DETECTING PHISHING URLS USING DATAMINING TECHNIQUES

THIS STUDY LEVERAGES DATA MINING TECHNIQUES AND MACHINE LEARNING TO DETECT PHISHING URLS, ACHIEVING 94% ACCURACY ON A 2024 DATASET BY ANALYZING FEATURES LIKE URL LENGTH AND DOMAIN INFORMATION. THE FINDINGS SHOWCASE A ROBUST, REAL-WORLD-READY MODEL FOR ENHANCING CYBERSECURITY.

AUTHORS

Mr. Htin Linn, Ms. May Khine Soe, Mr. Nyan Lin Aung

AFFILIATIONS

Rangsit University

INTRODUCTION

What is Phishing?
- A form of cybercrime where attackers deceive individuals into revealing sensitive information

Cybersecurity Challenge
- Financial losses, identity theft, and data breaches
- Advanced abilities of attackers & Traditional method fail to handle
- Manual analysis impractical for high volume of web traffic

DataMining & MachineLearning as a Solution
- Machine learning technique (Logistic Regression) for binary classification tasks, offering a balance of accuracy and efficiency
- Data mining, extraction of meaningful patterns, combined with logistic regression provides scalable, real-world-ready phishing detection model.

Literature Review and Gap

Previous study rely on blacklist-based approach & simple rule-based, unable to detect zero-day phishing attacks leading to emphasis of more sophasticated machine learning-based approaches. However, Jeeva & Rajsingh(2016) demonstrated importance of proper feature selection in URL analysis, identifying key characteristics such as URL length, special character frequency, and domain attributes and so on.

Previous Studies	Current Study
• Lack of updated, old datasets • Feature changes rarely analyzed • Lack of model adaptability • Limited real-world testing	• Comprehensive dataset (2019-2024) • Tracks feature evolution • Iterative training approach model adaptability • Focuses on practical applications

OBJECTIVE

To develop an effective phishing URL detection system using logistic regression to enhance accuracy and adaptability for real-world cybersecurity applications

METHODOLOGY

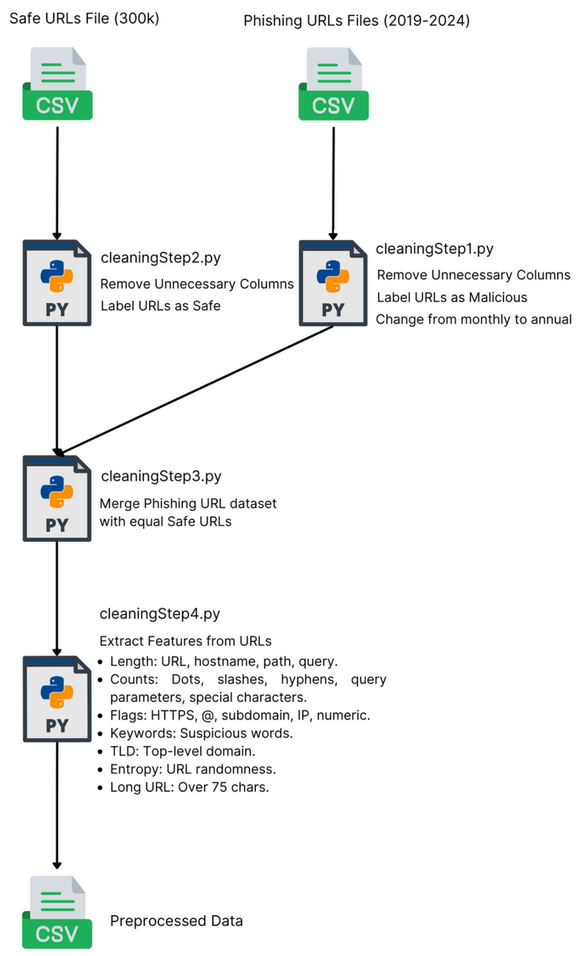
1. Data Processing
   - Combined & balanced data from Kaggle (safe URLs) and JPCERT/CC (phishing URLs)
   - Extracted key features: URL length, suspicious keywords, special characters, domain info

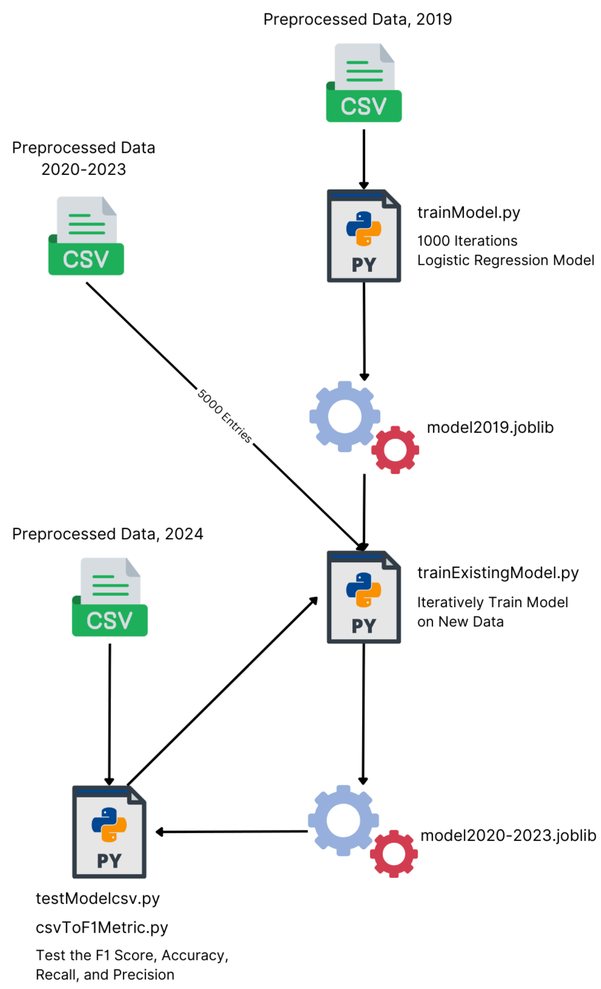
2. Model Development
   - Logistic Regression model
   - Initial training on 2019 data (80-20 split)
   - Incremental training: 5,000 entries/year (2020-2024)
   - Evaluated against 2024 benchmark data

RESULT/FINDINGS

- The final model achieved 94% accuracy and balanced F1-scores (0.94) for detecting both malicious and safe URLs, outperforming control models.
- Comparison with Control Models:
  - Control Model 1 (Combined Dataset): Similar accuracy (93%) but lower recall (88%) for malicious URLs.
  - Control Model 2 (Single Year): Similar accuracy (93%) but struggled with generalization, missing malicious URLs (recall 86%).
- Key Metrics (Final Model on 2024 Data):
- Precision: 94% for both malicious and safe URLs.
- Recall: 94% for both classes.

Visualizing the Data Pipeline: From Raw URLs to Machine-Learning-Ready Features Through Preprocessing and Feature Extraction





Visualization of the Iterative Model Training Process: This diagram outlines the step-by-step approach used in the study, starting with training a Logistic Regression model on 2019 data, iteratively updating it with data from 2020-2023, and evaluating its performance on 2024 data. Key scripts and processes for training, saving, and testing the model are highlighted.

Iterative Model Performance Across Different Years on 2024 Benchmark

Malicious URLs

Metric	2019 (Initial)	2019-2020	2019-2021	2019-2022	2019-2023
Precision	0.96	0.98	0.94	0.98	0.94
Recall	0.63	0.78	0.93	0.86	0.94
F1-Score	0.76	0.87	0.93	0.92	0.94

Safe URLs

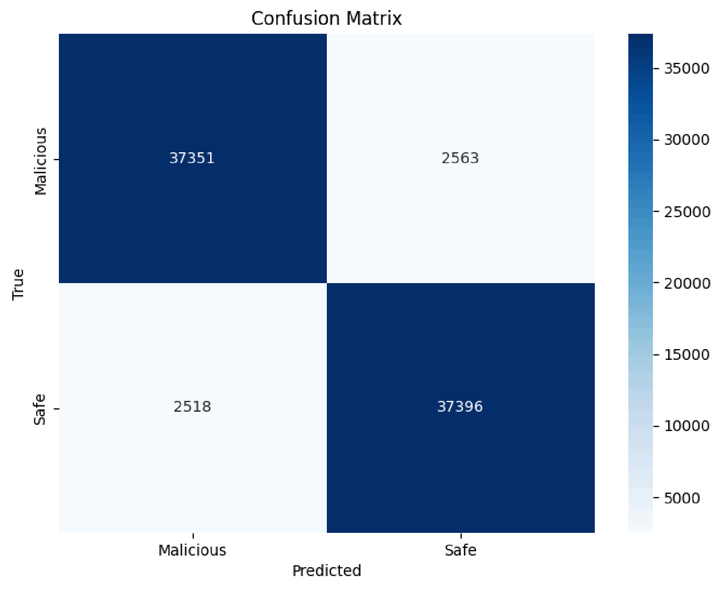
Metric	2019 (Initial)	2019-2020	2019-2021	2019-2022	2019-2023
Precision	0.73	0.82	0.93	0.88	0.94
Recall	0.97	0.99	0.94	0.99	0.94
F1-Score	0.83	0.89	0.94	0.93	0.94

Comparison of Model Performance on 2024 Benchmark

Malicious URLs

Metric	Iterative Model	Combined Data	Single Random Data
Accuracy	0.94	0.93	0.93
F1-Score	0.94	0.93	0.93
Malicious Class			
- Precision	0.94	0.98	0.99
- Recall	0.94	0.88	0.86
Safe Class			
- Precision	0.94	0.89	0.88
- Recall	0.94	0.99	0.99

Confusion Matrix of the Iterative Model's Final Performance on 2024 Benchmark



ANALYSIS

- Model Evaluation: The model's performance improved with each year of data added. The final model, using data from 2019–2023, achieved 94% accuracy and strong recall for both malicious and safe URLs.
- Control Models:
  - The combined dataset model showed high precision but missed many malicious URLs (low recall).
  - The 2022 dataset model had high precision for malicious URLs but also low recall, indicating missed phishing cases.
- Feature Analysis: Key features like contains_https, contains_subdomain, and unsafe keywords were important in identifying legitimate and phishing URLs, becoming more predictive as data from newer years was incorporated.

CONCLUSION

This study developed a machine learning model using logistic regression to detect phishing URLs, achieving 94% accuracy. The model improved by incorporating data from 2019 to 2023, adapting to evolving phishing tactics. The final model outperformed other approaches, such as using a combined dataset or training on a single year, with better balance in precision and recall. Key features like contains_https, contains_subdomain, and certain top-level domains were crucial for accurate predictions.

Future Work:
- Explore and compare other more intensive algorithms to further improve performance.
- Add more features (eg. webcertificates.)

REFERENCES

- Starbuck, C. (2023). Logistic regression. In The fundamentals of people analytics. Springer, Cham. https://doi.org/10.1007/978-3-031-28674-2_12
- 8. Jeeva, S. C., & Rajsingh, E. B. (2016). Intelligent phishing URL detection using association rule mining. Human-Centric Computing and Information Sciences, 6(10). https://doi.org/10.1186/s13673-016-0064-3
- Kumar, S. (2019). Malicious and Benign URLs [Data set]. Kaggle. https://www.kaggle.com/datasets/siddharthkumar25/malicious-and-benign-urls
- JPCERT/CC. (n.d.). phishurl-list. GitHub. https://github.com/JPCERTCC/phishurl-list/tree/main