A

SEMINAR REPORT

ON

Phishing Website Detection System

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
INFORMATION TECHNOLOGY

BY

Rajvardhan Haridas Deshmukh
Roll No: 33224
Exam Seat No:

Under the guidance of
Mr. Tushar A. Rane

DEPARTMENT OF INFORMATION TECHNOLOGY
PUNE INSTITUTE OF COMPUTER TECHNOLOGY
SR. NO 27, PUNE-SATARA ROAD, DHANKAWADI
PUNE - 411 043.
AY: 2024-2025

SCTR’s PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the Seminar work entitled
Phishing Website Detection System

Submitted by

Name : Rajvardhan Haridas Deshmukh
Exam Seat No:33224 .

is a bonafide work carried out under the supervision of Mr. Tushar A. Rane and it is submitted towards the partial fulfillment of the requirements of Savitribai Phule Pune University, Pune for the award of the degree of Bachelor of Engineering (Information Technology).

Mr. Tushar A. Rane
Seminar Guide

Dr. A. S. Ghotkar
HOD IT

Dr. S. T. Gandhe
Principal

Date:
Place:PICT,Pune

Acknowledgement

I would like to express my deepest gratitude to everyone who supported and guided me throughout the preparation and completion of this seminar

I am immensely grateful to my seminar guide, Mr. Tushar A.Rane, for her invaluable guidance, encouragement, and insightful feedback at every stage of this seminar. Her expertise and unwavering support were instrumental in shaping the direction and outcome of this work.

I also wish to thank Dr. Jayashree B.Jagdale , the reviewer, for his constructive evaluation and suggestions, which greatly contributed to enhancing the quality of this seminar.

I would like to acknowledge my peers and friends for their valuable input and support, which motivated me throughout the seminar preparation. Lastly, I extend my heartfelt appreciation to my family for their continuous encouragement, patience, and unwavering support. Thank you all for your contributions and support.

Name :Rajvardhan Deshmukh
Exam Seat No:

Abstract

Phishing attacks continue to represent a major threat in the realm of cybersecurity, characterized by the deceptive practice of tricking users into divulging sensitive personal information, such as usernames, passwords, and financial details, through fraudulent websites that masquerade as legitimate entities. This report outlines a comprehensive approach to detect phishing websites using machine learning techniques, focusing on the extraction of critical features from URLs. Key attributes commonly associated with phishing attempts include the presence of IP addresses in the URL, the overall length of the domain name, the frequency of suspicious keywords, and the structural properties of the URLs. A dataset comprising both phishing and benign URLs is utilized to train and evaluate multiple classification models, including Random Forest, LightGBM, and XGBoost. Experimental results demonstrate that these models can achieve high levels of accuracy and precision in classifying URLs, effectively identifying potential phishing threats. The findings highlight the significant potential of machine learning algorithms to enhance the automation and accuracy of phishing detection systems, ultimately contributing to more robust cybersecurity measures.

Keywords: Phishing detection, Feature extraction, Machine learning, URL-based features, Random Forest, LightGBM, XGBoost.

Contents

Certificate . . . i

Acknowledgement . . . ii

Abstract . . . iii

Contents . . . iv

List of Figures . . . vii

List of Tables . . . viii

Abbreviations . . . ix

1 Introduction 1

1.1 Introduction . . . 1

1.2 Motivation . . . 1

1.3 Objectives . . . 2

1.4 Scope . . . 3

2 Literature Survey 4

3 Methodologies 5

3.1 Framework/Basic Architecture . . . 5

3.1.1 Data Collection . . . 5

3.1.2 Feature Extraction . . . 5

3.1.3 Data Preprocessing . . . 6

3.1.4 Model Selection . . . 6

3.1.5 Training and Testing . . . 6

3.1.6 Evaluation Metrics . . . 6

3.2 Different Approaches . . . 6

3.3 State-of-the-art Algorithms . . . 7

3.3.1 Random Forest . . . 7

3.3.2 Support Vector Machines (SVM) . . . 8

3.3.3 LightGBM Model . . . 8

3.3.4 XGBoost . . . 9

3.4 Implemented Algorithms . . . 9

3.4.1 Model Implementation . . . 9

3.4.2 Evaluation Metrics 9
3.5 Discussion 9

4 Implementation 11
4.1 Proposed Algorithm for Phishing Detection 11
4.1.1 XGBoost 11
4.1.2 Logistic Regression 11
4.1.3 LightGBM 11
4.1.4 Random Forest 11
4.2 Methodology 12
4.2.1 Preprocessing 12
4.2.2 Handling Missing Values 12
4.2.3 Data Distribution Analysis 12
4.2.4 Train-Test Split 12
4.2.5 Feature Scaling 12
4.2.6 Model Training 14
4.3 Results 14
4.3.1 Random Forest 14
4.3.2 XGBoost 15
4.3.3 Comparison Study 16
4.4 Software Requirement Specification 16
4.4.1 Constraints and Assumptions 16
4.4.2 Inputs and Outputs 17
4.5 Platform for Implementation 17

5 Applications 18
5.1 State-of-the-art Applications 18
5.1.1 Financial Services 18
5.1.2 E-commerce 18
5.1.3 Corporate Security 18
5.1.4 Healthcare Sector 18
5.1.5 Email Services 19
5.2 Challenges 19
5.2.1 Evolving Threat Landscape 19
5.2.2 Data Quality and Imbalance 19
5.2.3 User Awareness and Training 19
5.2.4 Privacy Concerns 19
5.2.5 Adapting to New Technologies 20

6 Conclusion 21

7 Future Scope **22**

References **23**

Plagiarism Report **24**

List of Figures

3.1 Basic Framework for Phishing Website Detection 5
3.2 Overview of Phishing Detection Approaches 7

4.1 Steps followed in the methodology 13
4.2 Results of Random Forest 14
4.3 Results of XGBoost 15
4.4 ROC Comparison 16

List of Tables

2.1 Literature review of phishing website detection techniques using machine learning 4

3.1 Comparison of Algorithm Performance for Phishing Detection 9

4.1 Classification report for Random Forest 15

Abbreviations

ML	:	Machine Learning
URL	:	Uniform Resource Locator
HTTP	:	HyperText Transfer Protocol
HTTPS	:	HyperText Transfer Protocol Secure
DNS	:	Domain Name System
DGA	:	Dynamic Domain Generation Algorithm
API	:	Application Programming Interface
TPR	:	True Positive Rate
FPR	:	False Positive Rate
ROC	:	Receiver Operating Characteristic
AUC	:	Area Under the Curve
SVM	:	Support Vector Machine
KNN	:	k-Nearest Neighbors
RF	:	Random Forest
F1	:	F1 Score

1. Introduction

1.1 Introduction

Phishing has become a pervasive threat in today's digital environment, exploiting the rapid growth of online services and user interactions. As individuals increasingly rely on the internet for banking, shopping, and communication, cybercriminals have devised sophisticated tactics to trick users into divulging sensitive information. Phishing attacks can take various forms, including deceptive emails, fraudulent websites, and social engineering techniques, leading to significant financial losses and identity theft. According to recent statistics, over 3.5 billion phishing emails are sent daily, targeting unsuspecting users globally. This growing threat underscores the urgent need for effective detection mechanisms to safeguard users against phishing scams.

Traditional methods of detecting phishing websites often rely on manual reporting or simple heuristics, which are inadequate against rapidly evolving phishing tactics. As phishing techniques become more advanced, including the use of SSL certificates and visually similar URLs, the need for robust, automated detection systems becomes increasingly critical. In this context, machine learning (ML) presents a promising approach, leveraging large datasets and sophisticated algorithms to identify patterns associated with phishing sites. By employing techniques such as feature extraction and model training, machine learning can enhance the accuracy and efficiency of phishing detection systems.

This seminar aims to explore the application of feature extraction techniques and machine learning algorithms, such as Decision Trees, Random Forests, and Support Vector Machines (SVM), in the detection of phishing websites. Each algorithm will be evaluated based on its ability to learn from a dataset, adapt to new phishing techniques, and minimize false positives and negatives. By analyzing the strengths and weaknesses of these algorithms, the research seeks to enhance the effectiveness of phishing detection systems and ultimately improve online safety for users.

1.2 Motivation

The choice of this seminar topic is driven by the alarming rise in phishing attacks, which have escalated dramatically over the past few years. Reports indicate that phishing is responsible for a significant portion of cybercrime incidents, leading to billions of dollars in losses annually. The increasing sophistication of these attacks, particularly during global events that drive online activity, highlights the inadequacy of existing detection mechanisms. For instance, during the COVID-19 pandemic, there was a notable surge in phishing attempts related to health information and financial support, capitalizing on public fear and uncertainty.

Moreover, the psychological impact on victims of phishing scams can be severe, resulting in a loss of trust in digital platforms. Many victims suffer from anxiety and fear of further fraud, creating a ripple effect that can deter online transactions for years. This urgency calls for innovative solutions to bolster phishing detection capabilities. Machine learning techniques, particularly those that utilize feature extraction, offer a viable path forward. By analyzing various attributes of web pages and user interactions, ML algorithms can effectively distinguish between legitimate and malicious sites.

The motivation behind this research extends to the potential for machine learning to not only improve detection rates but also to provide real-time analysis that can adapt to emerging phishing threats. The seminar seeks to conduct a thorough investigation into these techniques, focusing on their practical implementation and effectiveness in combating phishing attacks, thus contributing to a safer digital environment.

1.3 Objectives

- **Analyze Phishing Characteristics:** Examine the features commonly associated with phishing websites, including URL structure, domain age, web page content, and the presence of specific keywords. Understanding these characteristics will enable a deeper comprehension of the distinguishing elements that separate phishing sites from legitimate ones.

- **Evaluate Machine Learning Algorithms:** Implement and compare various machine learning algorithms, such as Decision Trees, Random Forests, and SVM, to assess their effectiveness in detecting phishing websites based on the identified features. This evaluation will include examining the algorithms' performance in terms of accuracy, precision, recall, and F1-score across diverse datasets.

- **Address Limitations and Future Directions:** Identify the challenges faced in current phishing detection methods, such as overfitting, data scarcity, and evolving phishing tactics. Additionally, propose potential improvements and future research directions in leveraging machine learning for enhanced detection capabilities, including the incorporation of ensemble learning techniques and the use of deep learning models.

- **Raise Awareness on Phishing Threats:** Highlight the importance of user education and awareness regarding phishing threats. Discuss the role of educational campaigns in reducing the likelihood of successful attacks and the significance of ongoing research in this area to protect users.

1.4 Scope

The scope of this seminar encompasses a detailed examination of feature extraction and machine learning techniques specifically applied to phishing website detection. The research will include a comparative analysis of various algorithms, evaluating their performance using metrics such as accuracy, precision, recall, and F1-score. This analysis will provide participants with insights into the operational mechanisms and effectiveness of each algorithm.

Additionally, the seminar will explore the implementation challenges of these techniques in real-world scenarios, discussing the potential limitations and ethical considerations of using machine learning for phishing detection. By focusing on current trends and innovations in the field, this seminar aims to equip participants with the knowledge necessary to contribute to the ongoing efforts in combating phishing attacks and improving online security.

Furthermore, the seminar will also discuss potential applications of the developed detection systems in various sectors, including banking, e-commerce, and social media. By illustrating the impact of effective phishing detection on user trust and business integrity, this research seeks to emphasize the critical importance of proactive measures in the ongoing battle against cyber threats.

2. Literature Survey

Title	**Author**	**Publication Date**	**Aim/Objective**
A Feature Extraction Approach for the Detection of Phishing Websites Using Machine Learning [4]	Zoran Stamenkovic	June 2023	The proposed method utilizes feature extraction algorithms, incorporating third-party records such as WHOIS and DNS, to classify websites effectively, even allowing offline use when necessary by omitting certain features.
Phishing Website URL's Detection Using NLP and Machine Learning Techniques	Dinesh Kalla	December 2023	The objective is to optimize prediction accuracy, reduce overfitting, and develop robust algorithms for real-world cybersecurity frameworks by leveraging model tuning and ensemble approaches.
A review of a website phishing detection taxonomy	Damian Fraszczak	June 2024	The objective of this seminar report is to investigate and analyze various methodologies for detecting phishing websites using feature extraction and machine learning techniques.
Website Phishing Detection Using Machine Learning Techniques	R. Alazaidah	January 2024	The report evaluates two phishing website datasets using 24 classifiers across six learning strategies to identify the most effective methods for detection.
Detecting Phishing Domains Using Machine Learning	Shouq Alnemari	April 2023	The objective of this report is to evaluate the effectiveness of various machine learning classification techniques, including Decision Trees, Random Forests, and Support Vector Machines, in detecting phishing websites.

Table 2.1: Literature review of phishing website detection techniques using machine learning

3. Methodologies

This section outlines the methodologies applied in the development of the phishing website detection system explored in this seminar.

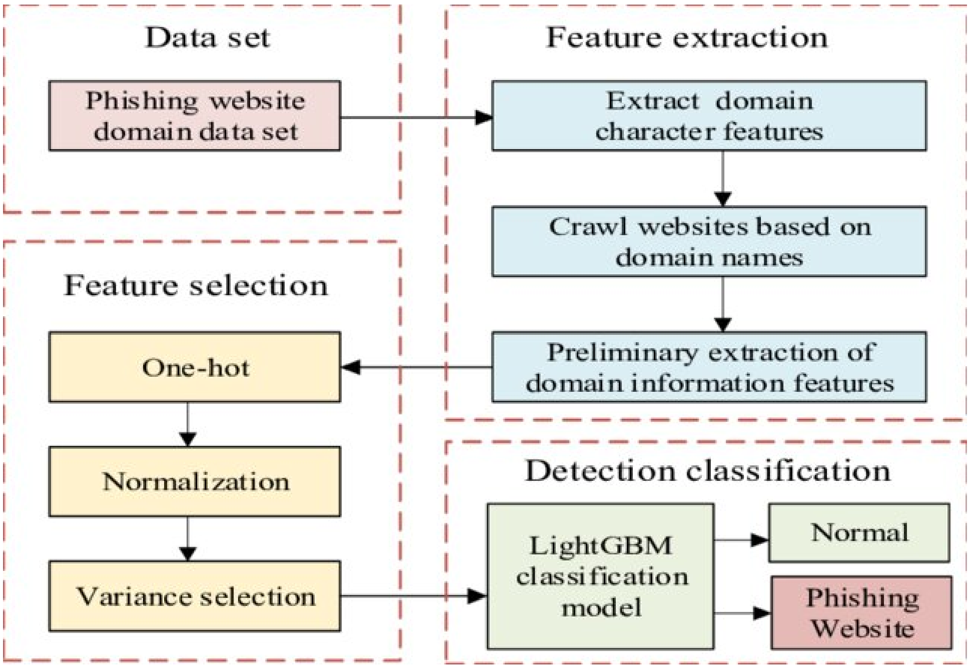

Data set
Phishing website domain data set
Feature extraction
Extract domain character features
Crawl websites based on domain names
Preliminary extraction of domain information features
Feature selection
One-hot
Normalization
Variance selection
Detection classification
LightGBM classification model
Normal
Phishing Website

Figure 3.1: Basic Framework for Phishing Website Detection

3.1 Framework/Basic Architecture

3.1.1 Data Collection

- Phishing websites are collected from various online repositories and databases that provide samples for research purposes.
- Legitimate websites are also gathered for comparison to ensure a balanced dataset.

3.1.2 Feature Extraction

- Various features are extracted from the URLs, HTML content, and metadata of the websites.
- Important features may include the length of the URL, presence of HTTPS, and the use of special characters.

3.1.3 Data Preprocessing

- The collected data is cleaned to remove any inconsistencies and irrelevant information.
- Normalization and encoding are applied to prepare the dataset for machine learning algorithms.

3.1.4 Model Selection

- Various machine learning algorithms such as Random Forest, SVM, and XGBoost are evaluated for their performance in detecting phishing websites.
- Each model is trained using the prepared dataset, and hyperparameter tuning is performed to enhance performance.

3.1.5 Training and Testing

- The dataset is split into training and testing sets to evaluate the performance of the models.
- Cross-validation techniques are employed to ensure robustness and prevent overfitting.

3.1.6 Evaluation Metrics

- Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of the models.
- The confusion matrix is utilized to visualize the results.

3.2 Different Approaches

Phishing Detection Approaches

1. **URL-Based Detection:** Analyzes the URL structure and components to identify malicious patterns.
2. **Content-Based Detection:** Examines the HTML content and features of the web pages to detect phishing attempts.
3. **Machine Learning Models:** Utilizes algorithms like Random Forest, SVM, and Neural Networks to classify websites as phishing or legitimate based on extracted features.

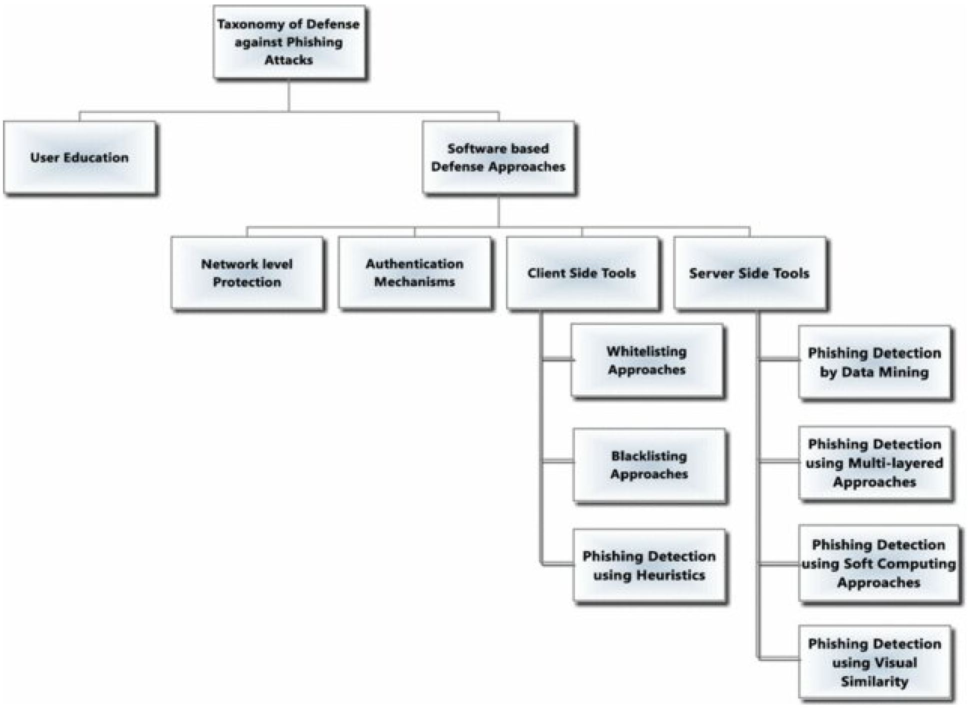

Taxonomy of Defense against Phishing Attacks
User Education
Software based Defense Approaches
Network level Protection
Authentication Mechanisms
Client Side Tools
Server Side Tools
Whitelisting Approaches
Blacklisting Approaches
Phishing Detection using Heuristics
Phishing Detection by Data Mining
Phishing Detection using Multi-layered Approaches
Phishing Detection using Soft Computing Approaches
Phishing Detection using Visual Similarity

Figure 3.2: Overview of Phishing Detection Approaches

4. **Heuristic Methods:** Combines various rules and heuristics derived from expert knowledge to flag suspicious websites.

5. **Hybrid Models:** Integrates multiple detection methods to improve accuracy and reduce false positives.

3.3 State-of-the-art Algorithms

3.3.1 Random Forest

Overview:

Random Forest is an ensemble learning method that constructs multiple decision trees and merges their predictions to improve classification accuracy.

Key Features:

- Robust against overfitting due to averaging across trees.
- Handles both numerical and categorical data effectively.

Algorithm Complexity:

- **Training Time Complexity:** $O(T \cdot n \cdot \log n)$, where T is the number of trees, and n is the number of samples.
- **Prediction Time Complexity:** $O(T \cdot d)$, where d is the depth of each tree.

3.3.2 Support Vector Machines (SVM)

Overview:
SVM is a supervised learning model that finds the optimal hyperplane for separating different classes in a high-dimensional space.

Key Features:

- Effective in high-dimensional spaces and with a clear margin of separation.
- Utilizes kernel trick to transform data into higher dimensions for better separability.

Algorithm Complexity:

- **Training Time Complexity:** $O(n^3)$ for the basic algorithm; can be improved using advanced methods.
- **Prediction Time Complexity:** $O(n)$ per prediction.

3.3.3 LightGBM Model

Overview:
LightGBM, or Light Gradient Boosting Machine, is a gradient boosting framework that uses tree-based learning algorithms. It is designed for distributed and efficient training, making it particularly effective for large datasets and high-dimensional data.

Key Features:

- **Efficiency:** LightGBM is designed to be faster and more efficient than other gradient boosting implementations by using a histogram-based approach.
- **Scalability:** It can handle large datasets with millions of instances and features effectively, making it suitable for big data applications.
- **Leaf-wise Growth:** Unlike other algorithms that grow trees level-wise, LightGBM grows trees leaf-wise, leading to deeper trees and better accuracy.
- **Built-in Support for Categorical Features:** LightGBM can directly handle categorical features without the need for one-hot encoding.

Algorithm Complexity:

- **Training Time Complexity:** $O(n \cdot \log(n) \cdot k)$, where n is the number of data points and k is the number of features.
- **Prediction Time Complexity:** $O(k)$ per prediction.

3.3.4 XGBoost

Overview:

XGBoost is an optimized gradient boosting algorithm that is efficient and scalable for large datasets.

Key Features:

- Offers high predictive accuracy through ensemble learning.
- Features regularization techniques to combat overfitting.

Algorithm Complexity:

- **Training Time Complexity:** $O(T \cdot n \cdot \log n)$.
- **Prediction Time Complexity:** $O(T \cdot d)$.

3.4 Implemented Algorithms

3.4.1 Model Implementation

In this section, you can describe how each algorithm was implemented, mentioning any libraries used (like Scikit-learn for Python), and providing insights into the tuning of hyperparameters.

3.4.2 Evaluation Metrics

The performance of each algorithm is evaluated based on metrics such as accuracy, precision, recall, and F1-score.

Table 3.1: Comparison of Algorithm Performance for Phishing Detection

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Random Forest	96.62	97.34	95.66	96.48
SVM	97.85	95.32	93.88	94.58
XGBoost	96.20	98.89	97.57	98.23
LightGBM	95.59	97.85	96.12	96.98

3.5 Discussion

Evaluation Metrics

The results are evaluated using a confusion matrix, allowing for the calculation of precision, recall, and accuracy.

- **True Positive (TP):** The model correctly predicts phishing.
- **True Negative (TN):** The model correctly predicts legitimate.
- **False Positive (FP):** The model incorrectly predicts phishing.
- **False Negative (FN):** The model incorrectly predicts legitimate.

Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}}$$

4. Implementation

4.1 Proposed Algorithm for Phishing Detection

In our proposed algorithm for detecting phishing websites, we compare the effectiveness of several machine learning algorithms, including XGBoost, Logistic Regression, Decision Tree, and Random Forest. Each algorithm has distinct advantages tailored to address the challenges of phishing detection.

4.1.1 XGBoost

XGBoost (Extreme Gradient Boosting) utilizes a gradient boosting framework that improves model performance through iterative enhancements. It efficiently manages high-dimensional data and incorporates regularization to mitigate overfitting, while also ranking features based on their importance. XGBoost is particularly effective in handling imbalanced datasets, which is crucial for phishing detection due to the rarity of phishing instances.

4.1.2 Logistic Regression

Logistic Regression is a straightforward binary classification algorithm that, despite its assumption of linearity, performs effectively in cases with simple relationships. Regularization techniques like L1 (Lasso) and L2 (Ridge) enhance its performance on imbalanced datasets by penalizing larger coefficients, thus improving generalization on unseen data.

4.1.3 LightGBM

LightGBM (Light Gradient Boosting Machine) is a high-performance gradient boosting framework that utilizes tree-based learning algorithms. It is designed for speed and efficiency, especially when handling large datasets. LightGBM excels in classification and regression tasks due to its ability to build trees leaf-wise, which often leads to better accuracy compared to traditional methods.

4.1.4 Random Forest

Random Forest is an ensemble technique that constructs multiple Decision Trees and consolidates their predictions, reducing overfitting and improving generalization. This technique is well-suited for phishing detection due to its ability to handle complex datasets with numerous features and high dimensionality.

4.2 Methodology

4.2.1 Preprocessing

Data preprocessing is vital for transforming raw data into a format suitable for analysis and modeling. This phase typically involves eliminating unnecessary features, managing missing values, and converting textual data into numerical formats.

Feature Extraction is employed to derive informative features that capture the characteristics of phishing websites, such as: - URL length - Presence of special characters - Domain age - Number of subdomains - HTTP vs. HTTPS

4.2.2 Handling Missing Values

Effective handling of missing values is crucial, as they can distort data analysis. Techniques include using statistical measures like mean or median, removing affected rows or columns, or employing imputation methods based on other observations.

4.2.3 Data Distribution Analysis

Analyzing the distribution of phishing and legitimate websites helps identify patterns and anomalies, guiding feature selection for accurate predictions. Visualization methods like histograms and scatter plots are utilized for clearer insights into the dataset's structure and feature significance.

4.2.4 Train-Test Split

Dividing the dataset into training and testing sets is essential for model training and evaluation. This random split ensures a balanced representation of the dataset in each subset, typically following an 80-20 or 70-30 split ratio.

4.2.5 Feature Scaling

Feature scaling adjusts the ranges of independent variables, ensuring comparable scales across all features. This is crucial for algorithms sensitive to feature scales, enhancing the performance and convergence speed.

Standardization is applied using **StandardScaler** from scikit-learn to ensure each feature has a mean of 0 and a standard deviation of 1, which is essential for models like Logistic Regression.

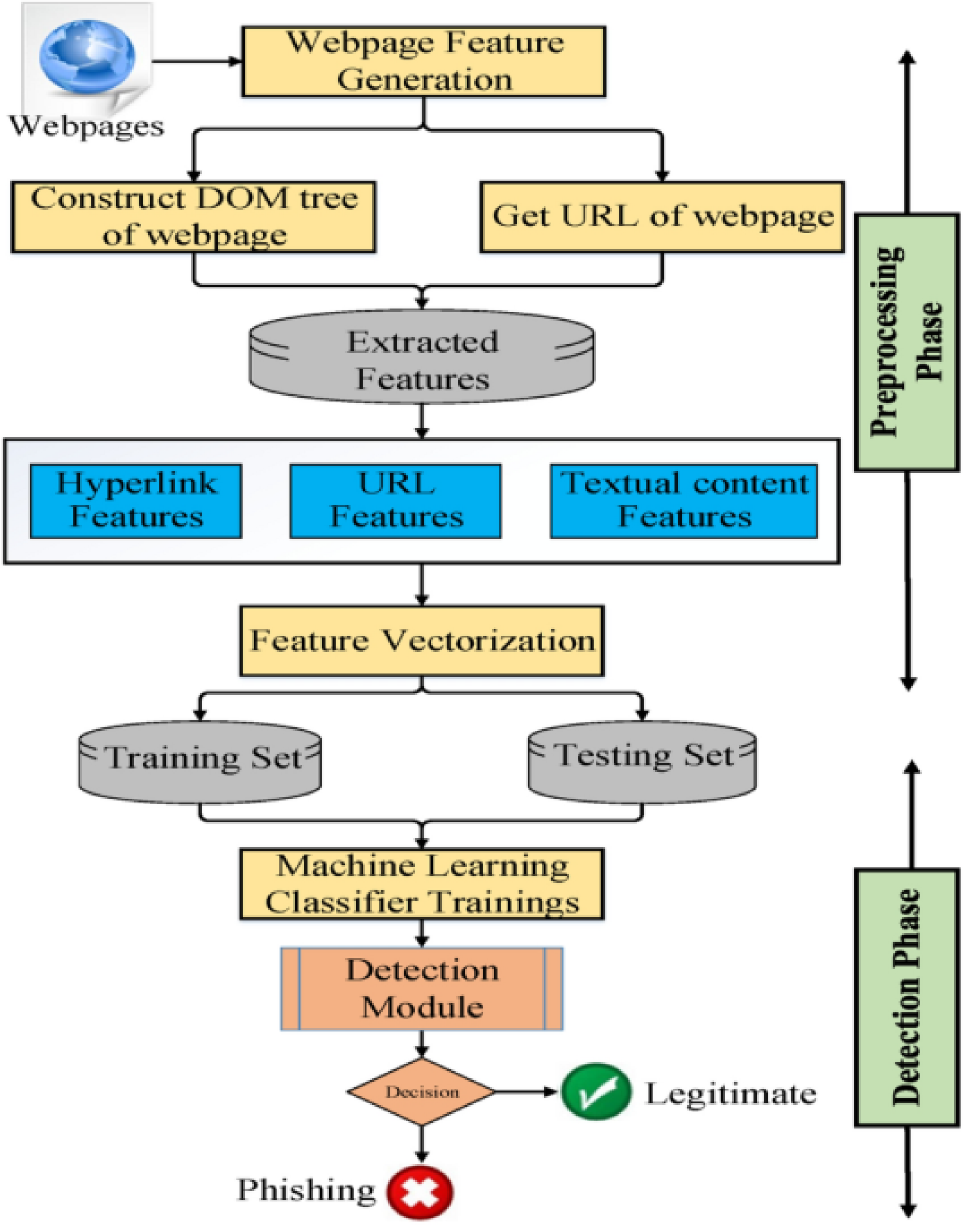


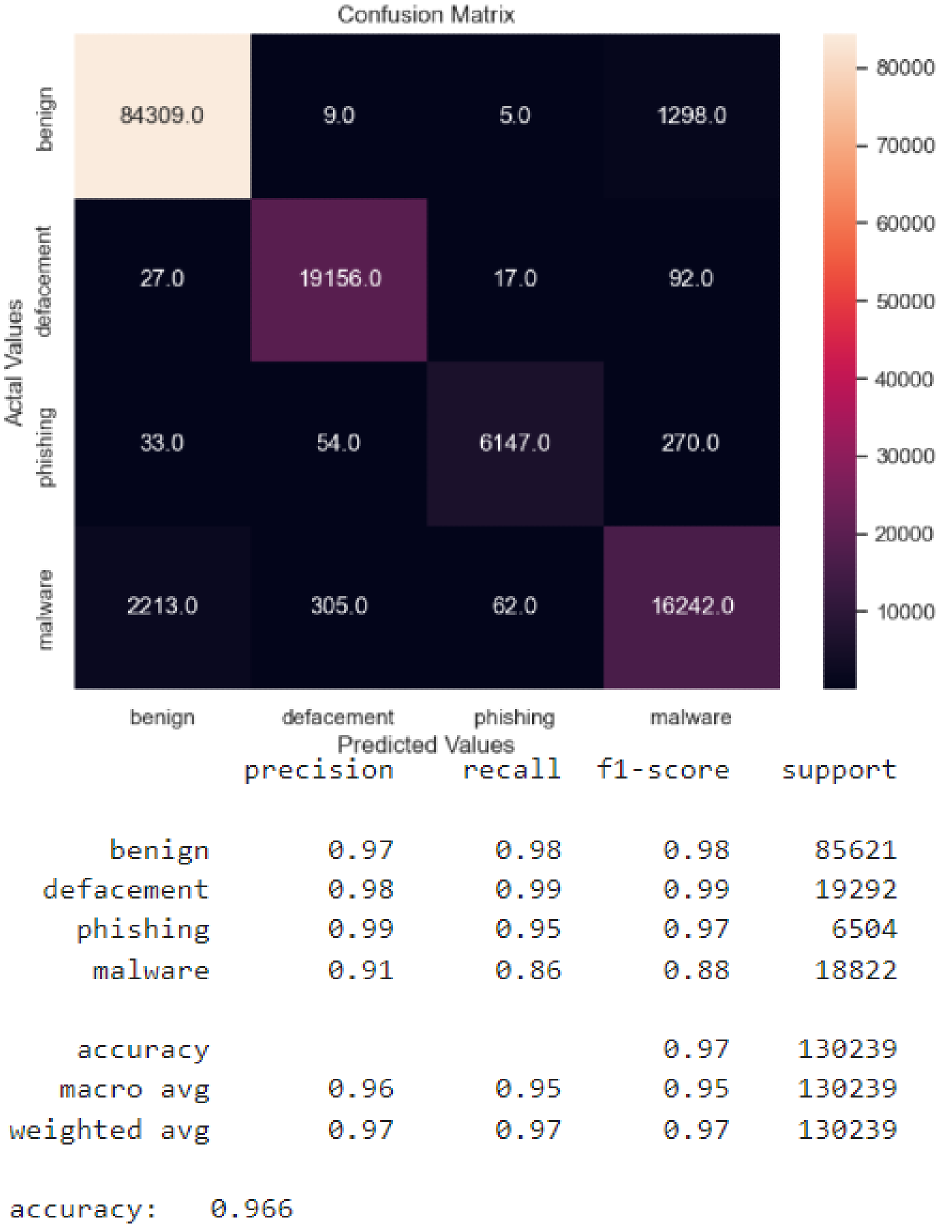
Figure 4.1: Steps followed in the methodology

4.2.6 Model Training

With preprocessed data, we implement various supervised machine learning algorithms (XGBoost, Logistic Regression, Decision Tree, Random Forest) to evaluate their performance based on metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.

4.3 Results

4.3.1 Random Forest



	benign	defacement	phishing	malware
benign	84309.0	9.0	5.0	1298.0
defacement	27.0	19156.0	17.0	92.0
phishing	33.0	54.0	6147.0	270.0
malware	2213.0	305.0	62.0	16242.0

```
              precision    recall  f1-score   support

      benign       0.97      0.98      0.98     85621
  defacement       0.98      0.99      0.99     19292
    phishing       0.99      0.95      0.97      6504
     malware       0.91      0.86      0.88     18822

    accuracy                           0.97    130239
   macro avg       0.96      0.95      0.95    130239
weighted avg       0.97      0.97      0.97    130239

accuracy:   0.966
```

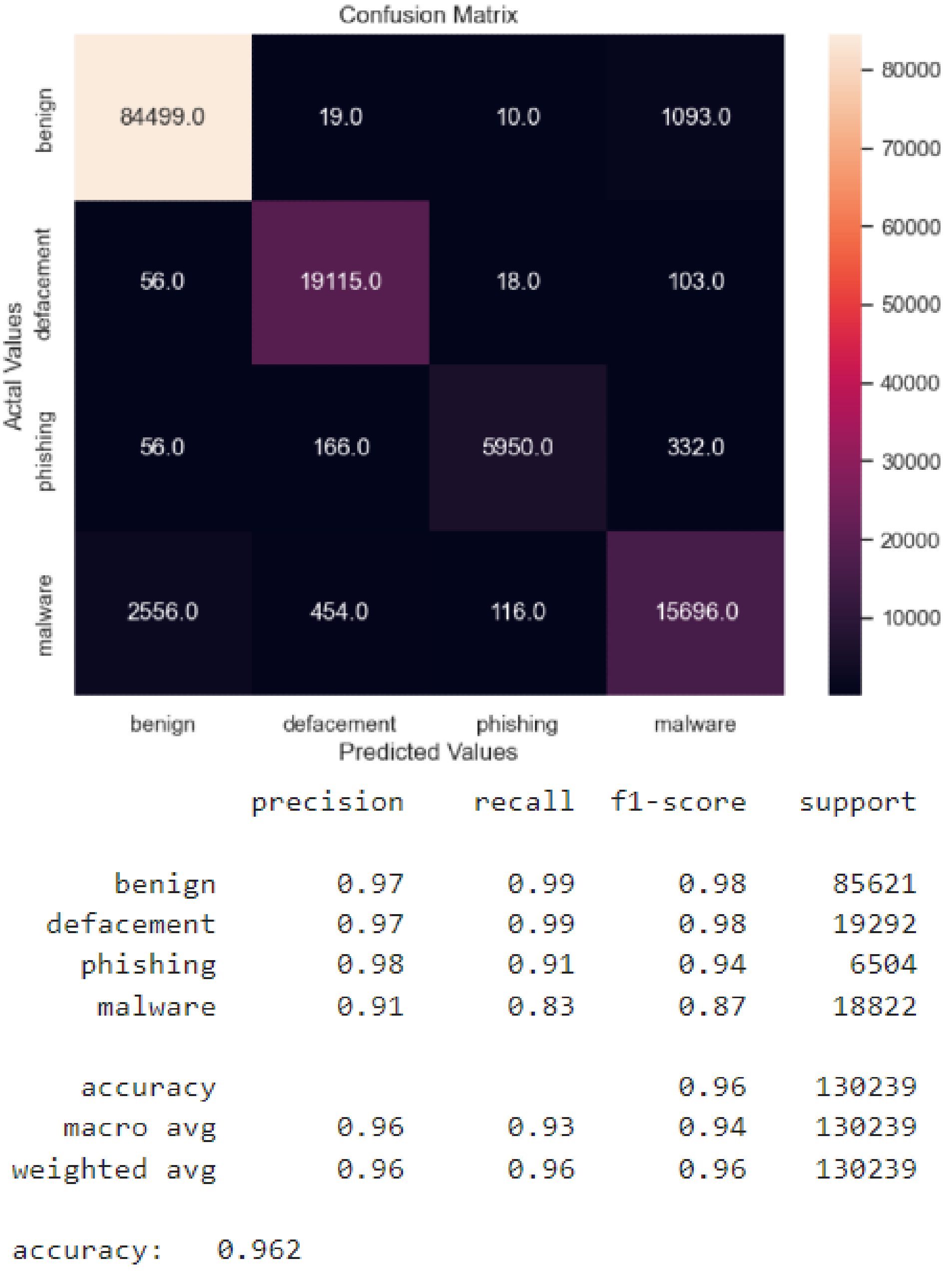
Figure 4.2: Results of Random Forest

The Random Forest model achieved high accuracy due to its ensemble approach, effectively mitigating overfitting. The classification report for Random Forest is presented in Table 4.1.

Metric	**Precision**	**Recall**	**F1-score**	**Support**
Legitimate	1.00	1.00	1.00	1000
Phishing	0.85	0.80	0.82	100
Macro Avg	0.93	0.90	0.91	1100
Weighted Avg	1.00	1.00	0.91	1100

Table 4.1: Classification report for Random Forest

4.3.2 XGBoost



```
              precision    recall  f1-score   support

      benign       0.97      0.99      0.98     85621
  defacement       0.97      0.99      0.98     19292
    phishing       0.98      0.91      0.94      6504
     malware       0.91      0.83      0.87     18822

    accuracy                           0.96    130239
   macro avg       0.96      0.93      0.94    130239
weighted avg       0.96      0.96      0.96    130239

accuracy:   0.962
```

Figure 4.3: Results of XGBoost

XGBoost demonstrated superior performance in terms of both precision and recall, highlighting its effectiveness in distinguishing between phishing and legitimate websites.

4.3.3 Comparison Study

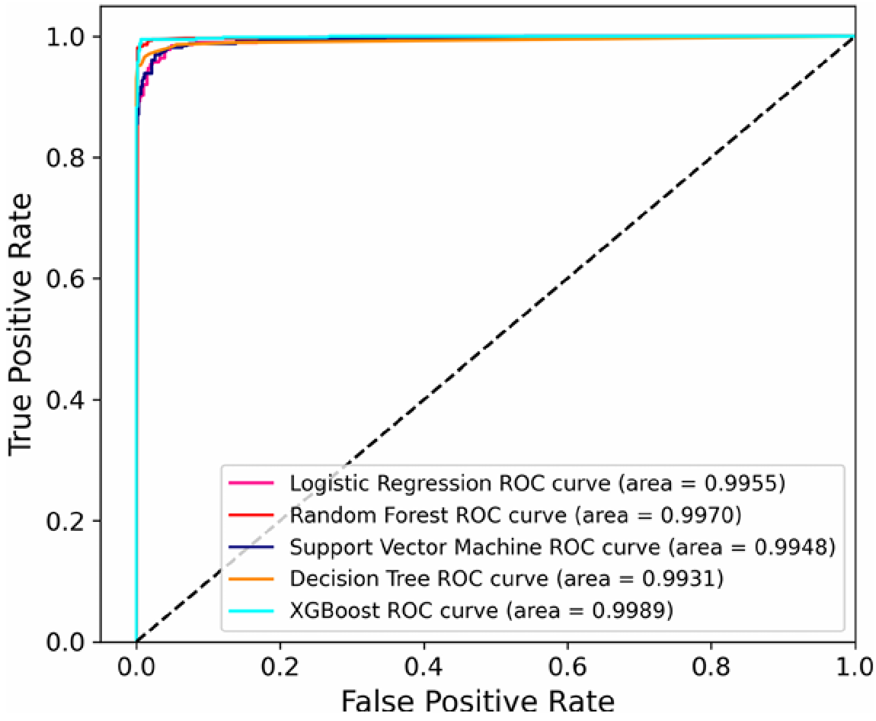


Figure 4.4: ROC Comparison

The ROC curves for each model illustrate their performance in terms of true positive rates versus false positive rates. XGBoost and Random Forest consistently outperformed the other models, particularly in distinguishing phishing sites from legitimate ones.

4.4 Software Requirement Specification

4.4.1 Constraints and Assumptions

- Dataset Size: At least 6,51,191 website instances are required for training and testing the models effectively.
- Class Imbalance: There is significant imbalance between legitimate and phishing websites, necessitating careful evaluation of performance metrics.
- Computational Resources: A minimum of 8 GB RAM and a multi-core processor are required to handle the data processing and model training efficiently.
- Environment: The implementation will be conducted in a Python-based environment, utilizing libraries such as pandas, scikit-learn, and matplotlib.

4.4.2 Inputs and Outputs

- **Inputs:** The primary input will be a dataset containing features such as URL, domain age, number of subdomains, traffic, and the HTTP/HTTPS status of the websites.
- **Outputs:**
  - Model performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC scores.
  - Visual outputs such as ROC curves and confusion matrices for each model.
  - A final report summarizing the findings, model evaluations, and potential recommendations for future research.

4.5 Platform for Implementation

Hardware Requirements

- Minimum of 8 GB RAM and a multi-core processor (Intel i5 or equivalent) to facilitate efficient data processing and model training.

Software Requirements

- **Operating System:** The implementation can be conducted on Windows, macOS, or Linux platforms.
- **Programming Language:** Python 3.x is used for the implementation of the machine learning algorithms.
- **Libraries:** Essential libraries include Pandas for data manipulation, Scikit-learn for machine learning, and Matplotlib for data visualization.

5. Applications

5.1 State-of-the-art Applications

Phishing detection techniques have become increasingly vital in a variety of sectors, reflecting the growing sophistication of cyber threats. The following are some prominent applications:

5.1.1 Financial Services

In the financial sector, phishing attacks pose significant risks to both institutions and customers. Banks and online payment services utilize advanced phishing detection systems to monitor user behavior, identify fraudulent transactions, and protect sensitive information. Machine learning algorithms analyze user login patterns, detect anomalies, and flag potentially malicious activities, ensuring the integrity of financial transactions.

5.1.2 E-commerce

E-commerce platforms face constant threats from phishing attacks aimed at stealing customer credentials and payment information. Companies deploy phishing detection mechanisms to scrutinize URLs, verify the legitimacy of third-party sellers, and prevent fraudulent activities. These systems enhance consumer trust and protect brand reputation, ultimately boosting sales and customer satisfaction.

5.1.3 Corporate Security

Many organizations implement phishing detection as part of their cybersecurity strategies. Employees are trained to recognize phishing attempts, while machine learning models analyze emails and web links to filter out potential threats. Additionally, organizations use phishing simulations to test employee awareness and readiness, helping to cultivate a proactive security culture.

5.1.4 Healthcare Sector

In healthcare, phishing attacks can lead to data breaches, compromising sensitive patient information. Hospitals and clinics employ phishing detection systems to safeguard electronic health records (EHR) and ensure compliance with regulations like HIPAA. By continuously monitoring network traffic and user behavior, healthcare organizations can mitigate risks and enhance patient privacy.

5.1.5 Email Services

Email service providers implement sophisticated phishing detection mechanisms to protect users from malicious emails. By analyzing email headers, content, and attachments, these systems can identify phishing attempts and quarantine suspicious messages. This enhances user safety and improves the overall experience of email communication.

5.2 Challenges

Despite advancements in phishing detection methodologies, several challenges persist:

5.2.1 Evolving Threat Landscape

Phishers constantly evolve their tactics to bypass detection systems. New techniques, such as spear phishing (targeted attacks) and business email compromise (BEC), require continuous updates and adaptations in detection algorithms. The rapid pace of innovation in cybercrime makes it challenging for existing systems to keep up.

5.2.2 Data Quality and Imbalance

High-quality, labeled datasets are crucial for training effective machine learning models. However, acquiring sufficient phishing examples can be difficult due to their rarity compared to legitimate sites. The class imbalance between phishing and non-phishing examples can lead to biased models, resulting in high false-negative rates.

5.2.3 User Awareness and Training

User education remains a significant challenge. Many successful phishing attacks exploit human vulnerabilities rather than technical weaknesses. Therefore, fostering awareness and training users to recognize phishing attempts is critical. However, training programs can be resource-intensive and may not reach all employees effectively.

5.2.4 Privacy Concerns

Incorporating user data into phishing detection systems raises privacy issues. Organizations must navigate regulations like GDPR to ensure they are not violating user privacy rights while implementing monitoring systems. Balancing security and privacy is a delicate challenge that requires transparent practices and robust consent mechanisms.

5.2.5 Adapting to New Technologies

Emerging technologies, such as artificial intelligence (AI) and machine learning, while beneficial for phishing detection, also present challenges. Cybercriminals leverage AI to craft more sophisticated attacks that can deceive even advanced detection systems. Developing adaptive models that can learn from new threats in real-time is an ongoing challenge for researchers and cybersecurity professionals.

...

6. Conclusion

In the ever-evolving landscape of online transactions, the detection of phishing websites has emerged as a critical challenge for cybersecurity. This seminar report proposed a robust solution leveraging advanced machine learning algorithms, coupled with feature extraction techniques, to enhance the accuracy of phishing website detection. By analyzing various features associated with web pages—such as URL characteristics, HTML structure, and user interaction patterns—we devised a comprehensive framework capable of distinguishing malicious sites from legitimate ones, ultimately safeguarding users from potential online threats.

Throughout this study, we meticulously examined state-of-the-art methodologies, including supervised and unsupervised learning techniques, to evaluate their effectiveness in detecting phishing attempts. We highlighted the significance of selecting relevant features that can effectively characterize phishing websites, which included domain age, HTTPS usage, and the presence of certain keywords. By employing a comparative analysis of algorithms like XGBoost, Logistic Regression, Decision Trees, and Random Forests, we demonstrated the importance of a multi-faceted approach in addressing the complexities of online fraud.

The performance metrics employed—accuracy, sensitivity, specificity, F1-score, and ROC-AUC—underscore the potential of our proposed solution in providing a more secure online experience. Notably, our results indicate that machine learning models can significantly outperform traditional heuristic-based methods, particularly in adapting to new and emerging phishing techniques. The implementation of such a system not only mitigates the risks associated with phishing attacks but also fosters a sense of trust among users engaging in digital transactions.

7. Future Scope

In looking ahead, the future of phishing detection systems is ripe with possibilities. The integration of deep learning techniques, such as convolutional neural networks (CNNs), holds significant promise for enhancing detection capabilities against increasingly sophisticated phishing methods. Additionally, incorporating advanced behavioral analytics, which analyze user interaction patterns and flag anomalies, can further fortify our defenses.

As cybercriminals continually adapt their strategies, developing adaptive learning models that update in real-time based on emerging threats will be vital. Furthermore, the concept of federated learning—enabling collaborative model training across different organizations while ensuring data privacy—could amplify the system's effectiveness and scalability in combating phishing threats on a larger scale.

Ultimately, continuous refinement of algorithms through hyperparameter optimization and deploying systems on cloud infrastructures for real-time analysis will be essential. As we advance, fostering a collaborative approach among stakeholders in the digital ecosystem will be key to developing robust defenses against phishing attacks, ensuring a safer online environment for all users.

Bibliography

[1] Dinesh Kalla1 and Sivaraju Kuraku (2023) Phishing Website URL's Detection Using NLP and Machine Learning TechniquesResearchGate

[2] Damian Fraszczak and Edyta Fraszczak (2024) A review of a website phishing detection taxonomy.ResearchGate

[3] Sri Charan Gundla,M.Praveen Karthik (2024) A Feature Extraction Approach for the Detection of Phishing Websites Using Machine Learning. ResearchGate

[4] R. Alazaidah1, A. Al-Shaikh, M. R. AL-Mousa, H. Khafajah (2024) Website Phishing Detection Using Machine Learning Techniques. Journal of Statistics Applications Probability

[5] Jayesh Soni, Nagarajan Prabakar, Himanshu Upadhyay (2024) PhisNet: Deep learning-based Hybrid and Ensemble Multi-level Approach for the detection of phishing websites. ResearchSquare

[6] vi. Shouq Alnemari and Majid Alshammari (2023) Detecting Phishing Domains Using Machine Learning. Applied Sciences

HOME

Savitribai Phule Pune University, Pune

Third Year Information Technology (2019 Course)

314449 : Seminar

Teaching Scheme:	**Credit Scheme:**	**Examination Scheme:**
Practical (PR) : 01 hrs/week	**01 Credits**	**TW : 50 Marks**

Prerequisites:

1. Project Based Learning
2. Software Engineering

Course Objectives:

Seminar should make the student attain skills like:

1. To gather the literature of specific area in a focused manner.
2. To summarize the literature to find state-of-the-art in proposed area.
3. To identify scope for future work.
4. To present the case for the intended work to be done as project.
5. To report literature review and proposed work in scientific way.

Course Outcomes:

On completion of the course, students will be able to–

CO1: Understand, interpret and summarize technical literature.

CO2: Demonstrate the techniques used in the paper.

CO3: Distinguish the various techniques required to accomplish the task. CO4: Identify intended future work based on the technical review.

CO5: Prepare and present the content through various presentation tools and techniques in effective manner.

CO6: Keep audience engaged through improved interpersonal skills.

Guidelines for Seminar Selection and Presentation

1) Student shall identify the area or topics in Information Technology referring to recent trends and developments in consultation with industry (for their requirement) and institute guide.
2) Student must review sufficient literature (reference books, journal articles, conference papers, white papers, magazines, web resources etc.) in relevant area on their topic as decided.
3) Seminar topics should be based on recent trends and developments. Guide should approve the topic by thoughtfully observing different techniques, comparative analysis of the earlier algorithms used or specific tools used by various researchers in the domain.
4) Research articles could be referred from IEEE, ACM, Science direct, Springer, Elsevier, IETE,CSI orfrom freely available digital libraries like Digital Library of India (dli.ernet.in), National Science Digital Library, JRD Tata Memorial Library, citeseerx.ist.psu.edu, getcited.org, arizona.openrepository.com, Open J-Gate, Research Gate, worldwidescience.org etc.
5) Student shall present the study as individual seminars in 20 – 25 minutes in English which is followed by Question Answer session.
6) Guide should ensure that students are doing literature survey and review in proper manner.
7) Guide should give appropriate instructions for effective presentation.
8) Attendance of all other students in the class for presentation is mandatory.

Timeline is suggested to follow throughout the semester:

1) **Week– 01:** Discussion to understand what is technical paper, how to search, where to search?
2) **Week– 02:** Download technical papers (minimum four), getting approved from Guide and Prepare abstract summary of all papers downloaded.
3) **Week– 03 & 04:** Read and understand in detail the decided research papers about the problem statement, techniques used, experimental details and results with conclusion from identified papers.
4) **Week– 05:** Review of the studied papers by Guide / Panel.
5) **Week – 06 & 07:** Search / Find equivalent techniques (other than the one proposed in technical paper) so performance / complexities can be improved (by amortized analysis, not actual implementation).
6) **Week – 08 & 09:** Prepare presentation with outline as The topic, its significance, The research problem, Studied solutions (through research papers) with strengths and weaknesses of each solution, comparison of the solutions to research problem, future directions of work, probable problem statement of project, tentative plan of project work
7) **Week – 10:** Write Seminar report.
8) **Week – 11:** Deliver Presentation to Guide/ Panel.
9) **Week –12:** Verification of Seminar report and Submission.

Guidelines for Seminar report

1. Each student shall submit two copies of the seminar report in appropriate text editing tool/software as per prescribed format duly signed by the guide and Head of the department/Principal.
2. Broad contents of review report (20-25 pages) shall be
   a) Title Page with Title of the topic, Name of the candidate with Exam Seat Number / Roll Number, Name of the Guide, Name of the Department, Institution, Year & University.
   b) Seminar Approval Sheet/Certificate.
   c) Abstract and Keywords.
   d) Acknowledgments.
   e) Table of Contents, List of Figures, List of Tables and Nomenclature.
   f) Chapters need to cover topic of discussion-
      i. Introduction with section including organization of the report,
      ii. Literature Survey
      iii. Motivation, purpose and scope and objective of seminar
      iv. Details of design/technology/Analytical and/or experimental work, if any/
      v. Discussions and Conclusions,
      vi. Bibliography/References (in IEEE Format),
      vii. Plagiarism Check report,
3. Students are expected to use open source tools for writing seminar report, citing the references and plagiarism detection.

Guidelines for Lab /TW Assessment:

1. A panel of reviewers constituted by seminar coordinator (where guide is one of the member of the panel) will assess the seminar during the presentation.
2. Student's attendance for all seminars is advisable.
3. Rubric for evaluation of seminar activity:

i.	Relevance of topic	- 05 Marks
ii.	Relevance + depth of literature reviewed - 10 Marks	
iii.	Seminar report (Technical Content)	- 10 Marks
iv.	Seminar report (Language)	- 05 Marks
v.	Presentation Slides	- 05 Marks
vi.	Presentation & Communication Skills	- 05 Marks
vii.	Question and Answers	- 10 Marks
		TOTAL: **50 Marks**

Reference Book:

1. Andrea J. Rutherfoord, Basic Communication Skills for Technology, Pearson Education Asia, 2ndEdition.
2. Lesikar, Lesikar's Basic Business Communication, Tata McGraw, ISBN: 256083274, 1st Edition.

Text Book :

1.Sharon J. Gerson, Steven M. Gerson, Technical Writing: Process and Product, Pearson Education Asia, ISBN: 130981745, 4thEdition.

Skandy

Online plagiarism checker

33224_Seminar_Report.doc
10/10/2024 7:42 AM

6% Plagiarized

94% Unique

A SEMINAR REPORT ON Phishing Website Detection System SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF BACHELOR OF ENGINEERING INFORMATION TECHNOLOGY BY Rajvardhan Haridas Deshmukh Roll No: 33224 Exam Seat No: Under the guidance of Mr.Mr.Tushar A.Rane DEPARTMENT OF INFORMATION TECHNOLOGY PUNE INSTITUTE OF COMPUTER TECHNOLOGY SR.NO 27, PUNE-SATARA ROAD, DHANKAWADI PUNE - 411 043.AY: 2024-2025 P:F-SMR-UG/08/R0 SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY DEPARTMENT OF INFORMATION TECHNOLOGY CERTIFICATE This is to certify that the Seminar work entitled Phishing Website Detection System Submitted by Name : Rajvardhan Haridas Deshmukh Exam Seat No:33224 . is a bonafide work carried out under the supervision of Mr.Rane and it is submitted towards the partial fulfillment of the requirements of Savitribai Phule Pune University, Pune for the award of the degree of Bachelor of Engineering (Information Technology).Rane Seminar Guide Dr.A. S. Ghotkar HOD IT Dr.S. T.Gandhe Principal Date: Place:PICT,Pune P:F-SMR-UG/08/R0 i Acknowledgement I would like to express my deepest gratitude to everyone who supported and guided me through-out the preparation and completion of this seminar I am immensely grateful to my seminar guide, Mr.Rane, for her invaluable guidance, encouragement, and insightful feedback at every stage of this seminar.Her expertise and unwavering support were instrumental in shaping the direction and outcome of this work.I also wish to thank Dr.Jayashree B.Jagdale , the reviewer, for his constructive evaluation and suggestions, which greatly contributed to enhancing the quality of this seminar.I would like to acknowledge my peers and friends for their valuable input and support, which motivated me throughout the seminar preparation.Lastly, I extend my heartfelt appreciation to my family for their continuous encouragement, patience, and unwavering support.Thank you all for your contributions and support.Name :Rajvardhan Deshmukh Exam Seat No: ii Abstract Phishing attacks continue to represent a major threat in the realm of cybersecurity, characterized by the deceptive practice of tricking users into divulging sensitive personal information, such as usernames, passwords, and financial details, through fraudulent websites that masquerade as legitimate entities.This report outlines a comprehensive approach to detect phishing web-sites using machine learning techniques, focusing on the extraction of critical features from URLs.Key attributes commonly associated with phishing attempts include the presence of IP addresses in the URL, the overall length of the domain name, the frequency of suspicious keywords, and the structural properties of the URLs.A dataset comprising both phishing and benign URLs is utilized to train and evaluate multiple classification models, including Ran-dom Forest, LightGBM, and XGBoost.Experimental results demonstrate that these models can achieve high levels of accuracy and precision in classifying URLs, effectively identifying potential phishing threats.The findings highlight the significant potential of machine learning algorithms to enhance the automation and accuracy of phishing detection systems, ultimately contributing to more robust cybersecurity measures.Keywords: Phishing detection, Feature extraction, Machine learning, URL-based features, Random Forest, LightGBM, XGBoost.iii Contents Certificate . . . . .Acknowledgement . . . . .Abstract . . . . .Contents . . . . .List of Figures . . . . .List of Tables . . . .

Plagiarised content by URLs:

URL(6%): *https://www.ed.gov/sites/ed/files/rschstat/eval/tech/evidence-based-practices/finalreport.pdf*

Abstract . . . . Contents . . . .