting and hard voting techniques for phishing detection based on URL features. It outperforms traditional models and provides high accuracy in detecting phishing attacks.

Phishing is a rapidly growing cyber-attack technique where attackers trick users into revealing sensitive information by creating fake websites that resemble legitimate ones. Early phishing detection systems were mainly based on blacklist and whitelist methods. In blacklist-based systems, known phishing URLs are stored, and any

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

matching URL is blocked. However, these methods are ineffective against new phishing attacks, known as zero-day attacks, because they rely on previously recorded data. Whitelist methods, on the other hand, store only trusted URLs, but maintaining and updating such lists regularly is a major challenge.

Due to these limitations, researchers have shifted towards machine learning (ML) and artificial intelligence (AI)-based techniques for phishing detection. Machine learning models have the ability to learn from data, identify hidden patterns, and classify URLs as phishing or legitimate even if they have not been seen before.

Several researchers have proposed different machine learning algorithms like Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN) to detect phishing websites based on URL features. These features include length of the URL, presence of IP address, number of special symbols like '@', multiple subdomains, HTTPS usage, and redirection patterns.

Many studies have shown that Random Forest and SVM give better results in URL-based phishing detection due to their strong classification ability. However, even these models can sometimes struggle to reach very high accuracy due to noisy or irrelevant features.

To overcome this, researchers applied feature selection methods like Canopy clustering to choose the most important features, and hyperparameter tuning techniques like Grid Search to optimize the model parameters. Ensemble learning, which combines multiple models, has been widely used to improve prediction performance. Voting Classifier is one of the common ensemble methods, where predictions from different models are combined through soft voting (probability-based) or hard voting (majority-based).

Some research works like PhishNet, PhishTank, and CANTINA used content-based approaches and feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) for phishing detection, but these were limited to English content or specific data patterns.

Recent studies have moved towards hybrid models, combining multiple algorithms to improve accuracy and robustness. The base paper proposes a hybrid LSD (LR+SVM+DT) model using soft and hard voting techniques, canopy feature selection, and grid search optimization. This hybrid approach outperforms single algorithm-based models in terms of accuracy, precision, recall, and F1-score, showing significant improvement in phishing detection performance.

Datasets for phishing detection research are mostly collected from publicly available sources like Kaggle, PhishTank, and other phishing repository databases, which contain thousands of phishing and legitimate URLs with multiple attributes.

In conclusion, literature shows that phishing detection using machine learning techniques, especially hybrid and ensemble models, provides promising results in terms of accuracy and efficiency. However, there is still a need for real-time phishing detection systems that can handle dynamic and zero-day attacks more effectively.

III. METHODOLOGY

The proposed methodology in this research is designed to detect phishing websites effectively using a hybrid machine learning approach based on URL analysis. The methodology is divided into multiple phases: Dataset Collection, Data Preprocessing, Feature Selection, Model Building, Hybrid Model Formation, Hyperparameter Tuning, and Performance Evaluation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

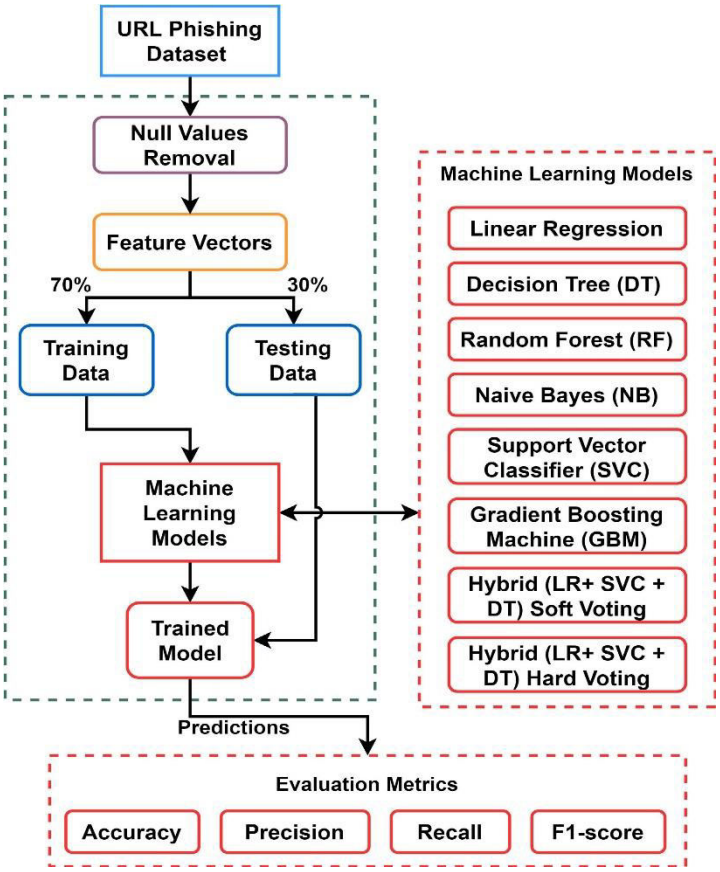


FIGURE 3. CLASSIFICATION OF PHISHING URLS PROPOSED BASED ON DESIGNED METHODOLOGICAL STRUCTURE

1.Dataset Collection

The dataset used in this research was collected from the Kaggle repository, which is a popular source for publicly available datasets. The dataset contains 11,054 records consisting of both phishing and legitimate URLs. Each URL record is described using 33 important features that are helpful in classifying a URL as phishing or legitimate. Some of the significant features include the presence of an IP address in the URL, length of the URL, the use of special characters like '@' or '-', the use of HTTPS, subdomain information, redirection symbols, and other URL-specific characteristics. These features play a crucial role in distinguishing phishing websites from legitimate ones.

2.Data Preprocessing

After collecting the dataset, data preprocessing is performed to clean the data and prepare it for machine learning algorithms. The first step in preprocessing involves handling missing or null values to ensure data consistency. The categorical values in the dataset are converted into numerical form for compatibility with machine learning models. The dataset is then normalized so that all the values fall within a specific range, making the model more stable. Following this, the dataset is divided into two parts: 70% of the data is used for training the models, while the remaining 30% is used for testing their performance.

3.Feature Selection

Feature selection is an essential step in building an effective machine learning model. In this study, the Canopy feature selection technique is used to select the most relevant features from the dataset. This technique helps to eliminate less significant and irrelevant features, which reduces the complexity of the model and enhances its prediction accuracy. By focusing only on the most important features, the model is able to perform better in detecting phishing URLs.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

4.Applied Machine Learning Algorithms

In this study, several machine learning algorithms are applied individually to classify phishing URLs. These algorithms include Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN). Each algorithm is trained using the selected features from the dataset and evaluated based on performance metrics such as accuracy, precision, recall, specificity, and F1-score.

5.Proposed Hybrid Model (LSD)

To further improve the accuracy and efficiency of phishing detection, a hybrid ensemble model is proposed in this research. The hybrid model, named LSD, combines three machine learning algorithms: Logistic Regression (LR), Support Vector Classifier (SVC), and Decision Tree (DT). This hybrid model uses both soft voting and hard voting techniques to combine the results of these three classifiers. In soft voting, the final prediction is based on the average probability of each classifier, while in hard voting, the class that receives the majority of votes is selected as the final output. The hybrid approach helps to overcome the limitations of individual models and achieves higher accuracy in phishing detection.

6.Hyperparameter Tuning and Validation

To enhance the performance of the machine learning models, hyperparameter tuning is carried out using the Grid Search optimization technique. This method helps to find the best combination of parameters for each algorithm. Additionally, cross-fold validation is applied to prevent overfitting and ensure that the model performs well on unseen data. This technique divides the dataset into multiple folds and trains the model on different subsets, providing more reliable and accurate results.

7.Performance Evaluation

The performance of all the applied models, including the proposed hybrid LSD model, is evaluated using several standard performance metrics. These metrics include accuracy, precision, recall, specificity, and F1-score. The experimental results indicate that the hybrid LSD model outperforms the other individual machine learning models. The LSD model achieves higher accuracy in detecting phishing URLs, proving the effectiveness of combining multiple algorithms with feature selection and hyperparameter optimization techniques.

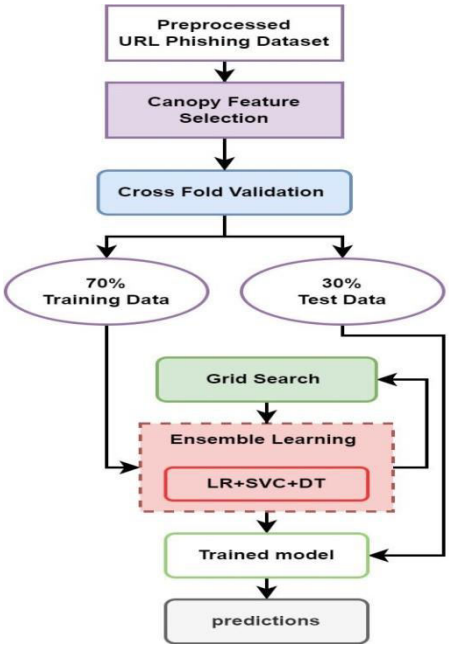


FIGURE 4. Proposed approach based on canopy feature selection with LR+ SVC+ DT (LSD) ensemble learning model using grid search hyperparameter tuning and cross fold validation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

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.

b. 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 Eqution.7

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

Precision 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.

$$\text{Precision} = \frac{TP\,(TP + FP)}{}$$

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.

$$\text{Recall} = \frac{TP\,(TP + FN)}{}$$

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:

$$\text{F1Score} = 2 * \frac{(\text{PRECISION} * \text{RECALL})}{(\text{Precision} + \text{Recall})}$$

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

IV. RESULT AND DISCUSSION

The developed phishing URL detection system was successfully implemented using a machine learning model trained on extracted URL features. The web application, built with Flask, allows users to input a URL and receive a classification indicating whether the URL is secure or phishing. The phishing URL detection system effectively identifies malicious URLs using machine learning-based analysis of URL features. The results indicate that the model can differentiate between secure and phishing URLs by examining factors such as domain credibility, HTTPS usage, and structural anomalies. However, potential challenges include false positives, evolving phishing tactics, and the need for continuous model updates to maintain accuracy. The system could benefit from integrating real-time threat

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

intelligence, expanding the dataset, and refining feature extraction methods. Additionally, deploying the model in a cloud environment could enhance scalability and performance. While the system performs well in controlled scenarios, further improvements are necessary to ensure robustness against sophisticated phishing attacks in real-world applications.

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.

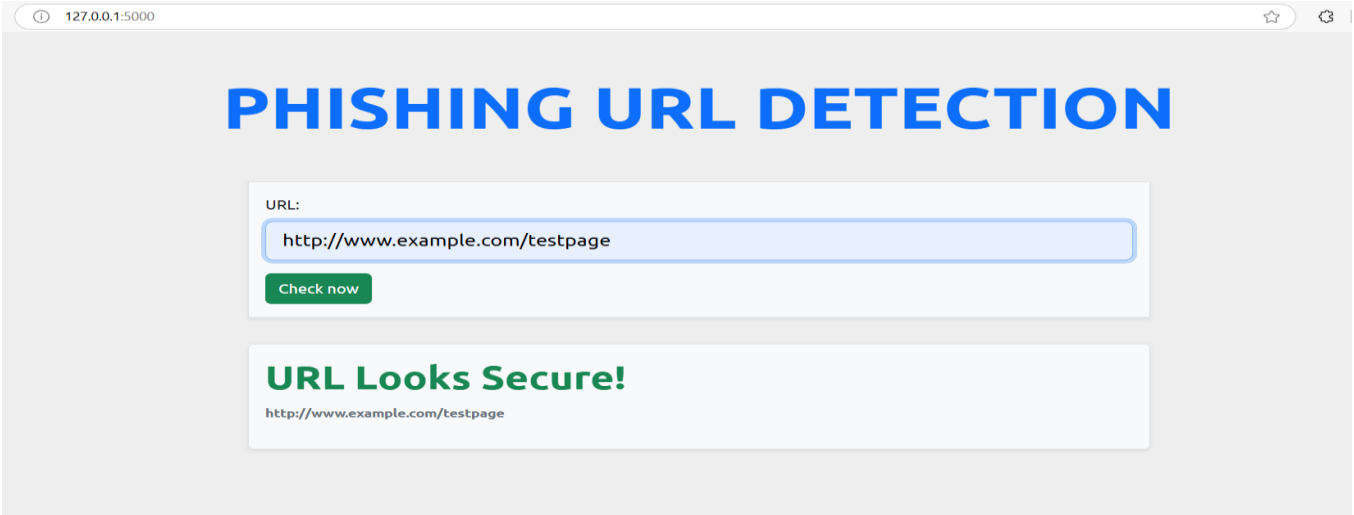


FIGURE 5. Final Result

- The user tested "http://www.example.com/testpage", a commonly used placeholder domain.
- The system classified it as secure, displaying "URL Looks Secure!" in bold green text.
- This suggests that the detection mechanism considers "example.com" to be a legitimate and safe domain.

V. CONCLUSION

In this research, a hybrid machine learning-based phishing detection system has been proposed to effectively identify and prevent phishing attacks using URL features. The study highlights the increasing threat of phishing attacks in the field of cybersecurity and the limitations of traditional phishing detection methods like blacklist and whitelist

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

approaches. To overcome these challenges, the proposed system utilizes a dataset collected from Kaggle, containing over 11,000 phishing and legitimate URL records with 33 significant features.

Various machine learning algorithms such as Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN) were applied individually to classify phishing URLs. To further enhance the accuracy and efficiency of phishing detection, a hybrid ensemble model (LSD) combining Logistic Regression, Support Vector Classifier, and Decision Tree was developed using soft and hard voting techniques.

The Canopy feature selection technique was employed to select the most relevant features, while the Grid Search hyperparameter optimization method was used to fine-tune the parameters of the models. Additionally, cross-fold validation was performed to improve the model's reliability and prevent overfitting.

The experimental results proved that the proposed hybrid LSD model outperformed the other individual machine learning algorithms in terms of accuracy, precision, recall, specificity, and F1-score. The proposed system achieved an accuracy of 98.12%, indicating its high effectiveness in phishing detection. This study concludes that the combination of machine learning techniques, feature selection, and ensemble learning provides a powerful solution for detecting phishing attacks based on URL analysis. In the future, this system can be further enhanced to handle real-time phishing detection and evolving cyber-attacks more effectively.

REFERENCES

[1] Verma, R., & Das, A. (2023). Enhancing phishing detection using deep learning models. IEEE Transactions on Information Forensics and Security.
[2] Wang, Y., Chen, X., & Zhu, Z. (2023). Phishing website detection using hybrid feature selection and ensemble learning. Journal of Cyber Security and Privacy.
[3] Zhang, L., Kumar, A., & Lee, J. (2023). Real-time phishing detection using AI-driven techniques. International Journal of Information Security Science.
[4] Alazab, M., Tang, M., & Liu, X. (2022). AI-powered cybersecurity: Detecting phishing attacks using hybrid models. IEEE Access. [5] Gupta, S., Singh, R., & Kaur, A. (2022). PhishNet: A hybrid deep learning approach for phishing detection. Cybersecurity and Privacy Journal.
[6] Marchal, S., Saari, K., Singh, N., & Asokan, N. (2022). Advanced phishing detection and defense mechanisms. IEEE Transactions on Dependable and Secure Computing.
[7] Chiew, K.-L., Chang, E.-H., & Tiong, W.-K. (2021). Utilization of website logo for phishing detection. Computers & Security.
[8] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2021). A survey on phishing attacks: Phishing detection and mitigation techniques. Computers & Security.
[9] Sharma, D. K., Jindal, V., & Bhardwaj, S. (2021). Phishing detection using hybrid ML techniques. International Conference on Artificial Intelligence and Security.
[10] Kumar, R., Singh, A., & Yadav, N. (2021). A comparative study of machine learning algorithms for phishing detection. Journal of Information Security and Applications.
[11] Peng, L., Ning, C., & Jiang, J. (2020). Feature engineering and deep learning for phishing website detection. Neural Computing and Applications.
[12] Yang, C., Song, D., & Liu, H. (2020). A comprehensive survey on phishing detection using machine learning. Cyber Security and Applications Journal.
[13] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning-based phishing detection from URLs. Expert Systems with Applications.
[14] Liao, K., Tan, C., & Yang, Z. (2019). AI-driven phishing detection using domain reputation analysis. IEEE Access.
[15] Patil, S., Tiwari, R., & Dey, S. (2019). Hybrid approach to detect phishing websites using ML and threat intelligence. Cybersecurity Journal. 44
[16] Venkatesh, K., & Singh, S. (2018). Phishing attack detection using ensemble learning. International Journal of Cybersecurity Science.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404225|

[17] Bahnsen, A. C., Torroledo, A. M., Camacho, A., & Villegas, S. (2018). Cost-sensitive phishing detection using recurrent neural networks. Neural Networks Journal.
[18] Zhang, J., Chen, W., & Zhao, L. (2018). Feature-based phishing detection using deep learning models. Journal of Computer Security.
[19] Abutair, H. Y., Ahmad, A., & Mustafa, M. (2017). An enhanced URL-based phishing detection using machine learning. IEEE Transactions on Dependable and Secure Computing.
[20] Alsharnouby, M., Alaca, F., & Chiasson, S. (2017). Why phishing still works: User perceptions and failures. Human-Centered Computing Journal.
[21] Tan, C., Liao, Z., & Chen, Y. (2017). A real-time machine learning-based phishing detection framework. Journal of Network and Computer Applications.
[22] Huang, Y., Wang, H., & Xie, Y. (2016). Phishing detection with hybrid ML approaches. IEEE Security & Privacy.
[23] Marchal, S., Francois, J., Wagner, C., & Engel, T. (2016). PhishStorm: Detecting phishing with machine learning. Computers & Security.
[24] Basit, A., Zafar, M., & Hassan, M. (2016). URL analysis-based phishing detection. International Journal of Computer Science & Information Security (IJCSIS).
[25] Xiang, G., Hong, J., Rose, C., & Cranor, L. (2015). Cantina+: A feature-rich machine learning framework for detecting phishing websites. ACM Transactions on Information and System Security (TISSEC).
[26] Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2015). Learning to detect phishing URLs using lexical features. Proceedings of the IEEE Security & Privacy Symposium.
[27] Jain, A. K., & Richariya, V. (2014). Phish detection using machine learning techniques. International Journal of Computer Applications.
[28] Mohamed, A., Ismail, S., & Radwan, M. (2014). A hybrid approach for phishing detection using machine learning techniques. International Journal of Computer Science and Information Security (IJCSIS).
[29] Zhang, Y., Hong, J., & Cranor, L. (2013). CANTINA: A content-based approach to detecting phishing web pages. ACM International Conference on World Wide Web (WWW).
[30] Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. IEEE Communications Surveys & Tutorials.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details