Phishing Website Detection using Machine Learning

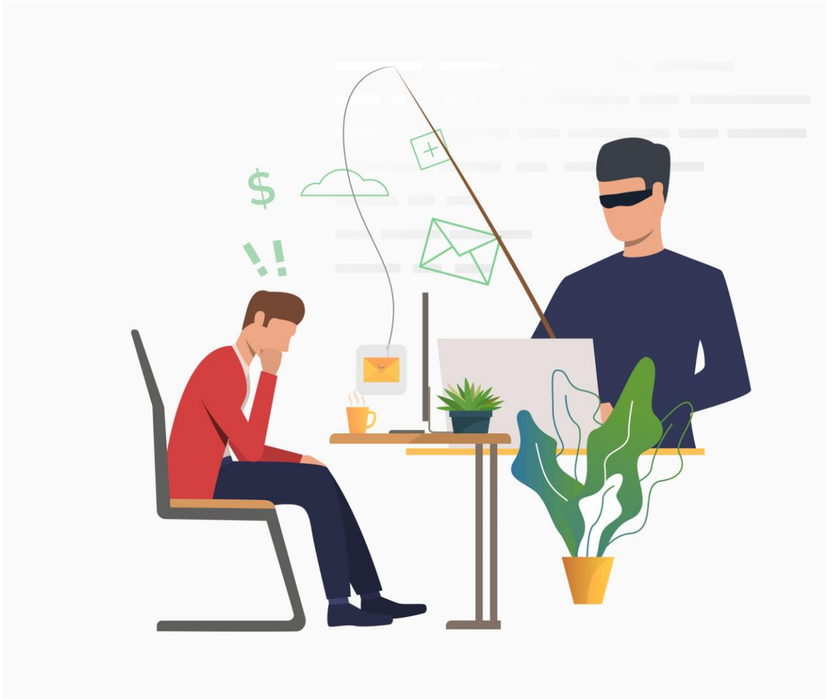
Phish Detector

Project Guide:
Mr. M V Pavan Kumar [M.Tech]

G.Varun Y20ACS455
E.Venkatesh Y20ACS444
Ch.Naveen Y20ACS427
B.Karthik L21ACS402

CONTENTS

- Abstract
- Existing System
- Proposed System
- Design
- Implementation
- Results
- Conclusion



ABSTRACT

- Phishing is a kind of worldwide spread cybercrime that uses disguised websites to trick users into downloading malware or providing personally sensitive information to attackers.

- With the rapid development of artificial intelligence, more and more researchers in the cybersecurity field utilize machine learning algorithms to classify phishing websites.

- This project proposes a comprehensive framework for detecting phishing websites using machine learning techniques. It involves data collection, preprocessing, feature extraction, model selection, training, evaluation, and deployment.

- Relevant features are extracted from URLs, and various machine learning models are explored for classification, with hyperparameter tuning for optimization.

EXISTING SYSTEM

- In existing systems Random Forest and Decision Tree algorithms are heavily employed in present systems, and their accuracy has to be enhanced. The existing models have low latency.

- Existing systems do not have a specific user interface.

- Consumers are led to a faked website that appears to be from the authentic company when the e-mails or the links provided are opened.

PROPOSED SYSTEM

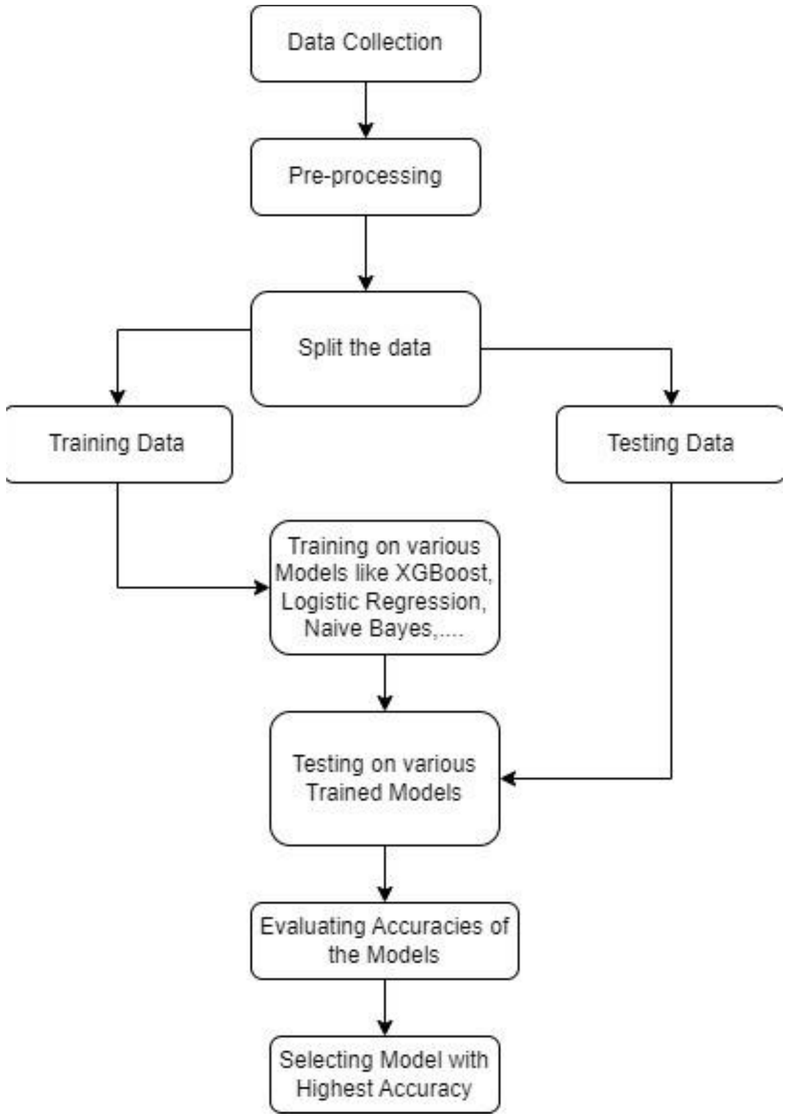
- In the proposed system, we introduce machine learning models such as Logistic Regression, Support Vector Machine, Gradient Boost, XGBoost, Naive Bayes alongside Random Forest and Decision Tree algorithms used in existing systems.

- By Performing the training over the all Machine Learning models, we choose the best performed Model based on accuracy score of model to perform test data. Here the Machine Learning model is Gradient Boost which having the accuracy score over 97%.

- These additions aim to diversify the model ensemble and improve accuracy. Moreover, we prioritize enhancing the system's latency performance.

- Additionally, a user-friendly interface will be developed to facilitate ease of use.

DESIGN


Data Collection
Pre-processing
Split the data
Training Data
Testing Data
Training on various Models like XGBoost, Logistic Regression, Naive Bayes,....
Testing on various Trained Models
Evaluating Accuracies of the Models
Selecting Model with Highest Accuracy

cont.

.

Data Collection:

- This initial phase involves gathering the data your machine learning model will be trained on. In phishing website detection, this might involve collecting URLs from publicly available datasets of legitimate and phishing websites.

Pre-Processing:

- Raw data often requires cleaning and preparation before it can be used for training. Pre-processing steps might include handling missing values, formatting inconsistencies, and transforming the data into a suitable format for your machine learning model.

Splitting the Data:

- The collected data is divided into two portions: training data and testing data. The training data, typically the larger portion, is used to train the machine learning model. The testing data is used to evaluate the model's performance on unseen examples.

cont.

Training the Model:

- Using the training data, various machine learning models are trained. During training,
- the model learns to identify patterns and relationships within the website features that differentiate between phishing and legitimate websites.

Testing Various Models:

- The training data and testing data is trained on Various Meachine learning models like logistic regression, K-Means, Gradient boost, XG Boost etc.

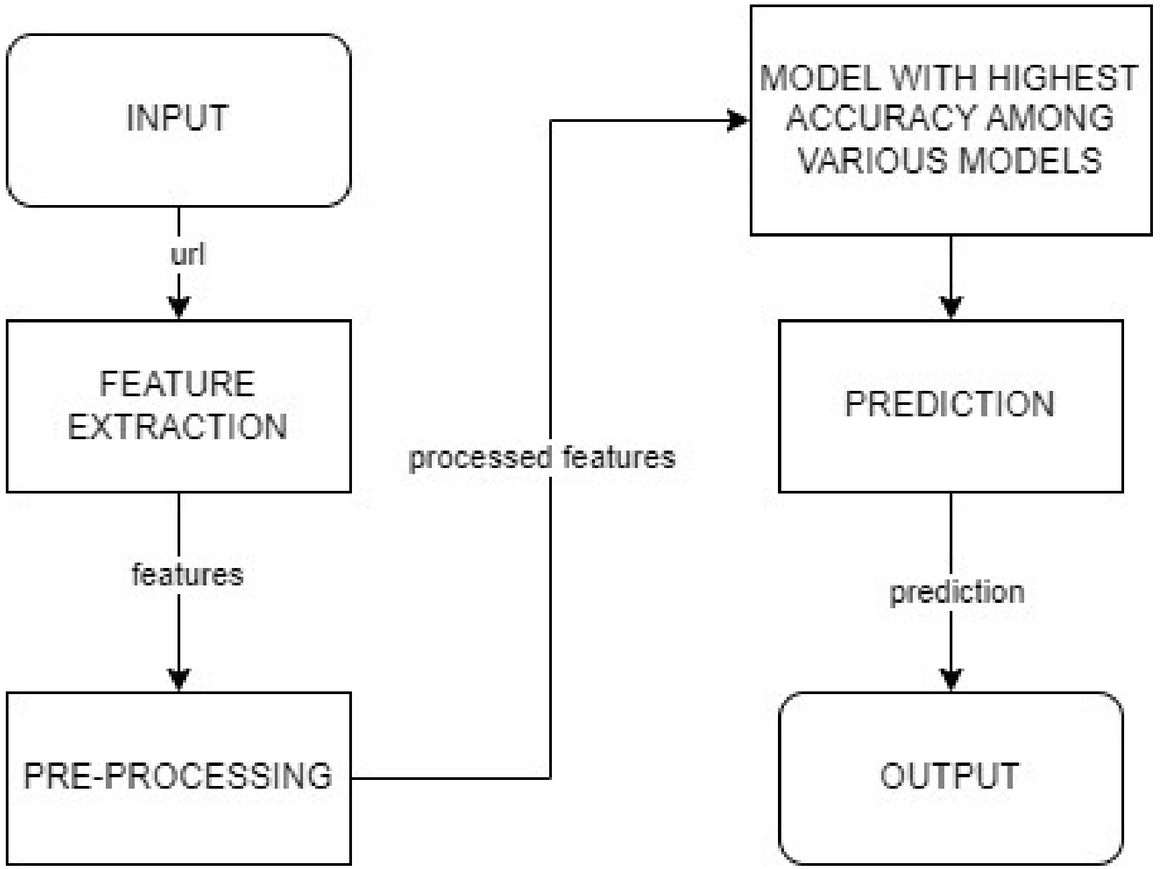
Evaluation Metrics:

- After training, the model's performance is assessed using various evaluation metrics. Common metrics include accuracy, precision, recall, and F1 score.

Deployment:

- If the evaluation results are satisfactory, the model can be deployed to a real-world setting.

IMPLEMENTATION



cont.
.

In this phase the design is translated into code

Libraires or API's used :

- **Pandas** - Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively.
- **NumPy** - NumPy is a general-purpose array-processing package. It provides a high- performance multidimensional array object, and tools for working with these arrays.
- **Matplotlib** - : Matplotlib is easy to use and an amazing visualizing library in Python. It is consists of several plots like line, bar, scatter, histogram, etc.
- **Scikitlearn** - It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, and clustering.
- **Pickle** - The pickle module is used for serializing and deserializing Python objects into a byte stream
- **Flask** - : Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications

Input Required :

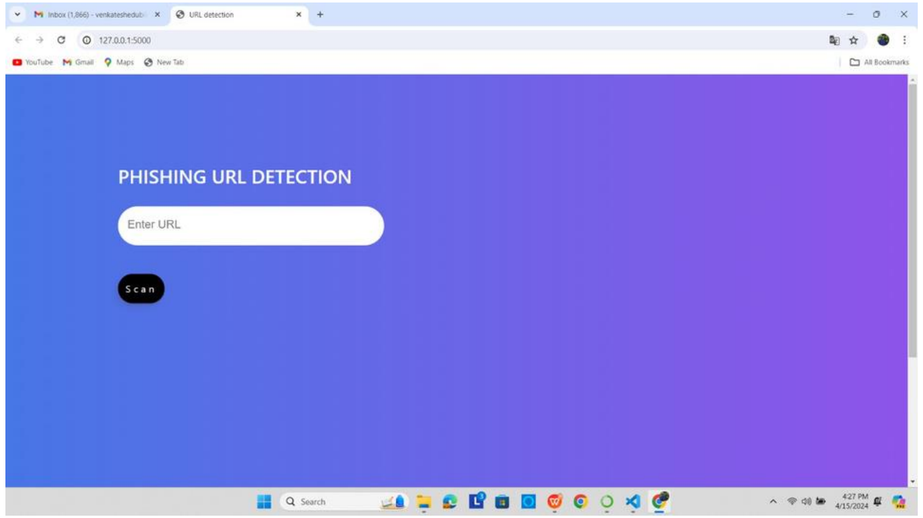
- **URL** - The URL provided by the user is typically entered into a web application or system interface.
- **Purpose** - The purpose of collecting the URL as input is to determine whether it is safe or potentially malicious (phishing).
- **Format** - URLs follow a specific format, starting with the protocol (e.g., "http://" or "https://"), followed by the domain name and optional path and query parameter

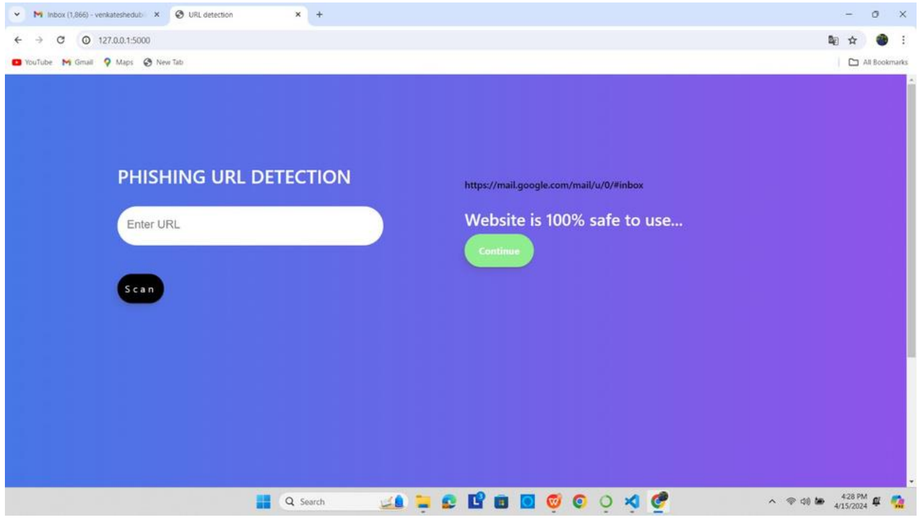
Feature Extraction :

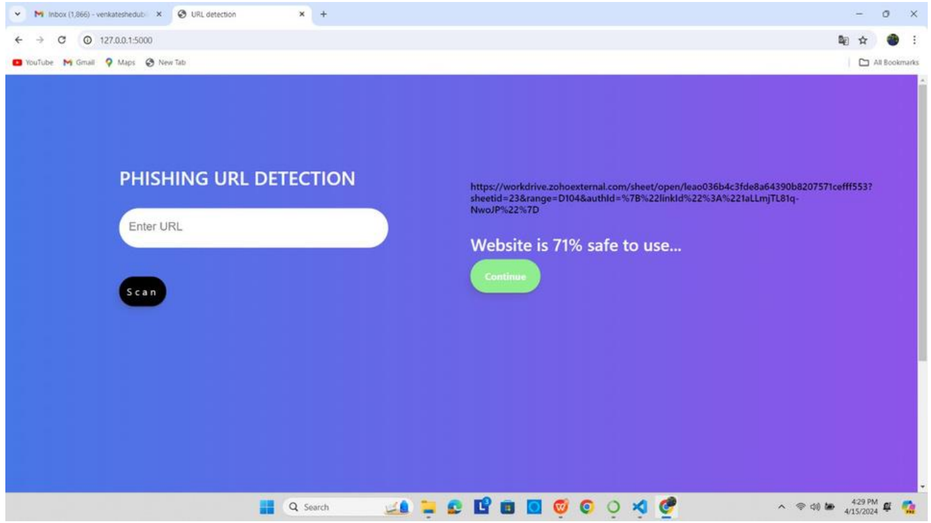
- We Extract 30 URL features from the given URL.
- The below table neatly arranges 30 URL features into two columns

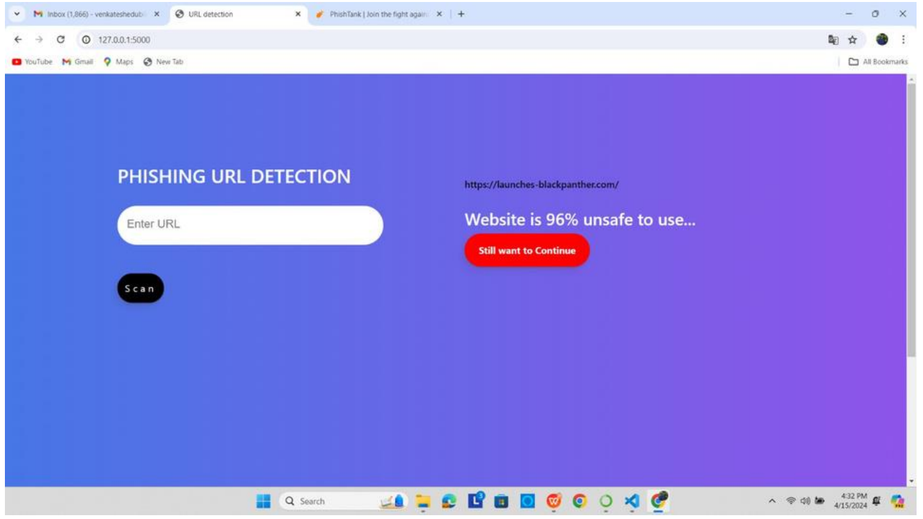
Features		
UsingIP	LongURL	ShortURL
Symbol@	Redirecting//	PrefixSuffix-
SubDomains	HTTPS	DomainRegLen
Favicon	NonStdPort	HTTPSDomainURL
RequestURL	AnchorURL	LinksInScriptTags
ServerFormHandler	InfoEmail	AbnormalURL
WebsiteForwarding	StatusBarCust	DisableRightClick
UsingPopupWindow	IframeRedirection	AgeofDomain
DNSRecording	WebsiteTraffic	PageRank
GoogleIndex	LinksPointingToPage	StatsReport

RESULT


PHISHING URL DETECTION
Enter URL
Scan


PHISHING URL DETECTION
Enter URL
Scan
https://mail.google.com/mail/u/0/#inbox
Website is 100% safe to use...
Continue


PHISHING URL DETECTION
Enter URL
Scan
https://workdrive.zohoexternal.com/sheet/open/1eao036b4c3fde8a64390b820757 1cefff553?sheetid=23&range=D104&authId=%7B%22linkId%22%3A%221aLLmjTL81q-NwoJP%22%7D
Website is 71% safe to use...
Continue


PHISHING URL DETECTION
Enter URL
Scan
https://launches-blackpanther.com/
Website is 96% unsafe to use...
Still want to Continue

CONCLUSION

- To the best of our knowledge, this study is the first analysis to include the findings of all other studies into the detection of phishing websites using machine learning techniques.

- The website functionality is used by ML-based phishing approaches to collect information that could be used to classify websites for the purpose of identifying phishing sites. Developing focused anti-phishing approaches and methods as well as minimizing their inconvenience are two waysto prevent phishing.

- We achieved 97% detection accuracy using Gradient Boost algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data.

THANK YOU