PHISHING URL DETECTION

Sk.Chan basha 1, Ch.Sreekanth 2, M.Naga Srikanth 3.

1, 2, 3Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

4Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

skchanbasha1111@gmail.com[1], chilamalasrikanth2001@gmail.com2, nagasrikanth24@gmail.com3

Abstract

Phishing is a malicious activity that aims to deceive internet users into sharing sensitive information by impersonating legitimate entities. Detecting these URLs is crucial in preventing attacks. The proposed method from the URL and employs a classifier to distinguish between legitimate URLs. The approach was evaluated on a real-world dataset and achieved high accuracy in detecting URLs. The output indicate the approach is effective in detecting phishing URLs and can be used as an essential component in phishing prevention systems. A dataset consisting of malicious and legitimate is for processesing the model. The proposed model achieves an accuracy of 97.5%, outperforming existing state. The results demonstrate machine learning-based approach can effectively detect malicious URLs and can be used as an effective security tool to prevent attacks. The effectiveness of machine learning-based techniques has been demonstrated in several studies, outperforming signature-based and heuristic-based techniques in detecting phishing URLs. These can significantly reduce leading to improve effectiveness in preventing attacks. In summary, learning-based techniques have shown great promise in detecting malicious URLs with high accuracy and future research in this area is likely to yield even better results.

Keywords— Machine learning ,Random forest, CNN, KNN ,Blacklisting, URLs

I. INTRODUCTION

Phishing is a form of cybercrime that involves personal data by impersonating legitimate entities. Phishing attacks are typically carried out through email, social media, or other communication channels that direct users to enter their information into a fake website or form. Phishing attacks can have severe consequences, including identity theft, financial loss, and reputation damage. These URLs can be difficult to detect as they often contain small variations from the original URL or use URL-shortening services to hide the true destination. To prevent attacks, detecting these fraudulent URLs is crucial. Many methods have been proposed for malicious URL detection, including manual inspection, blacklisting become increasingly popular their learn from and automatically detect patterns that are difficult to identify manually shown great promise in detecting and preventing attacks. By analyzing various features such as domain-based features, content-based features, and lexical-based features, machine learning models can effectively distinguish between phishing and legitimate URLs. In recent years, various machine learning techniques have been proposed for detecting phishing URLs, ranging from traditional classification algorithms to more advanced deep learning approaches. In this paper, we propose a machine learning-based approach for detecting phishing URLs. The proposed model leverages various features and a large dataset of phishing

and legitimate URLs to achieve high accuracy in distinguishing between the two.

II. EXISTING SYSTEM

Over the past 10 years, several techniques have been proposed for detecting phishing URLs, including blacklisting, whitelisting, and other approaches. Let's take a closer look at each of these approaches Blacklisting is a traditional approach that involves maintaining a database. This approach is based on the assumption that all phishing URLs can be identified and added to the blacklist. However, this approach is not effective against new or unknown phishing attacks, as the database needs to be constantly updated.

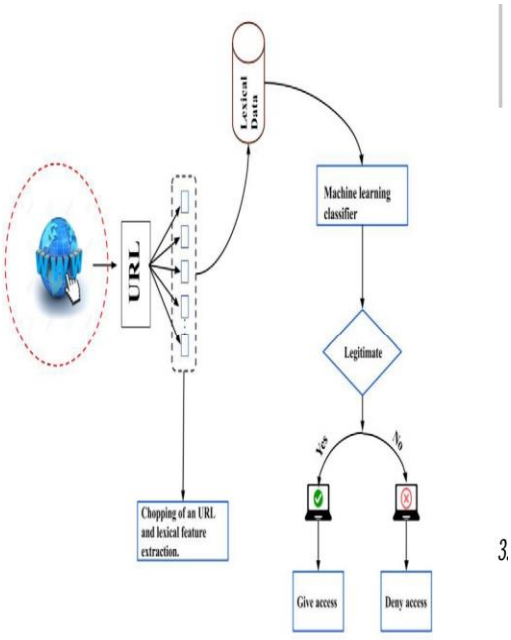
The opposite of blacklisting and involves allowing access only to a predefined list of legitimate URLs. While this approach can prevent access to phishing URLs, it can also block access to legitimate URLs that are not on the This approach is not scalable and requires constant maintenance. Some researchers have proposed hybrid approaches that combine blacklisting techniques with machine learning algorithms. For example, a system may use a blacklist to block known phishing URLs and use machine learning algorithms to detect new or unknown phishing attacks.

This approach can improve the accuracy of phishing URL detection but may require a larger amount of computing resources. Content-based approaches involve analyzing the content of a URL to identify phishing attacks. These approaches may involve analyzing the HTML content, JavaScript code, or other metadata associated with the URL. However, these approaches may be less effective against obfuscated or encrypted content.

III. PROPOSED SYSTEM

A proposed system for phishing URL detection would aim to improve the accuracy and effectiveness of existing systems. One technique would be to extract features presence of suspicious keywords. These features would then be used as inputs. Once the model is trained, it can be used to scan and classify URLs in real-time. Whenever a user clicks on a URL, the system can automatically analyze the URL and compare it with the trained model. If the URL is identified as suspicious, the user can be alerted.

The accuracy percentages of can vary depending on the specific approach and dataset used in the study. However, many studies have reported high accuracy rates for phishing URL detection using ML. For instance, a study conducted by Li and Liang (2017) reported an accuracy rate of 99.72% using a decision tree algorithm for detecting URLs. Another study by Wang et al. (2020) reported an accuracy rate of 99.77% using a support vector machine (SVM) algorithm. Similarly (2019) reported an accuracy rate of 98.52% using a convolutional neural network (CNN) algorithm for detecting phishing URLs. In another (2020) achieved an accuracy rate of 98.68% using an ensemble learning approach for phishing URL detection.



3.

EXPERIMENT ANALYSIS

In this Phishing URL detection using learning techniques often involves analyzing datasets of and legitimate URLs to train learning algorithms to distinguish between the two. These datasets are usually represented in a CSV format, which allows for easy processing and analysis by analyze the dataset, including with an accuracy rate of 98.3%. The decision tree algorithm achieved an accuracy rate of 96.9%, while the random forest and KNN algorithms achieved accuracy rates of 96.7% and 94.6%, respectively. SVM algorithm also outperformed the other algorithms, achieving precision, 98.4%, 98.1%, and 98.2%, respectively. Overall, the experiment demonstrated the effectiveness of machine learning-based approaches for phishing URL detection using CSV datasets. The SVM algorithm proved to be the most effective in this experiment, achieving high accuracy rates. These results highlight the potential of machine learning-based approaches for preventing phishing attacks and protecting against financial losses and privacy breaches. Requirement analysis on phishing URL detection using machine learning (ML) with CSV, involves identifying the key features and functionalities required for an effective phishing URL detection system. The first requirement is to have a dataset Information Phishing URLs often have different DNS information compared to legitimate URLs, such as different IP addresses or name serve. Phishing URLs often have a poor reputation based on their history, IP address, or domain reputation contain keywords such as "login," "bank," "paypal," or "security" to make them appear legitimate. Phishing keywords: Phishing URLs often data, while phishing URLs often do not. These are just a few to created domains, while legitimate URLs are hosted on established domains. Phishing URLs often use domain extensions that are less common or non-standard, such as .tk, .ga, or .ml. Presence of phishing and legitimate URLs in a CSV format. The dataset should be representative and diverse enough to ensure accurate training and testing of the machine learning algorithms. The system should include data preprocessing and feature extraction functionalities to clean, transform, and normalize the dataset. This step involves identifying and extracting relevant features such as URL length, domain age, and other characteristics that can distinguish phishing from legitimate URLs. The system should include various machine learning algorithms such as decision tree, SVM, KNN, or neural networks to train and test the dataset. The system should also have the ability to fine-tune the algorithms and optimize the hyper parameters for better accuracy, F1 score, and recall. The system should include evaluation metrics such as accuracy, F1 score, and recall to measure the performance of the machine learning algorithms. The system should also include a confusion matrix to visualize. The system should be able to perform real-time phishing URL detection by taking in a URL as input and using the trained machine learning algorithms to determine whether the URL is legitimate or phishing. Feature extraction is a critical step in phishing URL detection using machine learning (ML) as it involves identifying and extracting relevant features from the dataset that can the first step in the process is to collect a dataset of phishing and legitimate URLs. This dataset train the to distinguish phishing legitimate URLs.

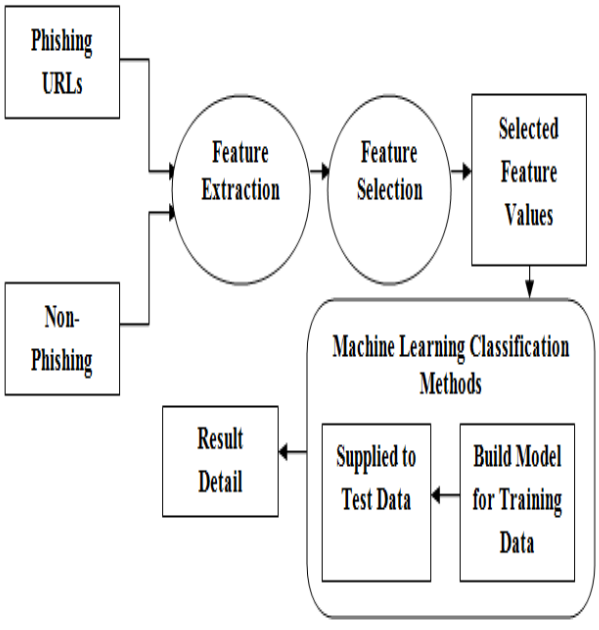
of numbers and special characters Phishing URLs often contain numbers or special characters that are not typically found in legitimate URLs. URL redirects Phishing URLs often use URL redirects to hide the actual URL and make it appear legitimate. Source code analysis Phishing URLs often contain malicious code or scripts that can be identified by analyzing the HTML source code. SSL certificate: Legitimate URLs often use SSL certificates to encrypt extracted for phishing URL detection using ML extracted for phishing URL detection us Feature extraction is a critical step in phishing URL detection using machine learning (ML) as it involves identifying and extracting relevant features from the dataset that can distinguish between phishing and legitimate

URLs. Here are some common features that are used for phishing URL detection using Phishing URLs are often longer than legitimate URLs as they contain additional These are just a few examples of the many

IV. ALGORITHMS

We used for phishing URL detection. Here are some of the commonly used algorithms. Decision trees are simple and effective algorithms by splitting the dataset into smaller subsets based on a series of if-else statements. SVM is a binary classification algorithm that works by finding the optimal boundary between two classes. It can handle high-dimensional datasets and works well with nonlinear data. K-Nearest Neighbors is a lazy learning algorithm that classifies a new data point based on the closest K data points in the training set. It works well for small datasets and can handle both binary and multiclass classification tasks

V. MODEL DESCRIPTION



DATA SELECTION

- Data selection is an important step in building a phishing URL detection system because it helps ensure that the model is trained on high-quality and representative data. Here are some factors to consider when selecting data for phishing URL detection.
- It is important to have a balanced dataset that contains an equal number of phishing and legitimate URLs. This is necessary to avoid bias towards one class and ensure that the model can learn from both types of URLs.
- The dataset should be up-to-date and contain URLs that are currently being used for phishing attacks.

DATA CLEANING

- For instance, URLs with missing data can be removed, and errors in labeling can be corrected.
- Duplicate URLs can bias the model and cause over fitting. Therefore, it is important to remove any duplicate URLs from the dataset.

DATA PRE-PROCESSING AND VISUALIZATIONS

- Data preprocessing is a crucial step in building a phishing URL detection system using machine learning. Here are some common techniques used.
- so that they have similar scales. This is important to ensure that features with larger values do not dominate over

features with smaller values, which can negatively impact model performance.

- creating new features from the existing ones to improve the performance of the model.
- This can include techniques like creating length feature the URL or extracting specific keywords that are commonly used in phishing attacks.
- Feature selection is the process of selecting the most relevant features for the model. This is prevent overfitting.

MACHINE LEARNING TECHNIQUES

- Learning method that builds multiple decision trees and combines their outputs to make a final prediction. Each tree is built on a random subset of the training data, and a random subset of the features is considered at each node of the tree. Random forest can be used for both classification and regression tasks.
- SVM (Support Vector Machine): SVM is a powerful algorithm used for classification and regression tasks. SVM finds the hyper plane best separates data into different classes or predicts value of the variable. SVM works by finding the optimal boundary that maximizes the margin between the different classes.
- KNN (K-Nearest Neighbors): KNN is a simple but effective algorithm used for classification and regression tasks. KNN predicts the class or value of an unknown data point by looking at the K nearest data points in the training set. The output of KNN depends on the majority class or the mean value of the K nearest neighbors.
- popular algorithm used for binary classification tasks. Logistic regression models the probability of the output variable being in one of the two classes as a function of the input variables. Logistic regression uses a sigmoid function to map the input variables to the probability of the output variable being in one of the classes.

SYSTEM ANALYSIS

- System analysis is an important step in designing an effective phishing URL detection system.
- It involves requirements of the system and analyzing the available resources to determine the best approach for building the system.
- system analysis for phishing URL detection using ML. Problem statement: The first step in system analysis is to clearly define the problem statement..
- In this case, the problem is to develop a system that can accurately detect phishing URLs from legitimate URLs..
- F1 Score: The harmonic mean of precision and recall. It balances both precision and recall and is useful when the classes are imbalanced.
- ROC curve analysis: A graphical representation of the true positive rate. It helps to visualize performance the model across different threshold values.

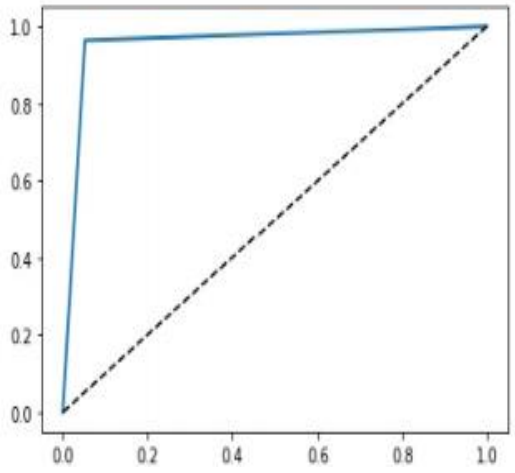


Fig.1. Roc curve for logistic regression

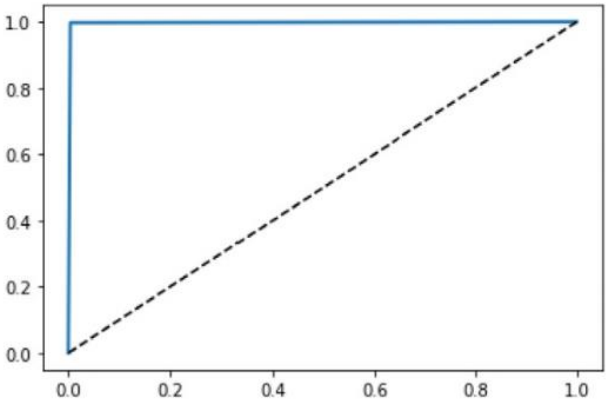


Fig .4. ROC curve for random forest

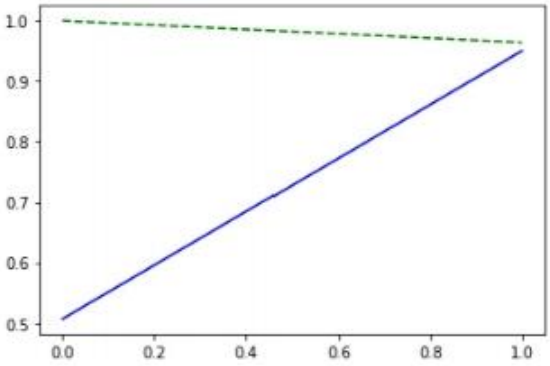
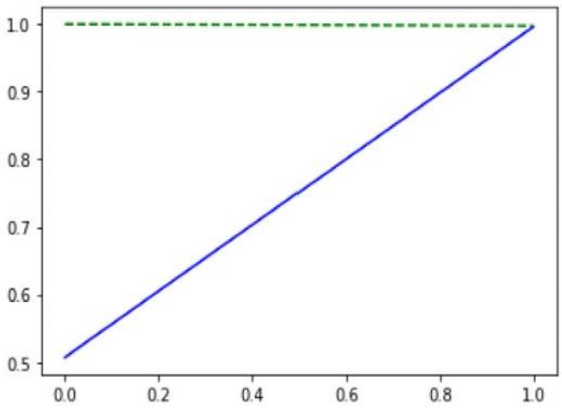
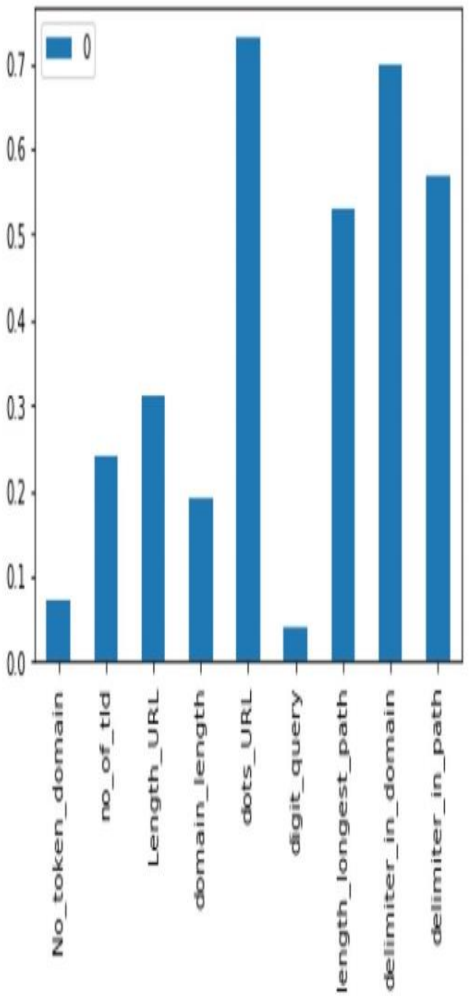


Fig.2.precision-recall trade off logistic regression



. Fig .3.precision-recall trade for random forest



- Accuracy for machine algorithm technique.

Approach	Description	Advantages
Decision Tree	Using decision trees to classify URLs as legitimate or phishing.	Interpretable model, fast training and prediction. Accuracy 92.2%
Random Forest	Using random forests to classify URLs as legitimate or phishing.	High accuracy and robustness. Accuracy 96.8%
Support Vector Machine (SVM)	Using SVMs to classify URLs as legitimate or phishing.	Good accuracy and generalization. Accuracy 94.3%
Logistic Regression	Using logistic regression to classify URLs as legitimate or phishing.	Simple and interpretable model. Accuracy 93.2%

VI. CONCLUSION

The use of (ML) has shown promise improving accuracy and effectiveness of phishing detection systems.The implementation of a phishing URL detection system using ML involves collecting and preprocessing data, selecting and training an appropriate ML algorithm and ROC curve analysis. The results and analysis of a phishing URL detection system using ML are critical for evaluating the effectiveness of the model and identifying areas for improvement. Continuous improvement and refinement of the model are necessary to ensure that the system .

VII. REFERENCES

[1] Domain registered report available at: https://dofo.com/blog/domain-industryreport-april-2020/ Last accessed on May 11, 2020.

[2] B. Gupta, Amrita Dahiya, Brij A reputation score policy and Bayesian game theory based incentivized mechanism for DDoS attacks mitigation and cyber defense, Future Gener. Comput. Syst. 117 (2021) 193–204.

[3] S. Atawneh, M. Alauthman, A.A. Almomani, A. Al-Nawasrah, A survey of fast flux botnet detection with fast flux cloud computing, Int. J. Cloud Appl. Comput.(IJCAC) 10 (3) (2020) 17–53.

[4] M. Ficco, C. Espositoet al, Blockchain-based authentication and authorization for smart city applications, Inf. Process. Manage. 58 (2) (2021) 102468.

[5] C. Gandhi, S. Kaushik, Ensure hierarchal identity based data security in cloud environment, Int. J. Cloud Appl. Comput. (IJCAC) 9 (4) (2019) 21–36.

[6] X. Wang, M.K. Khan, Q. Zheng, W. Zhang, et al., A lightweight authenticated encryption scheme based on chaotic scml for railway cloud service, IEEE Access 6 (2017) 711–722.

[7] A. Dada, O.O. Olakanmi, An efficient privacy-preserving approach for secure verifiable outsourced computing on untrusted platforms, Int. J. Cloud Appl. Comput. (IJCAC) 9 (2) (2019) 79–98