Phishing Website Classification Project

Group Members:
Faiza Mahek - 219X1A2910
J Jyothi-219X1A2914
V S S P Akshaya-219X1A2936

Introduction

Understanding Phishing and Its Implications

Significance of Phishing

With the rise of online activities, phishing has become a critical threat to users, businesses, and organizations.

Definition of Phishing

Phishing is a growing cyber threat where attackers disguise malicious websites as legitimate to steal sensitive information.

Objective of the Project

The goal is to create a machine learning model that accurately classifies websites as phishing or legitimate and integrates it into a Django-based web application.

Abstract

01 Problem

Phishing attacks are becoming more sophisticated, making it difficult for users to distinguish between legitimate and phishing websites.

02 Solution

Develop a machine learning-based phishing detection system using a Random Forest classifier.

03 Scope

The project encompasses three main areas:

04 Model Development

Build an ML model for website classification.

05 Web Application Development

Develop a user-friendly interface for phishing detection.

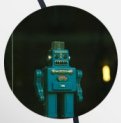
06 Deployment

Make the app globally accessible through cloud deployment (Heroku).

Technical Domain

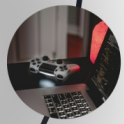
Machine Learning (ML)

Algorithm development for phishing detection.



Data Science

Feature selection and data preprocessing.



Web Development

Django-based application integration.

Cloud Computing

Deployment using Heroku.

Base Paper 1 Overview

Detection of Phishing Website Using Machine Learning Approach

01 Title

Detection of Phishing Website Using Machine Learning Approach

02 Authors

Mahajan Mayuri Vilas et al.

03 Conference

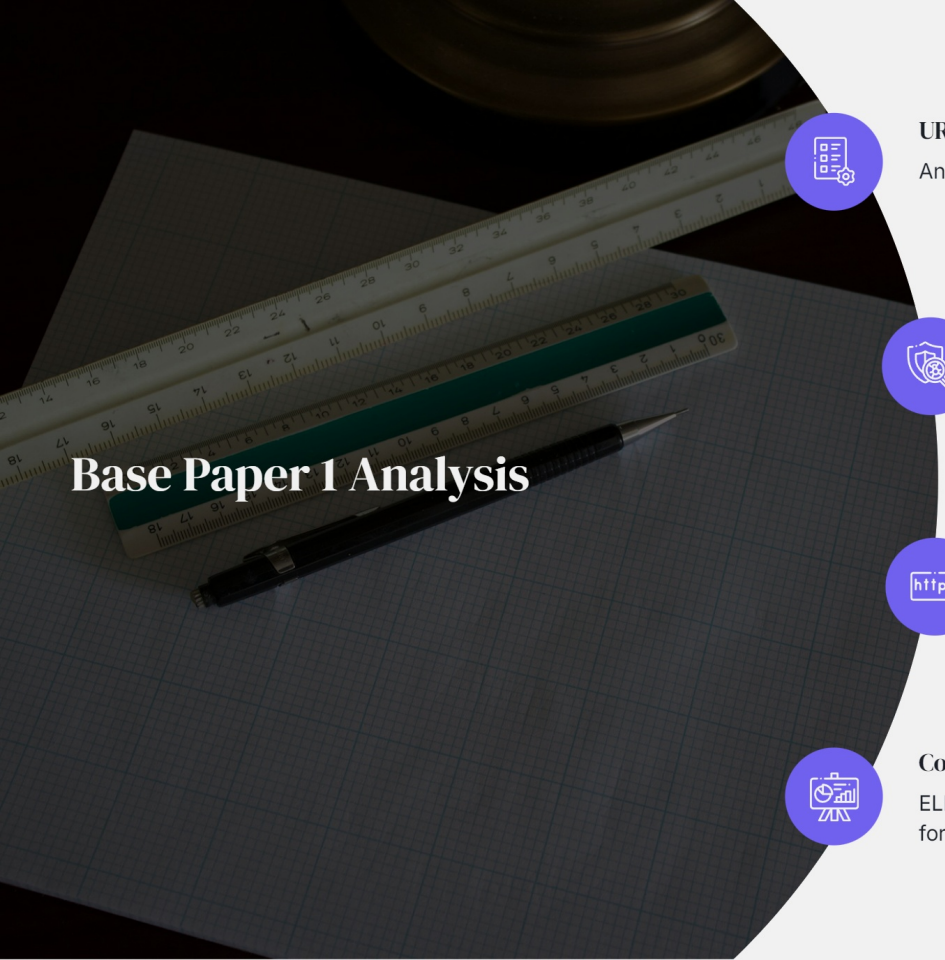
Presented at ICEECCOT 2019

04 Machine Learning Focus

Focuses on machine learning to identify phishing websites by analyzing URLs and website authority.

05 Extreme Learning Machine (ELM)

Explores the effectiveness of Extreme Learning Machine (ELM) and URL feature analysis.



Base Paper 1 Analysis

URL Evaluation

Analyzes URL structure.

Authority Checking

Checks website legitimacy and supervision.

HTTPS Use

Phishing websites often use HTTPS to appear secure.

Conclusion

ELM combined with URL-based analysis provides a reliable method for phishing detection.

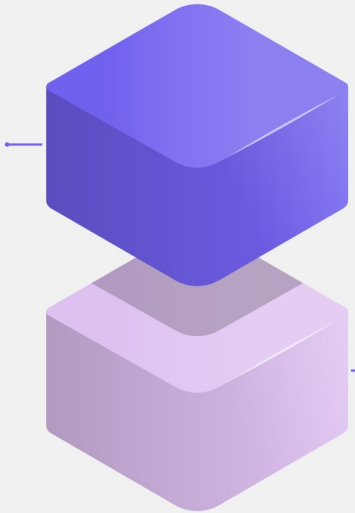
Base Paper 2 Overview

Comparative Study of Classification Algorithms for Website Phishing Detection

Comparison of Classification Algorithms

01

The study compares several classification algorithms, focusing on their performance across multiple datasets.



Datasets Used

02

The datasets utilized in the research include UCI Phishing Websites (2016) and Mendeley Phishing Websites (2018).

Base Paper 2 Analysis

Highest Accuracy Achieved

Random Forest achieved the highest accuracy, with 97.50% on Mendeley and 88.92% on UCI.

Performance Compared to Other Algorithms

Random Forest outperformed other algorithms, including SVM, Naive Bayes, and Decision Trees.

Conclusion on Model Effectiveness

Random Forest is the most effective model for phishing detection, especially on varied datasets.

Base Paper 3 Overview

Detecting Phishing Websites Using Naive Bayes Classification

The paper focuses on utilizing Naive Bayes classification techniques specifically for detecting phishing websites.

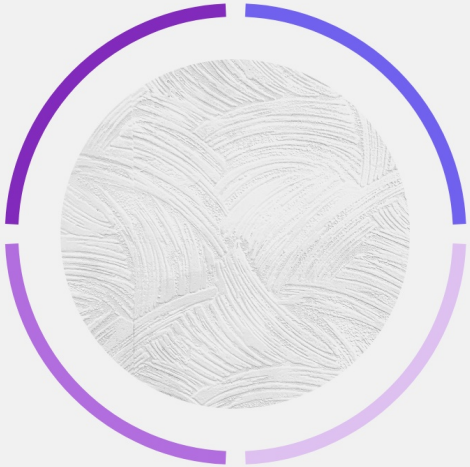
It compares three variants of Naive Bayes: Bernoulli, Multinomial, and Complement, evaluating their effectiveness in phishing detection.

Base Paper 3 Analysis

Results

04

Multinomial and Complement Naive Bayes achieved 96% accuracy, making them suitable for phishing detection.



01

Bernoulli Naive Bayes

Focuses on binary features.

Complement Naive Bayes

03

Improved performance for imbalanced data.

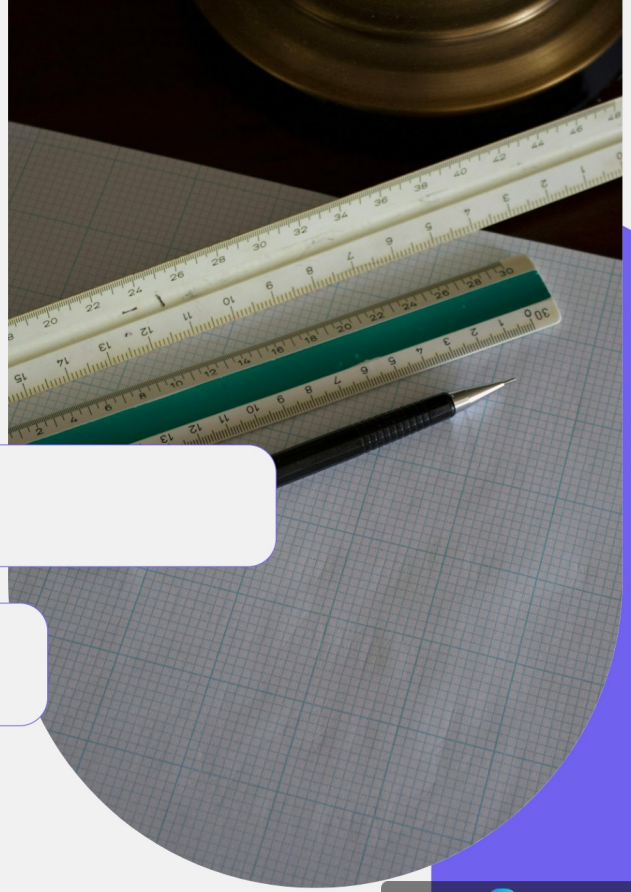
02

Multinomial Naive Bayes

Best for text classification.

Base Paper 4 Overview

Phishing Website Detection Using Random Forest and Support Vector Machine



Comparison of Detection Methods

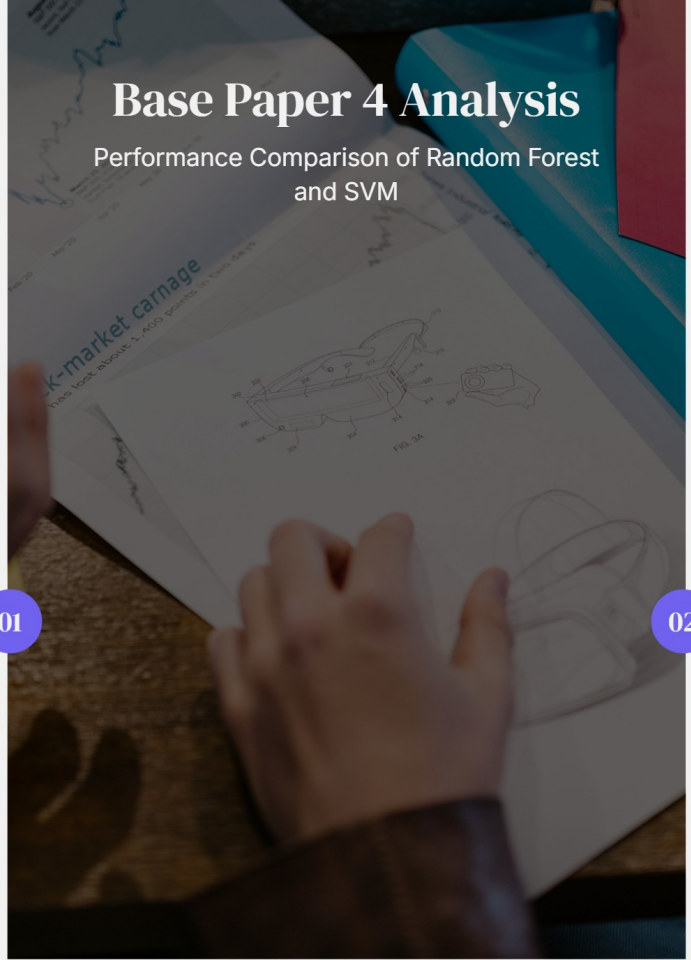
Compares Random Forest and SVM for phishing detection.

HTML and Hyperlink Analysis

Analyzes HTML code and hyperlink structure to identify phishing attempts.

Base Paper 4 Analysis

Performance Comparison of Random Forest and SVM



Random Forest Performance

01

Random Forest achieved 99.98% accuracy.

SVM Performance

02

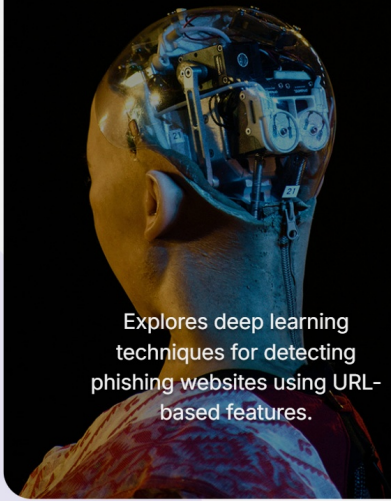
SVM achieved 84.73% accuracy, showing it is less effective for this problem.

Base Paper 5 Overview

Deep Learning for Phishing Website Detection

01

Deep Learning Techniques



Explores deep learning techniques for detecting phishing websites using URL-based features.

02

Performance Comparison

Compares the performance of Random Forest and SVM with deep learning models.

Base Paper 5 Analysis

Techniques for Phishing Detection

Feature Extraction

Analyzes URL characteristics to differentiate between phishing and legitimate sites.

Model Comparison

Random Forest performs well, but deep learning models show potential for further improvement.

Conclusion

Deep learning can enhance phishing detection, though Random Forest remains a reliable choice.

Algorithm: Random Forest Classifier

Understanding the Benefits and Features

Handles large datasets

Random Forest is capable of managing large volumes of data while effectively reducing overfitting.

High accuracy in phishing detection

This algorithm provides significant accuracy and stability when detecting phishing attempts.

Compatibility with data types

Random Forest works efficiently with both categorical and continuous data.

Key features in analysis

Important features for analysis include URL structure, domain age, and presence of HTTPS.

Design and Implementation Plan

01

Data Preprocessing

Clean dataset, handle missing values, and select relevant features.

02

Model Training

Train the Random Forest classifier with labeled phishing and legitimate websites.

03

Integration with Django

Implement the model into the Django framework for a web application.

Web Application Design

04

User Interaction

Users input the website URL, and the system processes it in real time to predict its legitimacy.

01

User Interface

Simple input form for website URL.

03

Back-End Technology

Utilizes Django framework integrated with a trained Random Forest model.

02

Instant Classification Results

Immediate feedback on whether the website is Phishing or Legitimate.

Expected Results

Impact

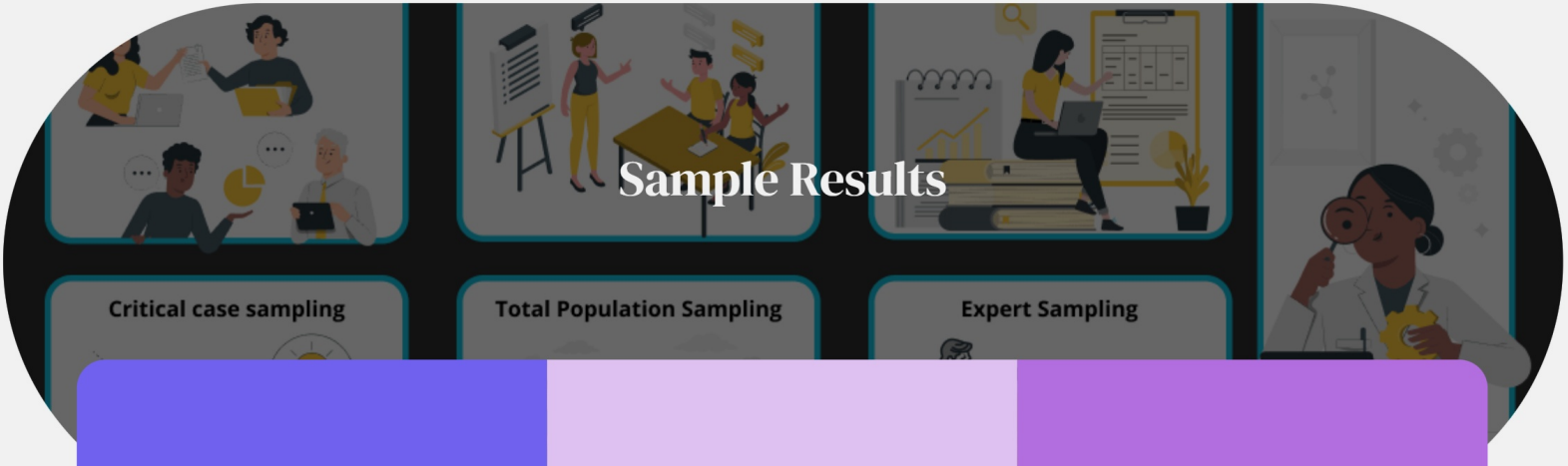
Improved protection for users and organizations against phishing attacks.

Performance

Fast, real-time predictions for users, scalable to handle a large number of queries.

Accuracy

Expected >95% accuracy in classifying phishing websites.



Confusion Matrix

Accurate classification

Displays accurate classification of phishing and legitimate websites.

ROC Curve

Robust performance

High true positive rate, low false positive rate, indicating robust performance.

Accuracy

High detection rate

The Random Forest model achieves 96-99% accuracy in detecting phishing websites.

Conclusion

Summary and Future Scope

Project Integration

The project successfully integrates a machine learning-based phishing detection model into a web application.

Best Performing Algorithm

Random Forest is the best performing algorithm for this task.

Phishing Detection Tool

The solution offers a simple yet effective tool to detect phishing websites.

Future Scope - Deep Learning

Implement deep learning models for further accuracy improvements.

Future Scope - Mobile Applications

Extend the tool for mobile applications to increase accessibility.