Strengthening Small and Medium-Sized Businesses' Cybersecurity: A Machine Learning-based Phishing Classification Model

by Steven C. Song

B.S. in Engineering Management, May 2006, United States Military Academy
S.M. in Engineering and Management, May 2016, Massachusetts Institute of Technology

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

August 31, 2023

Praxis directed by

John Fossaceca
Professorial Lecturer of Engineering and Applied Science

Shahryar Sarkani
Professorial Lecturer of Engineering and Applied Science

The School of Engineering and Applied Science of The George Washington University certifies that Steven C. Song has passed the Final Examination for the degree of Doctor of Engineering as of July 29, 2023. This is the final and approved form of the Praxis.

Strengthening Small and Medium-Sized Businesses' Cybersecurity:
A Machine Learning-based Phishing Classification Model

Steven C. Song

Praxis Research Committee:

John Fossaceca, Professorial Lecturer of Engineering and Applied Science
Praxis Director

Shahryar Sarkani, Professorial Lecturer of Engineering and Applied Science
Committee Member

Muhammad Islam, Professorial Lecturer of Engineering and Applied Science
Committee Member

© Copyright 2023 by Steven C. Song
All rights reserved.

Dedication

To my best friend and partner, Dr. Ella Kim, thank you for supporting me throughout my entire doctoral journey. Words cannot express my gratitude. You have been my most incredible supporter since the first day we met at MIT.

To my beautiful daughter, Miro, thank you. You spent countless hours waiting for Daddy to finish his dissertation. I have more time and energy now for the playground and bubbles. Also, thank you for saying, “Good luck Daddy” on the day of my defense!

Acknowledgments

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Drs. John Fossaceca and Shahryar Sarkani, for their feedback, and most of all patience throughout the entire process.

I would like to acknowledge and thank the Army training-with-industry fellowship. This broadening experience provided the time to think about my research and the practical application of this praxis. I would like to thank the Verizon Global Network Engineering and Network Security teams for sharing their insights on network engineering, data analytics, and cybersecurity. Also, special thanks to Mike and the FBI Internet Crime Complaint Center (IC3).

Abstract of Praxis

Strengthening Small and Medium-Sized Businesses' Cyber Readiness: A Machine Learning-based Phishing Classification Model

Small and medium-sized businesses (SMBs) are priority targets for cybercriminals due to lacking financial, human, and technical resources. Most SMBs cannot afford information security experts or cyber defense tools to defend against cyberattacks, necessitating cost-effective interventions to enhance cybersecurity. In 2022, cybercriminals caused over $10.3 billion in losses to SMBs using various cyberattack types like social engineering, phishing, and ransomware (FBI, 2023). Phishing is the top cyberattack for SMBs, often leading to more severe consequences like ransomware. A machine learning-based classification model can effectively determine phishing website attacks for SMBs.

This praxis demonstrates that a cost-effective phishing detection tool can be developed using open-source tools, historical phishing data, and machine learning to detect phishing website attacks proactively. Using a novel ensemble feature selection technique, this praxis develops a features-based uniform resource locator (URL) parser to identify significant lexical features of phishing websites and evaluates which machine learning classifiers have the best performance for detecting phishing websites.

The contributions of this praxis include the following: First, it demonstrates that the ensemble feature selection with the Random Forest classifier performed the best in classifying phishing websites with up to 97% accuracy. Second, it identifies and validates a global set of features using the latest phishing data sets. Finally, the research introduces a phishing data set for benchmarking machine learning models.

Table of Contents

Dedication ..... **iv**

Acknowledgments ..... **v**

Abstract of Praxis ..... **vi**

List of Figures ..... **xi**

List of Tables ..... **xiii**

List of Acronyms ..... **xv**

List of Glossary of Terms ..... **xvi**

Chapter 1—Introduction **1**

1.1 Background 1

1.2 Research Motivation 2

1.3 Problem Statement 3

1.4 Thesis Statement 3

1.5 Research Objectives 3

1.6 Research Questions and Hypotheses 4

1.7 Scope of Research 5

1.8 Research Limitations 5

1.9 Organization of Praxis 5

Chapter 2—Literature Review **7**

2.1 Introduction 7

2.2 Cybersecurity and Small and Medium-Sized Business 8

2.2.1 Limited Cybersecurity Research on SMBs 9

2.2.2 Insufficient Cyber Defense Resources for SMBs 10

2.2.3 Lack of Publicly Available Cybersecurity Data 11
2.3 SMB Cyber Threat Landscape 11
2.4 Cybersecurity Guidance and Resources for Small Businesses 13
2.5 Phishing – Brief History & Types 14
2.5.1 Email phishing 17
2.5.2 Website phishing 18
2.5.3 Vishing ("Voice Phishing") 18
2.5.4 Smishing (SMS Text + Phishing) 19
2.5.5 Pharming 19
2.6 Phishing Website Detection Techniques 19
2.6.1 User Awareness and Training 20
2.6.2 Heuristic-Based Detection Techniques 22
2.6.3 List-Based Detection Techniques 23
2.6.4 Machine Learning (ML) 24
2.6.4.1 Deep Learning (DL) 29
2.7 Summary of Gaps in Phishing Detection Methods 30
2.8 Summary and Conclusion of ML-based Phishing Detection 32
Chapter 3—Methodology 34
3.1 Introduction 34
3.2 Data Collection and Preprocessing 34
3.2.1 Phishing Website Data Set on Mendeley 2020 (Vrbančič et al., 2020). 36
3.2.2 Phishing Site URLs (Tiwari, 2022) 37
3.2.3 Phishing Website Detection Data Set (Pham, 2022) 37

3.2.4 Validation Phishing Data set 38
3.2.5 Data Preprocessing 38
3.3 Feature Selection Methods 41
3.3.1 Filter Method: Dimension Reduction using Correlation Analysis 43
3.3.2 Dimension Reduction using Principal Component Analysis 45
3.3.3 Wrapper Method - Recursive Feature Elimination 46
3.3.4 Feature Selection using SHAP values 47
3.3.5 Feature Selection by Importance Variable (Embedded) 48
3.3.6 Proposed Ensemble Feature Selection 49
3.3.7 Feature Selection Comparison and Analysis 50
3.4 Model Development 51
3.4.1 Naïve Bayes 52
3.4.2 Support Vector Machine (SVM) 53
3.4.3 K-Nearest Neighbors (KNN) 53
3.4.4 Random Forest (RF) 53
3.4.5 AdaBoost 54
3.4.6 XGBoost 54
3.5 Model Training and Validation 54
3.6 Model Evaluation 56
3.7 Summary and Conclusion 58
Chapter 4—Results 60
4.1 Introduction 60
4.2 Feature Selection Results (Research Question 1) 61

4.3 Summary of Features Selection 69
4.4 Phishing Classification Model Results (Research Questions 2 & 3) 70
4.4.1 Data Set #1 Results (Vrbančič et al., 2020) 71
4.4.2 Data Set #2 (Tiwari, 2022) 71
4.5 External Model Validation 74
4.5.1 Phishing Website Detection Data Set #3 (Pham, 2022) 74
4.5.2 Phishing Validation Data Set #4 77
4.6 Model Summary and Conclusion 79
4.7 Research Questions and Hypotheses Results 81
4.7.1 Hypothesis 1- Significant Features 81
4.7.2 Hypothesis 2 - Phishing Classification Model Evaluation 82
4.7.3 Hypothesis 3 - Random Forest Model 83
Chapter 5—Discussion and Conclusions 84
5.1 Discussion 84
5.2 Conclusions and Recommendations 85
5.3 Contributions to Body of Knowledge 88
5.4 Recommendations for Future Research 89
References 91
Appendix A 99

List of Figures

Figure 1: NIST Cybersecurity Framework Functions (Chidukwani et al., 2022) 10

Figure 2: Top Five Internet Crimes (FBI, 2023) 2022 12

Figure 3: Phishing Attack (Cloudflare, 2023) 15

Figure 4: Growth of Phishing Attacks (APWG, 2023) 16

Figure 5: Common Types of Phishing (Martin, 2022) 17

Figure 6: Real-world Example: Email Phishing Attempt 17

Figure 7: Google Safe Browsing: Malware and Phishing Sites (Rosenthal, 2022) ... 18

Figure 8: Smishing Example (Proofpoint, 2023) 19

Figure 9: Phishing Website Detection Methods 20

Figure 10: Heuristic-based Web Phishing Detection (Vijayalakshmi et al., 2020) ... 22

Figure 11: Whitelist-based Web Phishing Detection (Vijayalakshmi et al., 2020) ... 23

Figure 12: URL Components 25

Figure 13: ML-based Web Phishing Detection Process (Vijayalakshmi et al., 2020)26

Figure 14: Research Methodology 34

Figure 15: Phishing Data Sets 35

Figure 16: Data Pre-Processing 39

Figure 17: Feature Selection Approach 41

Figure 18: Correlation Matrix Heat Map Example 45

Figure 19: Scree Plot Example (Kuhn & Johnson, 2013) 46

Figure 20: Recursive Feature Elimination Example 47

Figure 21: SHAP Feature Selection Example (Lin, 2022) 48

Figure 22: Embedded Feature Selection by Importance 49

Figure 23: Model Training and Validation Split 55

Figure 24: Binary Confusion Matrix..... 56

Figure 25: Research Questions 60

Figure 26: Correlation Matrix..... 63

Figure 27: Features by Importance..... 65

Figure 28: Key Features using Random Forest Classifier 67

Figure 29: Top Features by SHAP values..... 68

Figure 30: SHAP values and Features 69

Figure 31: Principal Component Analysis (PCA) on Data Set #1 – Scree Plot..... 72

Figure 32: RF Important Features using Data Set #2 (Tiwari, 2022)..... 73

Figure 33: Phishing Classification Model with Random Forest Classifier..... 76

Figure 34: Five Important Features for Phishing Classification..... 77

Figure 35: ML Model Results - Accuracy..... 78

Figure A-1: Eigenanalysis of the Correlation Matrix..... 101

List of Tables

Table 1: U.S. Firms by Enterprise Size 8

Table 2: Common SMB Cyber Threats (SBA, 2023) 12

Table 3: Cybersecurity Planning and Assessment Tools 13

Table 4: Relevant ML-based Papers - Contributions and Gaps 29

Table 5: Summary of Phishing Detection Methods 31

Table 6: Literature Review Gaps and Praxis Contributions 32

Table 7: Phishing Data Sets 36

Table 8: Mendeley 2020 Features Grouped (Wei & Sekiya, 2022) 36

Table 9: Phishing Site URLs Features (Tiwari, 2022) 37

Table 10: Phishing Website Detection Data set (Pham, 2022) 38

Table 11: Phishing Data Set Preprocessing: Missing Values and NAs 40

Table 12: Feature Selection Methods Comparison Summary 50

Table 13: Classification Algorithms and Categories 52

Table 14: Features Filter due to Zero Variance 61

Table 15: Features with High Correlations 62

Table 16: Phishing Classification Model - Baseline Features 65

Table 17: Data Set #1 Model Results 71

Table 18: Data Set #2 Model Results 74

Table 19: Data Set #3 Model Results 75

Table 20: Random Forest Results using five predetermined features 79

Table 21: Data Set Records and Features 79

Table 22: Summary of Research Hypotheses 81

Table 23: Significant Features 82

Table 24: Summary of Research Questions and Hypotheses..... 84

Table A-1: Data Set, Description, and Sources 99

Table A-2: Correlation Coefficients and Significant Values for Phishing Features100

Table A-3: Phishing URLs Data Set – 27 Features Parsed 102

Table A-4: Estimated Cost for SMB Phishing Detection System 103

List of Acronyms

AI	Artificial Intelligence
APWG	Anti-Phishing Working Group
CISA	Cybersecurity and Infrastructure Security Agency
DL	Deep Learning
FBI	Federal Bureau of Investigation
FN	False Negative
FP	False Positive
FTC	Federal Trade Commission
ML	Machine Learning
NIST	National Institute of Standards and Technology
RF	Random Forest
SBA	Small Business Administration
SMB	Small and Medium-sized Businesses
TN	True Negative
TP	True Positive
URL	Uniformed Resource Locator

List of Glossary of Terms

- **Ensemble Feature Selection:** Identify the best feature subsets by combining different feature selection methods to overcome each of their single feature selection limitations.
- **Feature Engineering:** Use a combination of features to improve the model's performance.
- **Feature Extraction:** Reduce data set dimensionality by transforming existing features.
- **Feature Selection:** Choose a subset of existing features without a transformation.
- **Malicious website:** A compromised website that attempts to steal, ransom, access a computer, or install malicious malware.
- **Machine Learning:** A subfield of artificial intelligence broadly defined as teaching a machine or computer to learn from data and perform tasks that can mimic human behavior.
- **NIST Cybersecurity Framework functions:** Identify, Protect, Detect, Respond, Recover
  - **NIST Detect -** The detect function outlines the proper steps to detect cybersecurity incidents.
  - **NIST Respond** - The respond function prescribes the necessary response to contain the effects of a possible cybersecurity incident.
- **Phishing** is a cyberattack that uses email or a malicious web link to install malware on your computer or gather confidential data.
- **Random Forest (RF):** A tree-based algorithm that combines multiple trees known as an ensemble method; created by Leo Brieman and Adele Cutler in 2001.
- **R Caret Package:** R programming package that helps train and build machine learning models for classification and regression problems (short for **C**lassification **A**nd **RE**gression **T**raining)

Chapter 1—Introduction

1.1 Background

The global cyber threat landscape is rapidly changing. In 2022, cybercriminals caused over $10.3 billion in U.S. losses, with phishing as the number one crime type with 300,497 complaints (FBI, 2023). Small and medium-sized businesses[1] (SMBs, with fewer than 1,000 employees) cannot sacrifice cybersecurity anymore. The impact of a phishing attack can lead to more catastrophic consequences, such as malware or a data breach.

Cyberattacks are no longer a large enterprise problem, as small and medium-sized businesses are now more susceptible to cyberattacks. The frequency and severity of cyberattacks have increased, impacting SMBs more than the larger enterprises with sufficient budgets for premium tools, technologies, and talent. One of the most persistent cyberattack types for SMBs is phishing.

Phishing is the top cyberattack type for SMBs and is often the leading cause of data breaches (FBI, 2023). A phishing attack can result in much more than monetary losses because the theft of sensitive personal data may have longer-term consequences. Phishing attacks use email or malicious websites to infect a user's computer or to collect sensitive data like credit cards, banking, and passwords (SBA, 2023). In the first quarter of 2020, the anti-phishing working group (APWG) observed only 230,554 phishing attacks; however, in the 3rd quarter of 2022, the number of phishing attacks quintupled and set a record with 1,270,883 total attacks (APWG, 2022). A record high of 4.7 million phishing attacks were reported in 2022 (APWG, 2023).

[1] There are varying definitions of a small and medium-sized business, the praxis is using the Small Business Administration (SBA) 'small' for under 500 employees and 'medium' for under 1,000 employees.

1.2 Research Motivation

Cyberattacks have been growing partially due to the shift toward the digital economy and the adoption of remote work. In 2022, the FBI's Internet Crime Complaint Center (IC3) received 800,944 complaints, costing the American public over ten billion dollars in losses due to the growth of cybersecurity crimes (FBI, 2023). With key watershed events like the Colonial pipeline and SolarWinds, cybersecurity attacks and data breaches financially impact the economy. SMBs employ 90% of the U.S. population and are paramount to innovation and the larger economy. As a system is only strong as its weakest link, small and medium-sized businesses (SMBs) can be more accessible entry points due to lacking technical, financial, and human resources. Therefore, SMBs need to be educated, resourced, and protected against cyberattacks, and the most common attack vector - phishing attacks.

With emerging technologies like artificial intelligence and chatbots (ChatGPT), cybercriminals can overcome previous language barriers and shortcomings of phishing campaigns. Many cybercriminals, which may be non-native English speakers, are weaponizing chatbots to create more sophisticated phishing campaigns, especially with more natural phishing emails with embedded URL links (Fripp, 2022).

Current techniques like blacklisting and whitelisting for phishing website detection are time-consuming and risk becoming outdated quickly. Machine learning (ML) and cybersecurity analytics[2] can better protect SMBs against phishing attacks with minimal capital costs. Therefore, this research aims to demonstrate and test cost-effective models for SMBs to detect phishing website attacks.

[2] While there is no universal definition of cybersecurity analytics, it has been widely adopted that it combines data analytics like machine learning with cybersecurity.

1.3 Problem Statement

Small and medium-sized businesses are priority cyberattack targets resulting in $10.3 billion in losses in 2022 (FBI, 2023).

Small and medium-sized businesses are priority targets due to lacking technical staff, financial resources, and knowledge of safeguarding customers' data. Without proper cyber hygiene and training, SMBs are prime victims of social engineering and phishing attacks. Phishing remains the top cyberattack type impacting up to thirty-two million U.S. SMBs. Also, cybercriminals use phishing attacks to deliver additional malware and ransomware attacks. In Verizon's 2022 DBIR annual report, ransomware represents 25% of data breaches and impacts a larger proportion of SMBs than large organizations (Verizon, 2022a)

1.4 Thesis Statement

A classification model is required to help small and medium-sized businesses (SMBs) detect phishing website attacks.

Unlike large enterprises and government agencies, SMBs have limited budgets and resources to thwart and mitigate cyber threats. Using machine learning and cybersecurity analytics to build a classification model is a suitable and cost-effective cyber solution for preventing and mitigating phishing website attacks.

1.5 Research Objectives

The main objective of this research is to use machine learning to build a classification model for SMB owners to detect phishing websites with high accuracy. This phishing classification model will allow SMBs to use a cost-effective approach to detect phishing website attacks.

Research sub-objectives include:

1. Examine existing cybersecurity resources and solutions for SMBs.
2. Identify and assess the most significant features for classifying phishing website attacks.
3. Build machine learning models to classify websites as phishing or legitimate.
4. Evaluate which machine learning algorithm is most accurate in classifying phishing website attacks.
5. Using key features from the phishing classification model, develop an optimal features-based URL parser and validate on other publicly available phishing data sets.

1.6 Research Questions and Hypotheses

For the aforementioned research objectives, the research questions and corresponding hypotheses are:

RQ1: Which features are the most significant in classifying phishing website attacks?

RQ2: Can a machine learning model be built to classify phishing website attacks?

RQ3: Which machine learning algorithm has the highest accuracy in classifying phishing website attacks?

H1: URL length, domain length, and the number of slashes (‘/’) are the most significant features for classifying phishing website cyberattacks.

H2: A phishing classification model based on machine learning can be developed with at least 80% accuracy.

H3: Random Forest is the most accurate machine learning algorithm for classifying phishing website attacks.

1.7 Scope of Research

While there are numerous research efforts to improve users' awareness of phishing and phishing email detection, there is less work on developing and testing web phishing detection tools for SMBs. Therefore, this research focuses on machine learning solutions using the lexical features of a Uniform Resource Locator (URL) primarily to detect phishing websites. The scope of this research was limited to publicly available phishing data sets, using four publicly available phishing data sets published within the last two years. The most-cited phishing data set consists of 112 URL features and 88,647 websites labeled as legitimate or phishing (Vrbančič et al., 2020). The second data set has 507,195 websites with only a URL column and labels for features (Tiwari, 2022).

1.8 Research Limitations

This research has two significant limitations. First, no official and approved data set for benchmarking research in phishing URL detection for small and medium businesses exist. This research's data sets are derived from publicly available sources such as Alexa, Phish Tank, and OpenPhish and are voluntarily reported. Second, this research assumes the published phishing data sets are accurate and correctly labeled. Therefore, the results may not be representative of all phishing web attacks.

1.9 Organization of Praxis

This praxis consists of five chapters. Chapter two provides a literature review on the SMBs' cyber threat landscape focusing on phishing, cybersecurity gaps for SMBs, and gaps in machine learning applications for phishing detection. Chapter three describes

the machine learning model development from data collection, preprocessing, feature selection, and the performance metrics for model evaluation. Chapter four presents feature selection, phishing classification models, and model validation results. Chapter five provides the discussion, conclusions, contributions, and recommendations for future research.

Chapter 2—Literature Review

2.1 Introduction

Phishing attacks continue to be the top threat for small and medium-sized businesses (SMBs), resulting in losses of billions of dollars (FBI, 2023). A National Cyber Security Alliance report found that 60% of small businesses are out of business within six months after a cyberattack. Despite the persistent threat and potential to cause millions of SMBs to go bankrupt, SMB owners perceive their businesses as too small to be the target of a cyberattack (Felton, 2021). Furthermore, there is minimal research on cybersecurity solutions for SMBs and insufficient quantitative data on SMBs in cyber research (Chidukwani et al., 2022). Government agencies like Small Business Administration (SBA) and Cybersecurity and Infrastructure Security Agency (CISA) outline best practices and recommendations for SMBs to protect against cyberattacks. However, they do not provide sufficient guidance on mitigating web phishing attacks using open-source tools, machine learning, and publicly available data sets.

There are numerous concerns with the existing machine learning approaches, such as the large number of training features needed for feature selection and the type of data sets used in the past (Shirazi et al., 2018). Additionally, most of the published phishing data sets in the literature are not publicly available, and the feature selection methods are unclear. Much of the published literature evaluated the machine learning model's performance using accuracy (e.g., frequency of true positives and negatives). However, measuring model performance in terms of recall (false negative) and precision (false positive) may be more suitable and meaningful for SMB owners.

This chapter provides the analytical foundation for this praxis, by reviewing the SMBs' cyber threat landscape focusing on phishing and existing phishing website detection solutions, including machine learning, before identifying research gaps in phishing detection.

2.2 Cybersecurity and Small and Medium-Sized Business

Small and medium-sized businesses are critical to a country's innovation and economic growth. In the U.S., SMBs are responsible for employing about half of the available workforce and contribute more towards patenting activity per employee (SBA Office of Advocacy, 2022). Based on the definition provided by the U.S. SBA office, small and medium-sized businesses include companies that employ up to 500 employees for “small” and 1,000 employees for “medium.”

Table 1 below shows the number of SMBs by employee size, total employed, and total annual payroll in the U.S. It also shows 6,091,892 registered SMBs in the U.S. (U.S. Census Bureau, 2023).

Table 1: U.S. Firms by Enterprise Size

Enterprise Size	Number of Firms	Total Employed	Total Annual Payroll
01: Total U.S Firms	6,102,412	132,989,428	$ 7,428,553,593.00
06: <20 employees	5,431,541	21,318,425	$921,388,644.00
07: 20-24 employees	139,326	3,039,508	$128,650,298.00
08: 25-29 employees	93,975	2,521,503	$108,557,923.00
09: 30-34 employees	67,658	2,156,146	$94,148,987.00
10: 35-39 employees	50,836	1,872,965	$82,769,036.00
11: 40-49 employees	71,670	3,158,849	$144,066,088.00
12: 50-74 employees	88,667	5,336,278	$254,072,554.00
13: 75-99 employees	42,914	3,677,614	$183,605,909.00
14: 100-149 employees	41,419	5,002,843	$259,831,097.00
15: 150-199 employees	19,977	3,436,427	$181,916,146.00
16: 200-299 employees	19,052	4,617,205	$253,368,176.00
17: 300-399 employees	9,085	3,138,367	$176,542,905.00
18: 400-499 employees	5,424	2,417,778	$136,476,822.00
19: <500 employees	6,081,544	61,693,908	$2,925,394,585.00

20: 500-749 employees	6,961	4,228,530	$241,907,874.00
21: 750-999 employees	3,387	2,921,096	$169,612,267.00
Total SMBs	**6,091,892**	**68,843,534**	**$3,336,914,726.00**

Despite the importance of SMBs, there is limited cybersecurity research on SMBs, insufficient cyber defense resources for SMBs, and minimal publicly available data sets to allow SMBs to harness the value of machine learning and cybersecurity analytics.

2.2.1 Limited Cybersecurity Research on SMBs

There is limited cybersecurity research focused on SMBs. According to a survey paper on the cybersecurity of small-to-medium businesses, SMB cybersecurity research is primarily qualitative and focuses on NIST's Cybersecurity Framework (CSF) "identify" and "protect" functions (Chidukwani et al., 2022). Additionally, cybersecurity research prioritizes information security polices and operational security over incident detection and response (Chidukwani et al., 2022).

Figure 1 below shows that there is more published research on NIST CSF "identify" and "protect" functions and less on the "detect, respond, and recover" functions (Chidukwani et al., 2022).

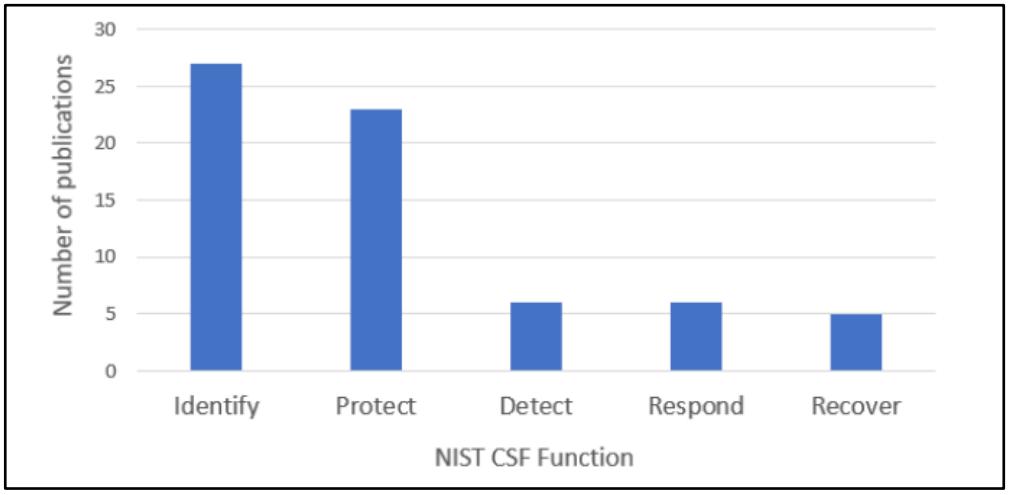

Figure 1: NIST Cybersecurity Framework Functions (Chidukwani et al., 2022)

In general, "identify" functions spans asset management, risk management, and governance and compliance, while "protect" is related to identity management and information protection processes and procedures (NIST CSF). This research contributes to the cybersecurity research on SMBs by using machine learning and historical phishing data sets to better "detect" and "respond" functions of NIST CSF.

2.2.2 Insufficient Cyber Defense Resources for SMBs

Compared to large organizations, SMBs are not well resourced to pay for information professionals, cannot meet regulatory compliances, and lack technologies to defend against cyberattacks (Chidukwani et al., 2022). While large organizations have other challenges like a larger attack surface with more employees and connected devices, a large organization has a professional information technology staff, a dedicated budget for cybersecurity (Heidt et al., 2019), and an appropriate education level (Zec, 2015). However, many of these SMB cybersecurity challenges can be addressed with open-source tools, historical cyber threat data, and machine learning, which will be discussed in-depth in **Section 2.6** - Phishing Detection Techniques.

2.2.3 Lack of Publicly Available Cybersecurity Data

A significant challenge with phishing detection is the lack of relevant and timely publicly available phishing data sets. Much of the published literature uses only a single data source which can be biased and are from publicly available databases such as Kaggle, Alexa, Mendeley, OpenPhish, and PhishTank.

Phishing website attacks do not discriminate by organization size. It is well known that many organizations do not want to share their cybersecurity incident data or any data to advance the security research community. Since this research's data sets are labeled, this research applies supervised learning to classify whether a website is phishing or legitimate.

2.3 SMB Cyber Threat Landscape

SMBs are often the easier targets for cybercriminals due to antiquated infrastructure and lax protection plans against cyberattacks. A 2022 report from Verizon shows that the attack vectors for small and large enterprises to be similar. While SMBs may not be the intended target for advanced persistent threat (APT) groups, SMBs are a prime target for social engineering (e.g., phishing), hacking (e.g., stolen credentials), and malware (e.g., ransomware) (Verizon, 2022a). Verizon's 2022 Data Breach Investigations Report reported that 82% of all security incidents involve a human element, which shows humans' susceptibility to social engineering attacks. Of note, many of the social attacks began with phishing.

According to the Small Business Administration (SBA), **Table 2** outlines the most common SMB cyber threats:

Table 2: Common SMB Cyber Threats (SBA, 2023)

Cyberattack Types	Description (Adapted from SBA, 2023)
Phishing	Uses email or a malicious website to install malware on computers to gather sensitive data
Malware	An umbrella term for software created to harm a computer, server, or computer network. Ransomware and viruses are examples of malware.
Viruses	Malicious programs can spread like a disease to other connected devices.
Ransomware	A type of malware that impacts and limits access to a computer until a ransom is paid. Ransomware generally spreads via phishing emails and takes advantage of unpatched software flaws.

Figure 2 depicts phishing as the most prominent type of attack and the exponential growth of phishing attacks since 2018 (FBI, 2023).

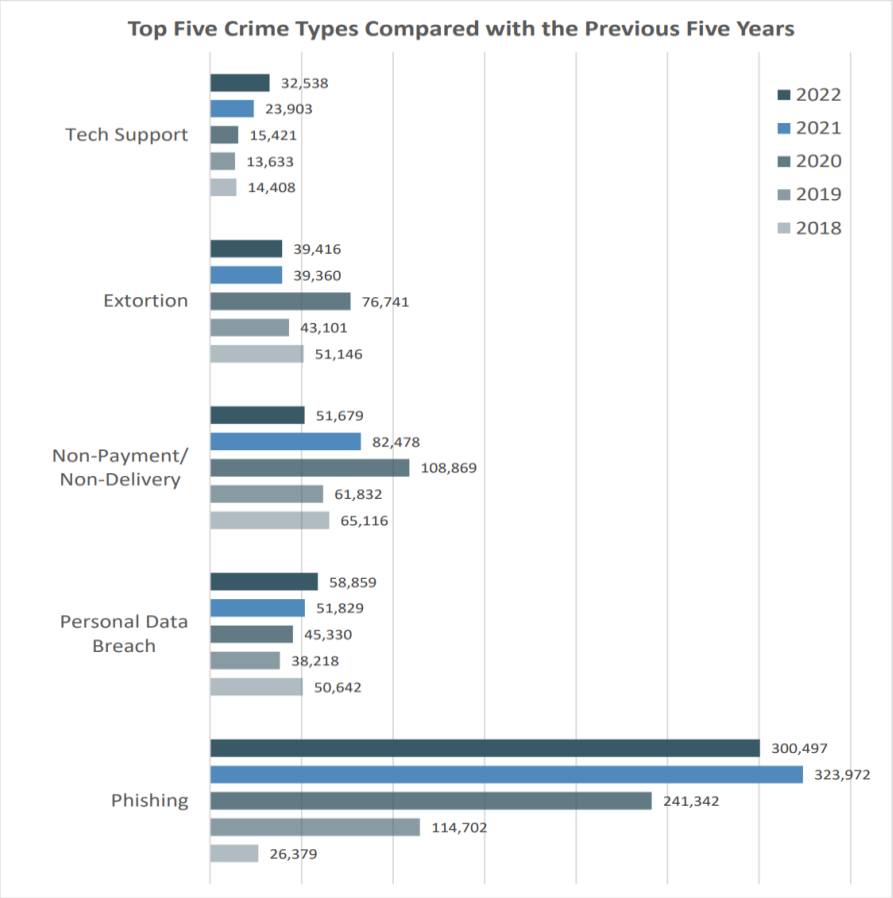

Figure 2: Top Five Internet Crimes (FBI, 2023) 2022

2.4 Cybersecurity Guidance and Resources for Small Businesses

The Small Business Administration (SBA), Federal Trade Commission (FTC), CISA, and NIST provide a comprehensive list of free cybersecurity tools and resources. The SBA's "strengthen your cybersecurity" webpage provides an overview of cybersecurity best practices, common threats, and links to resources to improve cybersecurity. The SBA outlines a list of free measures with specific resources. However, these free cybersecurity tools and resources lack a predictive modeling component and phishing detection capability.

Table 3 outlines the free government-provided resources and tools.

Table 3: Cybersecurity Planning and Assessment Tools

Cybersecurity Planning and Assessment Tools	
Create a Cybersecurity Plan	https://www.fcc.gov/cyberplanner
Conduct a Cyber Resilience Review	https://www.cisa.gov/uscert/resources/assessments
Conduct a Vulnerability Scan	https://www.cisa.gov/cyber-hygiene-services
Manage information communication technology (ICT) supply chain risk	https://www.cisa.gov/ict-supply-chain-library
Other free cybersecurity services and tools	https://www.cisa.gov/free-cybersecurity-services-and-tools
Maintain DoD industry partner compliance	https://dodcio.defense.gov/CMMC/
Learn the basics of cybersecurity	https://www.ftc.gov/business-guidance/small-businesses; https://www.cisa.gov/small-business; https://www.nist.gov/itl/smallbusinesscyber https://www.sbdcnet.org/small-business-cybersecurity/cyber-attacks-threats-defenses-small-business/

SMBs have access to various resources that emphasize awareness, education, and even the distribution of free cybersecurity tools (SBA, 2023). These tools frequently function as self-evaluation tools, primarily aiding in identifying the crucial assets that must be safeguarded. The resources listed above can help SMBs improve their cyber hygiene and preparedness, but it does not directly enable SMBs to detect phishing websites.

Analyzing phishing websites can be tedious and time-consuming to evaluate every URL. With today's technologies, SMBs can leverage free, open-source tools (e.g., R programming) to automate and build a phishing classification model. One could take several basic steps to protect against personal information phishing by simply deleting email, text, and other forms of communication. However, these essential tips ignore the root of the problem, and training only incrementally improves the detection of phishing URLs.

Current federal cybersecurity guidance and resources are insufficient to protect against phishing websites. SMBs need more than a list of resources and training to deter cyberattacks and phishing. For example, security awareness and training do not automatically stop the growth of phishing websites and attacks. Cybercriminals continually adapt and can socially engineer new sophisticated attacks that exploit human weaknesses. SMBs need a cost-effective open-source tool to detect and classify if a URL is phishing or legitimate, not a list of resources.

2.5 Phishing – Brief History & Types

A phishing attack attempts to steal sensitive information primarily for financial gain. Cybercriminals lure the victim by trickery and seek to steal credential information.

The term phishing comes from a hacking group that tricked AOL users into providing their passwords in 1996 (Jakobsson & Myers, 2006). One of the earliest known citations on phishing was from a hacker posting on Usenet newsgroup alt.2600 (Jakobsson & Myers, 2006).

> "It used to be that you could make a fake account on AOL so long as you had a credit card generator. However, AOL became smart. Now they verify every card with a bank after it is typed in. Does anyone know of a way to get an account other than phishing?"
>
> mk590, "AOL for free?," alt.2600, January 28, 1996

The phishing attack method was not limited to the AOL website but evolved to other social networking and financial platforms. A phishing attack can occur using fake email, cloning legitimate websites, and other social media applications (e.g., SMS, voicemail). While the motive is primarily financial, there are other reasons why cybercriminals use phishing attacks which include notoriety and recognition. **Figure 3** depicts an email-based web phishing attack.

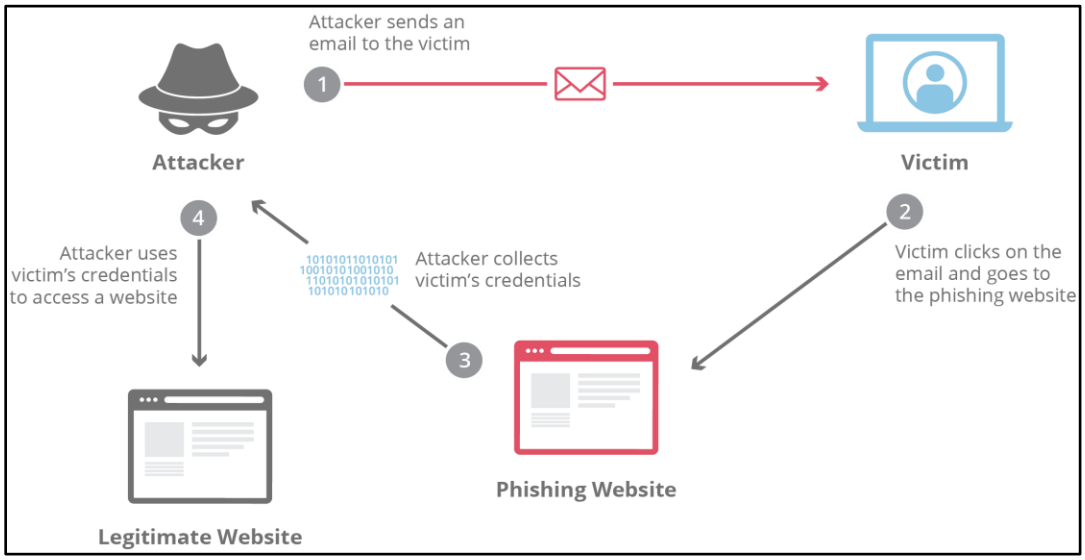

Figure 3: Phishing Attack (Cloudflare, 2023)

Phishing attacks remain a severe threat today and have become more sophisticated in the last two decades. For example, cybercriminals register new domains for phishing within 14 days (Aaron et al., 2022). The number of phishing attacks continues to grow at an alarming rate which indicates that there is still a critical need for detecting phishing websites.

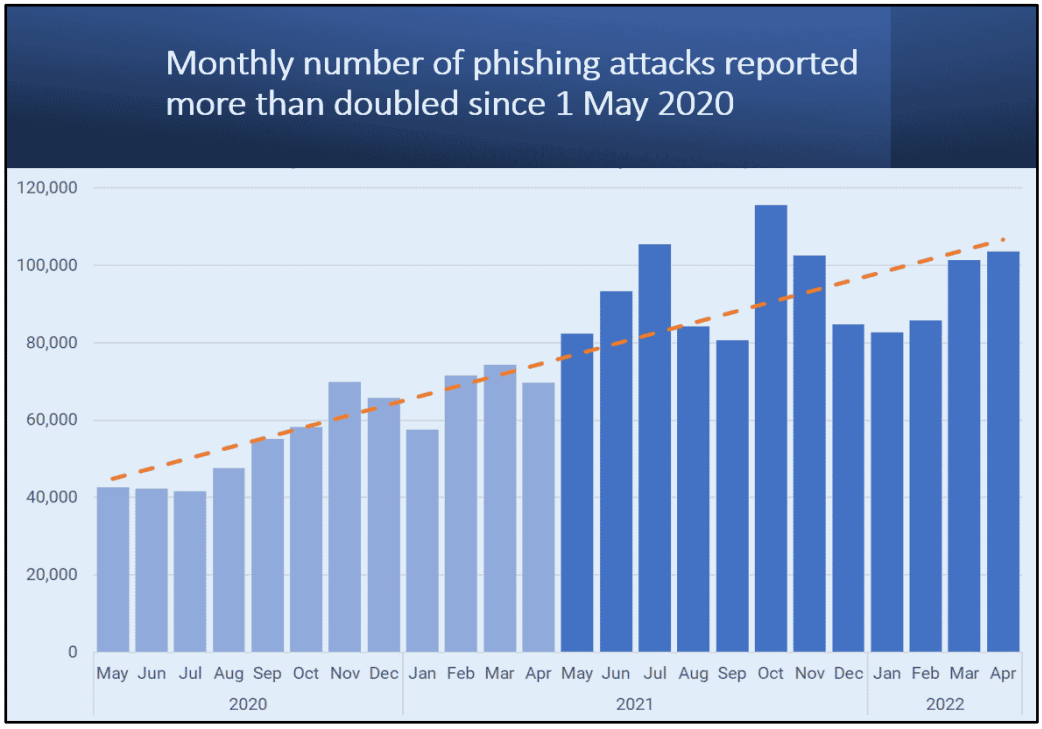

Figure 4: Growth of Phishing Attacks (APWG, 2023)

Figure 4 shows a 61% increase in phishing attacks since May 2020, and the number of phishing attacks has not decreased (APWG, 2022).

There are generally five types of phishing attacks. The primary goal of the attack is to gain information and access to sensitive information for the users. **Figure 5** shows the common types of phishing attacks (Martin, 2022).

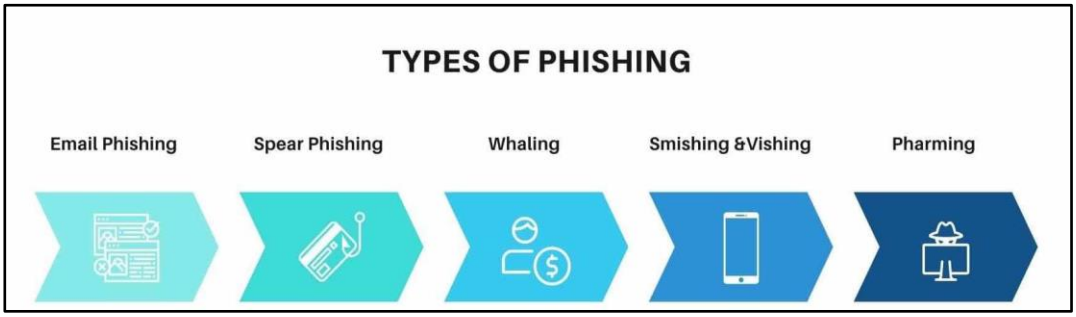

Figure 5: Common Types of Phishing (Martin, 2022)

2.5.1 Email phishing

One of the most common phishing methods is via email. Cybercriminals impersonate fake domains and send generic requests with an embedded URL link. Once the user clicks the link, the malicious website steals credentials or installs malicious code.

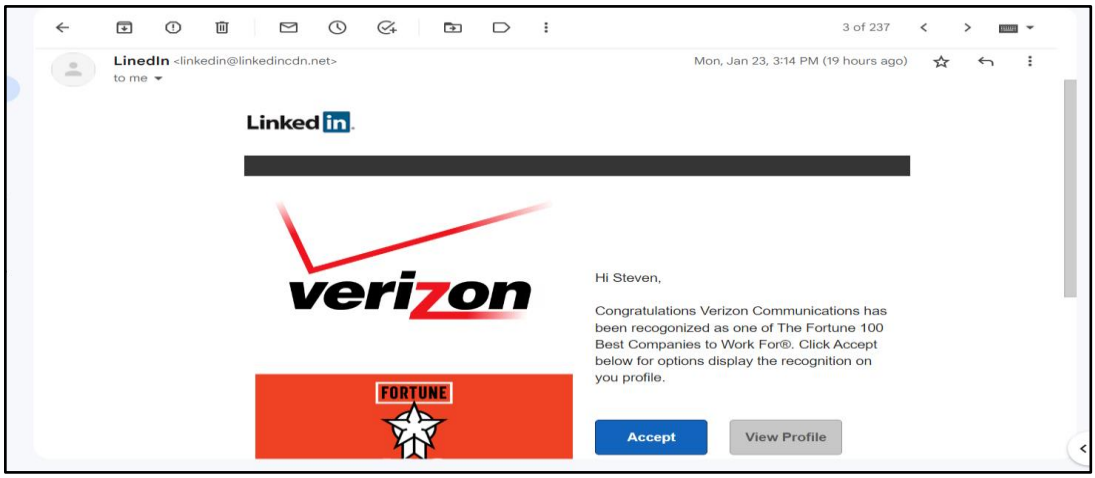

Figure 6: Real-world Example: Email Phishing Attempt

Figure 6 shows a recent email phishing attempt sent to a work email address. It attempted to mimic an official LinkedIn email. One can see the fake "LinedIn" and incorrect email address ending with "cdn.net"; however, with generative AI (e.g., ChatGPT), cyber criminals are easily able to generate a more sophisticated phishing campaign without typos and grammar mistakes.

2.5.2 Website phishing

Phishing websites pretend to be legitimate and attempt to steal users' credentials to a banking account, credit card, and other passwords. Phishing and legitimate websites may look the same, but the key differences include URLs, website content, and hosting domain. This problem persists because end-users click on URLs that they believe are trustworthy and websites are easy to create.

According to Google, the number of registered phishing sites in 2021 increased by 27% since 2020. **Figure 7** shows the continued growth of phishing sites since 2020.

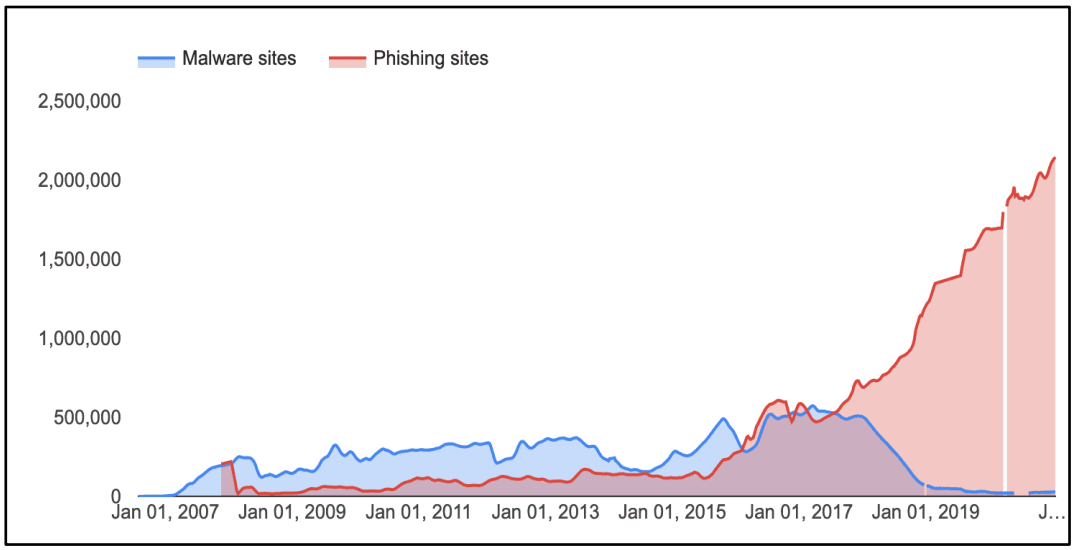

Figure 7: Google Safe Browsing: Malware and Phishing Sites (Rosenthal, 2022)

A more in-depth discussion and review of the literature on phishing website detection is discussed in **Section 2.6 – Phishing Detection Techniques**.

2.5.3 Vishing ("Voice Phishing")

Hackers attempt to steal victims' data by posing as a legitimate authority or agency on the phone. A typical example is a fraud call posing as IRS and urgently requesting social security numbers and other personal information.

2.5.4 Smishing (SMS Text + Phishing)

Smishing is another form of phishing using text messages to get someone to act. Often, these texts are junk messages and include a link that, when clicked, can install malware on the end-user's phone. Millions of junk-like smishing messages attempt to direct people to malicious phishing websites daily. For example, the number of spams, abuse, and smishing reports related to the Super Bowl increased by over 860% during the 2023 playoffs.

Figure 8 shows a screenshot of a smishing text.

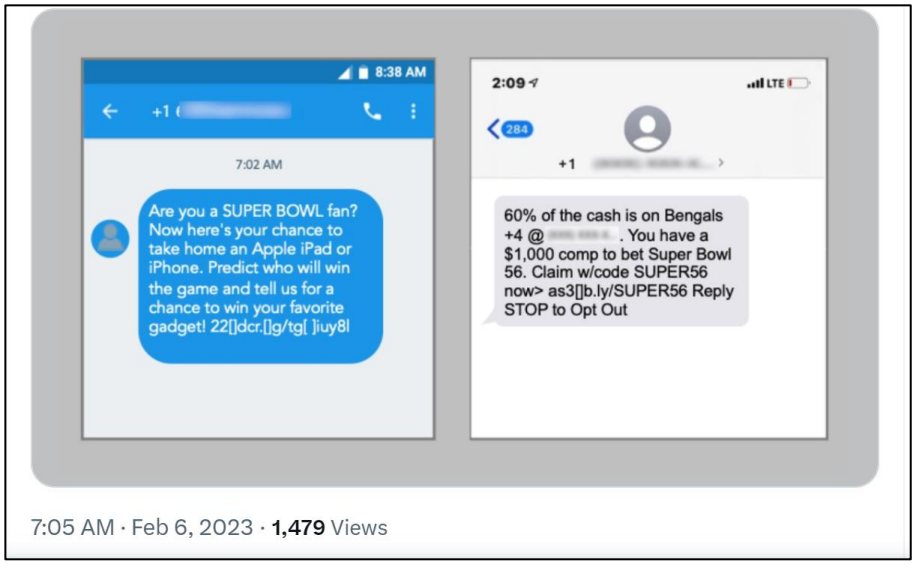

Figure 8: Smishing Example (Proofpoint, 2023)

2.5.5 Pharming

Cybercriminals hack a domain name server (DNS) and redirect the user to a clone website. In a more sophisticated URL attack, the user gets redirected to a fraud or cloned website that may look legitimate.

2.6 Phishing Website Detection Techniques

Many recent studies focus on education and training (Kara et al., 2022), traditional machine learning (Shahrivari et al., 2020), and deep learning (Wei & Sekiya,

2022), yet there is still minimal interpretability and feature selection remain arduous. Due to the ever-changing nature of phishing attacks, the ability to proactively detect phishing remains a challenge today. There are several methods for detecting phishing attacks based on user awareness and training (Sumner et al., 2022), heuristics (Nguyen et al., 2014), list-based (Vijayalakshmi et al., 2020), and machine learning (Aljabri & Mirza, 2022; Joshi et al., 2019; Kara et al., 2022; Sarker, 2021; Shahrivari et al., 2020). However, the literature proves no singular solution to address webpage phishing attacks. Each phishing detection technique has advantages and limitations, but a machine learning approach is a cost-effective solution for phishing detection with high accuracy (Safi & Singh, 2023).

The following section describes the various detection techniques for phishing sites, as seen in **Figure 9.**

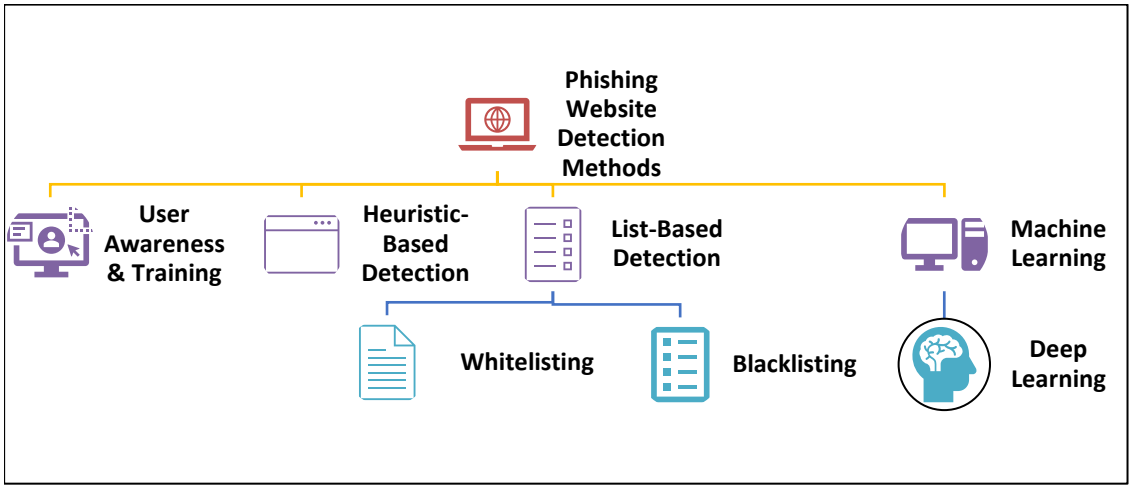

Figure 9: Phishing Website Detection Methods

2.6.1 User Awareness and Training

Raising user awareness is one of the most challenging aspects of studying the effects of phishing websites (Kara et al., 2022). This is challenging because the characteristics of phishing and legitimate webpages may look very similar. User error is

the primary cause or contributes to most security flaws (Alavi et al., 2016; Whitten & Tygar, 1999). While there is ongoing research on integrating "human-as-solution" and the need to shift cybersecurity as a problem to "Cybersecurity, Currently." Studies proposed that a shift in mindset could improve cybersecurity across the social-technical system (Zimmermann & Renaud, 2019).

Even with modern antivirus software, spam filters, and integrated "safe" browsers, phishing websites continues to threaten users and SMBs. Some companies are actively testing end-users to help bring awareness to phishing attacks (e.g., KnowBe4). A study examined the factors impacting the effectiveness of KnowBe4 anti-phishing training using survey data for pre- and post-training. It showed that anti-phishing training improved the detection of phishing by 8% but, at best, achieved only 68% accuracy (Sumner et al., 2022). Some researchers take a psychology-based approach to identify users' personalities and susceptibility (Sloan, 2020). However, this technique requires the user to build phishing knowledge and the training time to detect phishing attacks may not be practical.

While education and training may be helpful to some users, this training requires time and funding. A study compared the costs for three anti-phishing software-as-a-service solutions provided by KnowBe4, PhishMe, and InfoSec (Sloan, 2020), but this may not be within SMBs' budget for security. These human-based phishing mitigation strategies lack urgency and do not directly detect phishing with high accuracy. Since cybercriminals continue developing sophisticated alternative phishing techniques, end-user training is not up to date to protect the user. Education and training cannot eliminate

human errors like clicking on a phishing URL. It is a marginal and resource-intensive solution to a perpetual problem.

2.6.2 Heuristic-Based Detection Techniques

Heuristic-based detection techniques are an extension of listing techniques and include webpage content, visual similarities, and URL-based features (Kalaharsha & Mehtre, 2021). This method extracts website features by analyzing the structure of phishing websites to classify websites. These new websites are classified as phishing sites if they match by visually comparing phishing websites. Research shows that heuristic-based is considerably more expensive than updating a list (Silva et al., 2020). However, heuristic-based detection can be more accurate results than list-based techniques. A study used a novel model using six heuristic features from URL and website ranking. This study evaluated 11,660 phishing websites and tested on ten testing data sets with 1,000 legitimate websites each. This study detected phishing web pages at 97.16% (Nguyen et al., 2014).

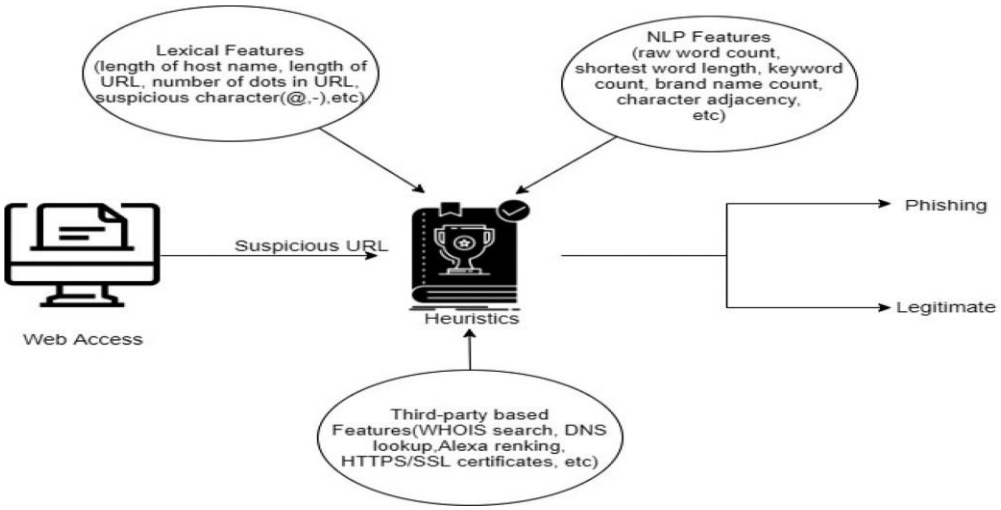

Figure 10: Heuristic-based Web Phishing Detection (Vijayalakshmi et al., 2020)

Jain and Gupta (2017) analyzed visual similarity-based web phishing detection and found that most approaches had limitations against accuracy, new phishing websites, and detecting embedded objects.

2.6.3 List-Based Detection Techniques

The list-based approach is known as the traditional or database-oriented method. These managed lists either allow the listing of websites (e.g., whitelisting) or block the listing (e.g., blacklisting); however, require active management and domain knowledge.

Whitelisting. Whitelists are safe websites and block access to websites not on the "safe" list. The legitimate URLs are kept in the database and used to verify new URLs. In this method, whenever a new URL is entered, it is first checked to see if it already exists in the database. If not, the entire URL's information, including its domain name, age, secure sockets layer (SSL) certificate, and any linked websites, is verified before being saved in the database. The disadvantage of the whitelisting technique is knowing whether the websites registered as "legitimate" are legitimate. The drawback of listing techniques is their high space requirement (Rao & Pais, 2019), depicted in **Figure 11.**

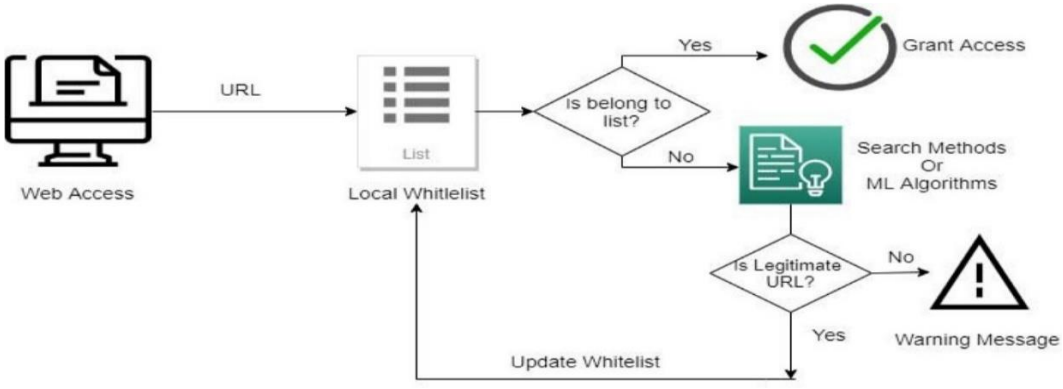

Figure 11: Whitelist-based Web Phishing Detection (Vijayalakshmi et al., 2020)

Blacklisting. In this method, the URLs thought to be phishing sites are stored in a database. As new URLs are entered, the database is checked against the URLs entered, and if a match is found, the browser blocks the URL and stores it in the database for future use (Vijayalakshmi et al., 2020). A key advantage of the blacklist is the low false-positive rates. However, a significant drawback is the time required to register the URL as phishing. This may take hours or months to be registered on a blacklist. Additionally, this technique cannot detect newly created phishing or zero-day phishing. Some of the well-known blacklisting websites include URLHaus, PhishTank, and OpenPhish.

There are several problems with list-based methods, which include not being able to detect new phishing URLs called zero-phishing (Vijayalakshmi et al., 2020) and time to maintain the list. However, this method provides high precision, resulting in a low false positive rate.

The following section describes how machine learning uses various features of the URL and webpage to predict phishing websites.

2.6.4 Machine Learning (ML)

[Machine learning is the] field of study that gives computers the ability to learn without being explicitly programmed. — Arthur Samuel, 1959

Machine learning can identify phishing websites using host-based and lexical features extracted from a URL (Kara et al., 2022). Lexical features are based on the URL's text, and examples include domain length, URL length, the number of special characters, and many statistical characteristics (Aljabri & Mirza, 2022). There are several features that one can extract from a URL. However, there is a lack of research and consensus on the optimal number of lexical features for phishing website detection. Generally, URLs will include different components and levels such as protocol,

subdomain, domain, top-level domain, directory, parameter, and query string (Berners-Lee et al., 2005). **Figure 12**Figure 1 depicts an example of a URL and its components.

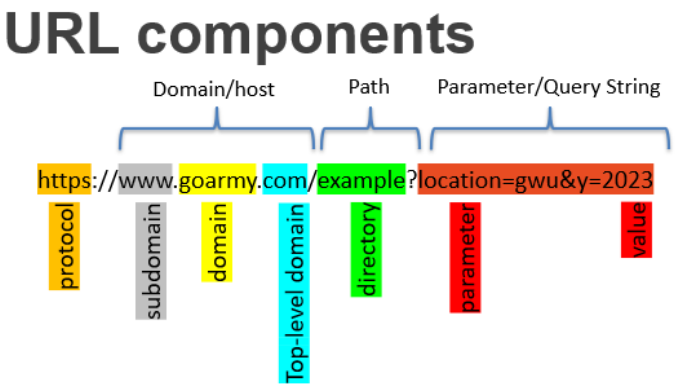

Figure 12: URL Components

Compared to other URL features (e.g., host-based, content-based), lexical feature extraction is relatively easier to apply than content-based and visual-similarity. Host-based features indicate who manages it, website location information, and other website characteristics (Kara et al., 2022).

Phishing website detection's central premise is that the structure and components of phishing URLs are distinguishable from good URLs. Joshi et al. (2019) found that malicious URLs have a second level of the subdomain, special characters in the URL that look like the whitelisted domain, and a higher average domain length. Machine learning has been proven to be an effective tool for classifying phishing websites (Shahrivari et al., 2020) since the machine learning model is a powerful tool for classification problems.

Machine learning can quickly identify patterns and build models from multiple data sets derived using various website URL features, then test them using various machine learning classifiers, including the support vector machine, naïve bayes, decision tree, Random Forest classifier, and logistic regression. A comprehensive review of

machine learning algorithms and real-world applications is discussed (Sarker, 2021). However, most published research lacks clarity on the feature selection methods, and the data sets are not publicly available. Additionally, machine learning algorithms are limited to URL features or webpages as features (Vijayalakshmi et al., 2020).

One of the challenges with machine learning-based solutions is feature identification and selection. Numerous time-consuming tasks exist to find the most important features to train and build an ML model. Lastly, there are few publicly available training data sets for phishing websites (Tang & Mahmoud, 2021). Compared to list-based methods, machine learning-based methods generally provide a higher accuracy rate and a higher false positive rate depending on the classifier and data set.

This section covers the related work and literature review on machine learning applications for detecting phishing websites. **Figure 13** depicts a high-level machine learning-based web phishing detection process**.**

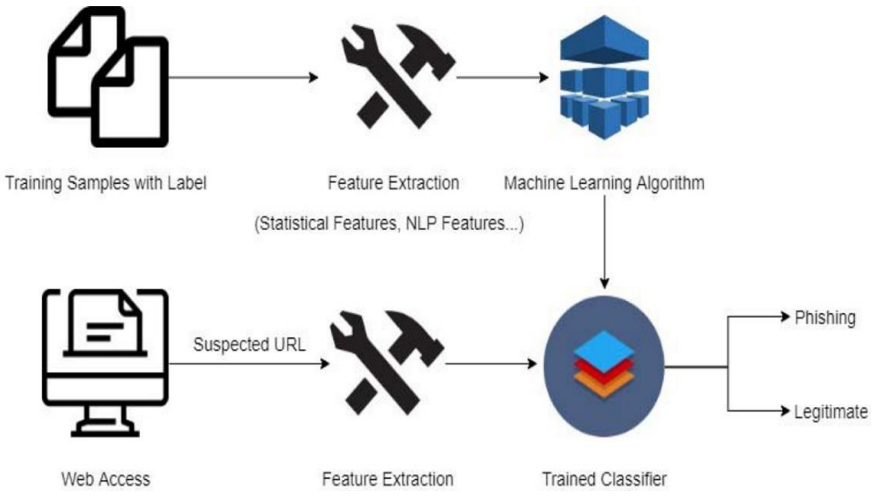

Figure 13: ML-based Web Phishing Detection Process (Vijayalakshmi et al., 2020)

Several machine learning-based phishing website approaches have been published in the literature. The main difference between the approaches are the selected features, the data set, and the machine learning algorithms. The most relevant and recent literature on ml-based phishing website detection is summarized and critiqued:

Chiew et al. (2019) used various machine learning techniques to detect phishing websites using Machine Learning Repository (UCI) data. The main contribution was the novel feature selection framework called hybrid ensemble feature selection (HEFS). The study used 5,000 URLs from the PhishTank, OpenPhish, Alexa, and Common Crawl archives. The best model used a Random Forest algorithm and had an accuracy of 94.6%; however, the study did not validate its finding using a validation data set (Chiew et al., 2019). It did not show if the ML model was robust with multiple data sets.

Shahrivari et al. (2020) used twelve machine learning models (logistic regression, AdaBoost, Random Forest, KNN, neural networks, SVM, Gradient boosting, and XGBoost) for detecting phishing websites using a small data set. The authors used 30 features from 11,000 sampled websites from PhishTank, and 10% for the testing phase. While the results showed high accuracy for the ensemble classifiers (e.g., Random Forest, XGBoost), this is expected with ensemble models since it combines weak learners for classification problems. The study did not discuss the feature selection methods and why the 30 features were used to train the model.

Similarly, Wei and Sekiya (2022) used a data set of 88,647 websites with 112 features from available Mendeley 2020. They proposed a new framework for feature selection by combining existing methods and exploring deep learning-based models. This study found that the Random Forest (RF) performed better than DL-based models in

accuracy, training time costs, precision, recall, and AUC. While DL-based is considered the state-of-the-art solution, the model requires high training time, computation time, and the most complex model architecture, which can be challenging to explain. For the classification of phishing websites, the ensemble ML with RF provided an accuracy of up to 96% and on average, significantly lower training time (Wei & Sekiya, 2022).

Kara et al. (2022) proposed a new method by categorizing a website's Internet URLs and domain names using six classifier algorithms. The study's data set had 12,134 legitimate websites and 20,614 phishing websites while achieving an accuracy of up to 98.90%. However, the methodology remains contentious as the data set access is restricted and lacks transparency on the data collection methodology from the Turkish government's TR-CERT website. The validation of the model on a publicly available data set would have enhanced the study's claim.

Odeh et al. (2021) examined traditional ML techniques and outlined the challenges with conventional ML techniques such as overfitting, low accuracy, and ineffective without enough training data. This can be improved if the training data is sufficiently large enough to mitigate overfitting. Additionally, ensemble models have proven to be very effective at accurately classifying phishing.

Table 4 below provides an overview of recent and relevant papers highlighting their contribution and research gaps.

Table 4: Relevant ML-based Papers - Contributions and Gaps

Paper	Contributions	Gaps
Shirazi et al., 2018	- Domain Name-Based Features (not using third-party servers, search engines, suspicious words, and URL-specific features) - Bias in data sets: usage and URL features	-Seven features 97% accuracy -Insufficient data sources - did not validate data on new phishing data sets.
Chiew 2019	- Feature selection framework – hybrid ensemble feature selection (HEFS) - RF at 94.6% accuracy	-Limited testing data set (5,000 phishing URLs) from 5/2015-6/2017 -Outdated ML UCI data; small experimental data set
Shahrivar et al, 2020	- Ensemble classifiers (RF, Adaboost, XGBoost) best accuracy	-Vague features selection method -Did not use multiple different data sets to train, test, and validate
Chen et al, 2020	- Demonstrated the validity of ensemble feature selection method	-Data, findings, and algorithms were only applied to medical domain
Kara et al, 2022	- Categorizing Internet URLs and domain names of a website using six classifier algorithms	-Single feature selection method -Cannot replicate findings as data set is unavailable
Bhowmik 2022	- Hybrid feature selection using filter and wrapper methods. - Two phishing data sets	-XGBoost at 96.08% accuracy -No comparison of the results to a different data set
Wei & Sekiya, 2022	- Novel feature selection approach: development of anti-phishing score - Deep learning comparison	-One data set -Did not validate findings on another data set

2.6.4.1 Deep Learning (DL)

Deep learning is a subset of machine learning that can provide state-of-the-art performance for various research problems (Catal et al., 2022). Recently, deep learning has been a well-known application for non-linear data such as face recognition, natural

language process, image classification, and other unstructured data. Catal et al. (2022) found that the Deep Neural Network is the most preferred deep learning algorithm, which provided the best performance. However, deep learning involves using artificial neural network models which are complex and computationally complex.

While deep learning performs well for the classification of phishing websites, a recent systematic literature review on the application of deep learning for phishing showed that there are several challenges with deep learning for phishing detection including model efficiency, model interpretability, a model for specific cases, model design consideration, and data sets (Catal et al., 2022). These DL-based models have long training times, high computing resources, and require fine-tuning of hyperparameters. Of note, many DL-based models have a "black-box" approach where it lacks interpretability and features are not explainable.

Despite deep learning's advantage with automatic feature selection, traditional machine learning is more appropriate because it allows the researchers to control and better understand the feature selection process. About half of the literature on deep learning did not mention any framework other than Keras and TensorFlow (Catal et al., 2022).

2.7 Summary of Gaps in Phishing Detection Methods

The literature review provided a comprehensive survey of the existing phishing detection methods and highlighted the key limitations of user awareness, list-based, heuristics, deep learning, and machine learning. User awareness and training are essential, but it does not significantly improve phishing detection accuracy. Every human is prone to making mistakes, even with effective phishing training or susceptible to

mandatory training fatigue. List-based detection methods are time-consuming to maintain, and they do not detect new phishing URLs (e.g., zero-day phishing). Heuristics phishing detection improved accuracy over list-based, but it requires high space and is expensive. Machine learning continues to be the preferred and dominant method for phishing detection but is highly dependent on feature selection and the quality of the data set.

The literature shows that the machine learning-based phishing classification model used only one data set and often, did not explain the feature selection methodology. Specifically, research findings were data set specific, lacked generalization, and were not reproducible. Research lacks a standard set of features to be used as a benchmark for ML model performance. Within machine learning, deep learning methods continue to mature, but it lacks interpretability and is computationally complex.

A summary of the phishing detection methods and attributes is described in **Table 5** below.

Table 5: Summary of Phishing Detection Methods

Phishing Detection Method	Summary	Papers
User Awareness/Training	Requires time, and budget, and provides an incremental improvement in phishing detection accuracy	Kara et al., 2022; Kalaharsha & Mehtre, 2021; Whitten & Tygar, 1999, Alavi, Islam, & Mouratidis, 2016; Zimmermann & Renaud, 2019; Sumner et al., 2022; Sloan, 2020; Silic & Lowry, 2020
List-based	Time-consuming to maintain; does not detect zero-day phishing	Rao et al., 2019 Vijayalakshmi et al., 2020
Heuristics	Slightly better accuracy than list-based; challenges with visual-similarities; high	Kalaharsha & Mehtre, 2021; Silva, Feitosa, & Garcia, 2020; Nguyen et al., 2014

	space requirement and expensive	
Machine Learning	Customizable; high accuracy with ensemble classifiers; feature selection can impact the model's performance; minimal costs	Kara et al., 2021; Aljabri & Mirza, 2022; Joshi et al., 2019; Shahrivari et al., 2020;Sarker, 2021;Vijayalakshmi et al., 2020;
Deep Learning	Lack of interpretability; requires large training data; resource intensive and computationally expensive	Safi & Singh, 2023; Catal et al., 2022

2.8 Summary and Conclusion of ML-based Phishing Detection

The literature review shows that the machine learning approach for detecting phishing is challenged with a large number of features for training, quality data sets, and interpretability of features to other data sets.

First, machine learning-based phishing requires sound feature selection methods for building high-performing models. Unlike deep learning, the machine learning model is preferred since it can help SMBs better understand and explain the features. This research will address the limitations and biases of training data sets using multiple phishing and curated data sets to validate the model's performance.

Second, the research will integrate explanatory algorithms, or "explainable AI" to help explain the "why" in the model prediction. Rather than using the model for only phishing classification, this research will close the gap in interpretability by showing how each independent variable impacted the dependent variable in the prediction model.

Table 6: Literature Review Gaps and Praxis Contributions

Gaps	Praxis Contributions
Vague feature selection method	Multiple feature selection method analysis including Shapley Approximation (SHAP) to improve the interpretability of the prediction model

Lack standard set of features for benchmarking	Identify an optimal standard set of features that can be used with multiple phishing data sets
Bias in data set using only one data source	Train, Test, and Validate three publicly available and recent phishing data sets
Outdated data set	Contribute a new data set using multiple data sources for detecting the latest techniques
Improve accuracy and minimize false positive	Ensemble feature selection

In summary, the gap for ML-based phishing detection is the feature selection method while obtaining high accuracy and validation on multiple publicly available data sets as shown in **Table 6**. Also, there is a lack of consensus on the optimal number of lexical features for phishing website detection. Many of the research feature selection methods were unexplained and lacked access to the data set for replicating research. To that end, this praxis will build a phishing classification model using open-source tools, publicly available data sets, and address the identified limitation of the machine learning model with a novel feature selection method and validate on multiple publicly available data sets.

Chapter 3—Methodology

3.1 Introduction

The research methodology chapter describes the research process used for the praxis. As mentioned in previous chapters, this praxis aims to build a machine learning-based phishing classification model using multiple phishing data sets and ensemble feature selection techniques.

A high-level overview of the research process and the rest of the chapter is depicted below in **Figure 14**. The figure highlights the key tasks (e.g. "What"), the method (e.g. "Why"), and "How" each tasks enable the primary objective of this research.

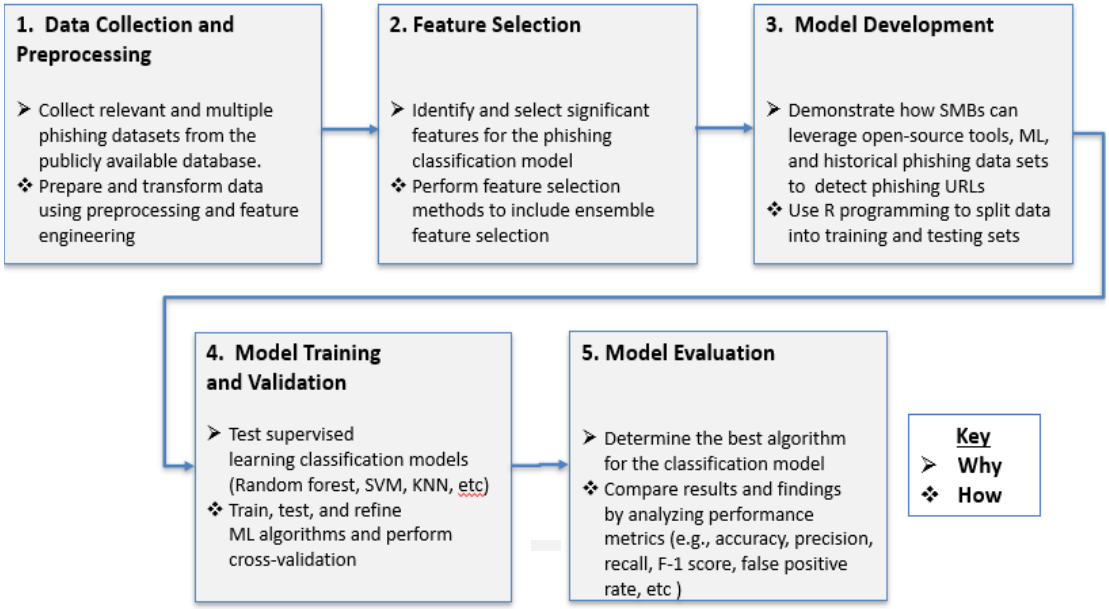

Figure 14: Research Methodology

The following section summarizes data collection, preprocessing, feature selection methods, and model development.

3.2 Data Collection and Preprocessing

Out of all the phishing data sets publicly available and developed after 2020, four phishing data sets that met the following criteria: provenance, quality of metadata, size of the data set, and availability (La Trobe University, 2020) were used in this praxis.

Figure 15 depicts the latest published phishing website data sets available: 1) Phishing Website Detection (Vrbančič et al., 2020) available on Mendeley; 2) Phishing Site URLs from Kaggle (Tiwari, 2022); 3) Phishing Website Detection Data set (Pham, 2022); 4) Validation research data set built from OpenPhish, MajesticMillion, and other publicly available databases for this praxis.

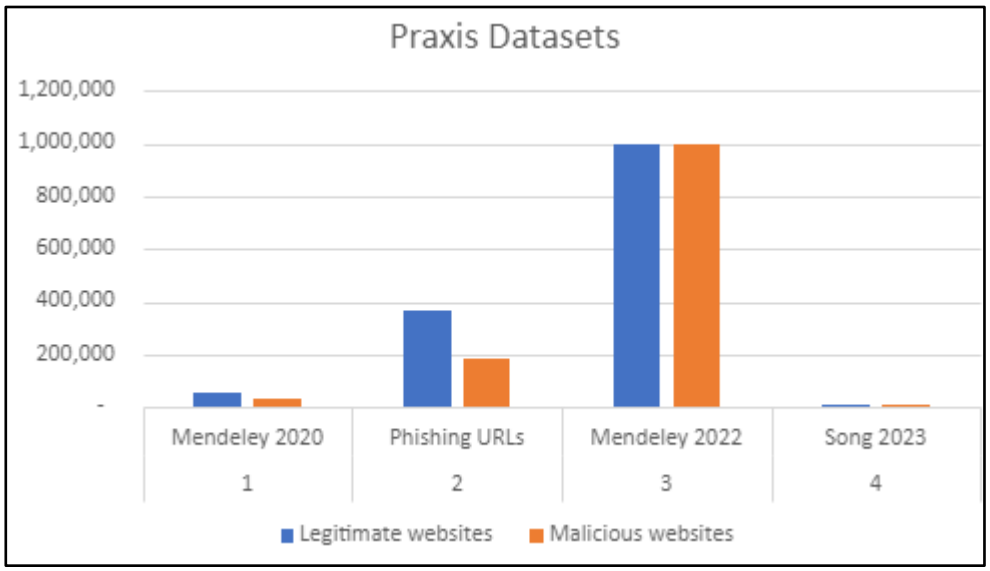

Figure 15: Phishing Data Sets

To address the limitations of the previous research, multiple phishing data sets were used to demonstrate and validate the phishing classification model's robustness and identify the optimal subset of features for benchmarking phishing classification. **Table 7** provides the purpose and distribution of the classes in each phishing data set.

Table 7: Phishing Data Sets

Short Name: (Source)	**#1 Mendeley 2020** (Vrbančič et al., 2020)	**#2 Phishing URLs** (Tiwari, 2022)	**#3 Mendeley 2022** (Pham, 2022)	**#4 Song 2023** My data set
Purpose	Learn key features to replicate from the most-cited phishing data set; Baseline Data set	Build a URL parser to select and extract features from Mendeley 2020	Test and validate the phishing classification model	Validate optimal set of features
Number of Features	112	2	2	2
Phishing	30,000	142,015	1,000,000	10,0000
Not Phishing	50,000	365,180	1,000,000	10,0000
Total	88,647	507,195	2,000,000	20,0000

3.2.1 Phishing Website Data Set on Mendeley 2020 (Vrbančič et al., 2020)

Vrbančič et al. (2020) data set was selected because it was well-known and most cited in the literature within the last two years. The data set consists of 112 URL features. The first variation is a small and balanced data set with 58,645 websites labeled as legitimate or phishing. The second variation contains 88,647 websites. Each URL consists of multiple components such as protocol, domain, directory, and parameter. **Table 8** depicts the phishing data set 112 features organized into eight groups and serves as the baseline for comparison.

Table 8: Mendeley 2020 Features Grouped (Wei & Sekiya, 2022)

Feature #	**Feature Group**	**Description**
1 ~ 17	URL	number of ". - _ / ? = @ & ! ~ , + * # " $ % " in URL
18 ~ 34	Domain	number of ". - _ / ? = @ & ! ~ , + * # " $ % " in Domain
35 ~ 51	Directory	number of ". - _ / ? = @ & ! ~ , + * # " $ % " in Directory
52 ~ 68	File	number of ". - _ / ? = @ & ! ~ , + * # " $ % " in File
69 ~ 85	Parameters	number of ". - _ / ? = @ & ! ~ , + * # " $ % " in Parameter

86 ~ 96	Other	number of vowels, number of parameters, time_response, asn_ip, time_domain_activation, time_domain_expiration, number of resolved Ips, number of resolved NS, number of MX servers, Time-To-Live, number of redirects
97 ~ 102	URL Components	Top-level domain character length, number of characters in the whole URL, number of domain characters, number of directory characters, number of file characters, number of parameter characters
103 ~ 112	Index	email present, URL domain in IP areas format, "server" or "client" in the domain, TLD present in parameters, domain has SPF, URL has valid TLD/SSL certificate, URL indexed on Google, domain indexed on Google, URL shortened

3.2.2 Phishing Site URLs (Tiwari, 2022)

The data set has 507,195 unique observations with only two columns (e.g., URL and label). The first column is the URL, and the second is labeled “good” or “bad.” Of note, the data set is imbalanced, with 72% labeled as “good” and 28% as “bad.” During preprocessing, the data set was resampled to create a more balanced data set resulting in 392,024 bad and 392,024 good. Furthermore, three URL string features for the predictive classification model were removed because they did not provide any numerical value (i.e., URL, Domain Name, and Directory Name)

Table 9: Phishing Site URLs Features (Tiwari, 2022)

#	Feature Name	Description
1	URL	URL (Character String)
2	Label	Good or Bad (Categorical)

3.2.3 Phishing Website Detection Data Set (Pham, 2022)

Like the Mendeley 2020 data set, this was a more recent publicly available and published data set on Mendeley in 2022 (Pham, 2022). This data set contains 1,000,000 phishing URLs and 1,000,000 legitimate URLs. The phishing URLs were extracted from phishtank.org while the legitimate URLs were from commoncrawl.org.

Table 10: Phishing Website Detection Data set (Pham, 2022)

#	Feature Name	Description
1	URL	URL (Character String)
2	Label	Good or Bad (Categorical)

3.2.4 Validation Phishing Data set

This data set includes 10,000 legitimate URLs from the majestic millions and 10,000 phishing URLs from PhishTank and OpenPhish.

The primary reasons for creating a validation data set were to validate the developed ML model and see if the model results could be replicated on a new data set. As discussed in the literature review, many of the prior studies built ML models for phishing detection that was not reproducible. This is problematic because there is bias in the data set usage and URL-based features (Shirazi et al., 2018).

In the validation phishing data set, legitimate websites were obtained from a variety of sources to mitigate bias in the existing phishing detection research. Kalaharsha and Mehtre (2021) showed that most research used Alexa.com to create a list of legitimate websites. While this may be a suitable approach, many of these academic findings cannot be generalized beyond the data set used in their research. Similarly, much of the phishing detection research exclusively used a list of websites from PhishTank. This research addresses these shortcomings of previous research by using a current data set that is more representative of the actual website population for phishing and legitimate URLs. For that purpose, 10,000 phishing websites were collected from the first week of February 2023.

3.2.5 Data Preprocessing

Data preprocessing was critical for improving the data quality before building the model by removing any inconsistencies, noise, missing values, overfitting, and other data

transformations essential for building a machine learning classification model. For example, transformation helped mitigate the impacts of outliers on the predictive classification model.

All four phishing data sets were preprocessed: data cleaning, transformation, and normalization. Preprocessed means that the data handled missing values, NAs, and categorical values. With the large data sets, removing observations with missing values was insignificant to the model outcome. The process of transforming raw data into relevant and valuable features in machine learning is called feature engineering. In other words, this research used feature engineering and ensemble feature selection techniques to build the best phishing classification model.

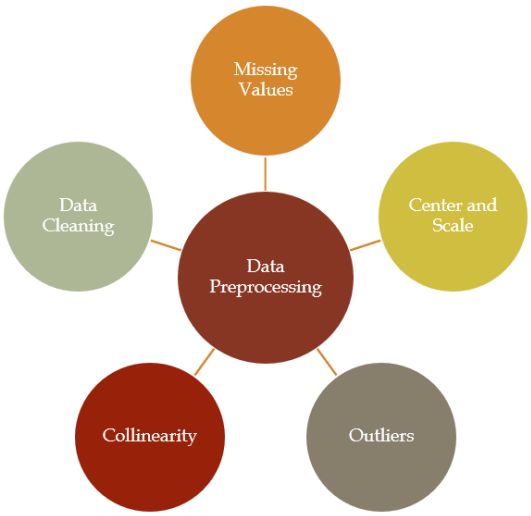

Figure 16: Data Pre-Processing

A crucial step for preprocessing was centering and scaling because features have different units. Centering and scaling are transformation techniques that do not change the intrinsic meaning of the variables. It simply centers and scales a variable to mean 0 and standard deviation 1. In centering, the independent variable is centered by removing

the average predictor variable from all the values (Kuhn, 2008). Scaling sets a common standard deviation of one (Kuhn, 2008).

All preprocessing was implemented using the caret (short for Classification and Regression Training) package in R programming 4.6, an open-source tool. The caret package provides a holistic framework to build machine learning models including tools for splitting data, different ML models, and tools to select the best models (Kuhn, 2008).

Table 11: Phishing Data Set Preprocessing: Missing Values and NAs

Data sets	Missing Values	NAs
1 Phishing Website (Vrbančič et al., 2020	2	1
2 Phishing Site URLs (Tiwari, 2022)	2	3
3 Phishing Website Detection (Pham, 2022)	4	5
4 Validation Data Set	0	0

The praxis utilized R programming on a desktop running Windows 10 with Intel Core vPRO i5-11500 2.7 GHz processor with 16 GB RAM. The caret package assisted with the following options for preprocessing range, centering, and scaling. To prepare the data, the following steps were completed:

- Reduce redundant information
- Remove missing values
- Preprocess features using ***range*** (e.g., normalize values between 0 and 1), ***center*** (subtract by mean), and ***scale*** (divide by standard deviation) [
- Remove information about the target variable from the training data
- Split the data set into 80% training and 20% testing

3.3 Feature Selection Methods

Due to the ongoing changes in website phishing attack tactics and methodology, the cybersecurity community lacks a universally accepted standard set of features for benchmarking and building machine learning models. Additionally, larger data sets create complex high-dimensional data sets, making feature selection even more challenging. Therefore, multiple feature selection methods were systematically analyzed to identify an optimal set of features that can be used universally across any phishing data set.

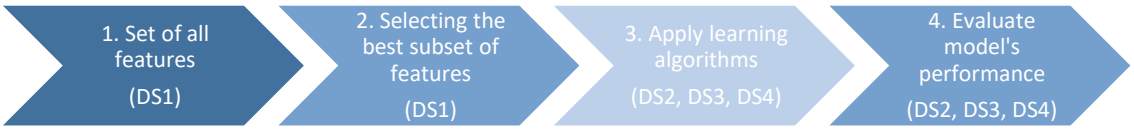

Figure 17: Feature Selection Approach

Feature selection helps improve the model's accuracy, reduces complexity, and enables the model to train faster. To identify an optimal subset of features, this research compared the performance of existing feature selection methods: filter, embedded, and ensemble on a well-known phishing data set (e.g., Mendeley 2020), then applied these learned features to other data sets for evaluation and validation. The ensemble feature selection method aided in selecting the best feature subset for phishing classification.

In machine learning, selecting meaningful and relevant features is the best way to mitigate overfitting and enhance the accuracy of the machine learning-based phishing classification model. For example, if the model is overly complex on the given training data, it might not be able to generalize on the new data or testing set. A common technique is dimension reduction. It is critical to improve learning performance and accuracy, lower computation costs, and enhance model interpretability. Selecting features relevant to the model outcome while minimizing redundancy and overfitting necessitates

a novel ensemble feature selection technique to quickly identify the optimal feature subsets that can be universally applied to all phishing data sets.

Chiew et al. (2019) find that the Hybrid Ensemble Feature Section (HEFS) with RF classifiers provide 94.6% of phishing and legitimate using 20.8% of the original features. However, HEFs were only tested using the UCI data set. As such, the HEFs worked well with the UCI data set but may not be generalized for another phishing data set. Most phishing detection research centers on improving classifiers, not feature selection techniques. Chiew et al. (2019) solely focused on filter measures-based feature selection to identify the best cut-off rank.

Identifying and extracting important key features is key to building an accurate phishing classification model. We will need to use the characters and components of the URL string to predict phishing URLs. Specifically, the features of URL will be extracted which can be used as predictor variables in the model. For example, the number of special characters, URL length, domain length, etc. A common feature selection method is removing features with low variance, as a feature with low variance may not be helpful for a predictive classification model. This research adopts an ensemble feature selection method to build the most accurate phishing classification model.

This section discusses feature selection techniques for determining the best subsets of features that optimize the model's performance. It provides an overview of the existing feature selection methods and highlights the advantages and disadvantages of each approach. Feature selection can be grouped into the following categories:

Filter Methods: features are selected using statistical measures such as correlation or chi-squared test (Gupta, 2020). This method helps understand each feature with minimal limitations from any specific classifier.

Wrapper Methods: features are selected usings combination of feature subsets using a predictive model (Gupta, 2020) such as recursive feature elimination and Shapley Additive exPlanations (SHAP). The wrapper technique iteratively produces and evaluates a feature subset via classification during each run (i.e., training and testing). The number of iterations in the wrapper approach will increase rapidly while assessing a more extensive set of characteristics and become computationally unfeasible for real-time applications (Chiew et al., 2019).

Embedded Methods: features are selected through learning their importance during model training (Gupta, 2020). This method uses the models to select the best or most important features such as Random Forest, KNN, and XGBoost.

3.3.1 Filter Method: Dimension Reduction using Correlation Analysis

In machine learning, correlation is important because it helps the relationship of the features in a data set, helps with dimension reduction, and can help simplify the model. Therefore, the model only retains features necessary to represent a different aspect of the data. Of note, correlation does not explain causation.

This filter-method identifies which features impact the model's outcome the most using correlation analysis (e.g., statistics). This feature selection technique helps reduce many features by removing features that are not important or statistically significant for predicting the model's outcome. A simple approach is to remove features that are not important to the model's outcome. Correlation analysis help with dimension reduction,

understanding the relationship between features, and improving the accuracy of a machine learning model by keeping only necessary features. For example, if two variables are highly correlated, a change in one variable will change the other. Thus, highly correlated predictors provide the same underlying information so it can be removed.

The figure below depicts the feature correlation matrix. The sensitive predictors were eliminated by applying a threshold of 0.75. In short, they eliminate the pairwise correlation predictors with less than 0.74. As shown, this heat map help analyze the correlations between the features. Dark blue circles denote high correlation. For example, there is a high correlation between features and the number of slashes in the file and equal in the file. Also, a high correlation between the number of slashes and the number of exclamations in the directory.

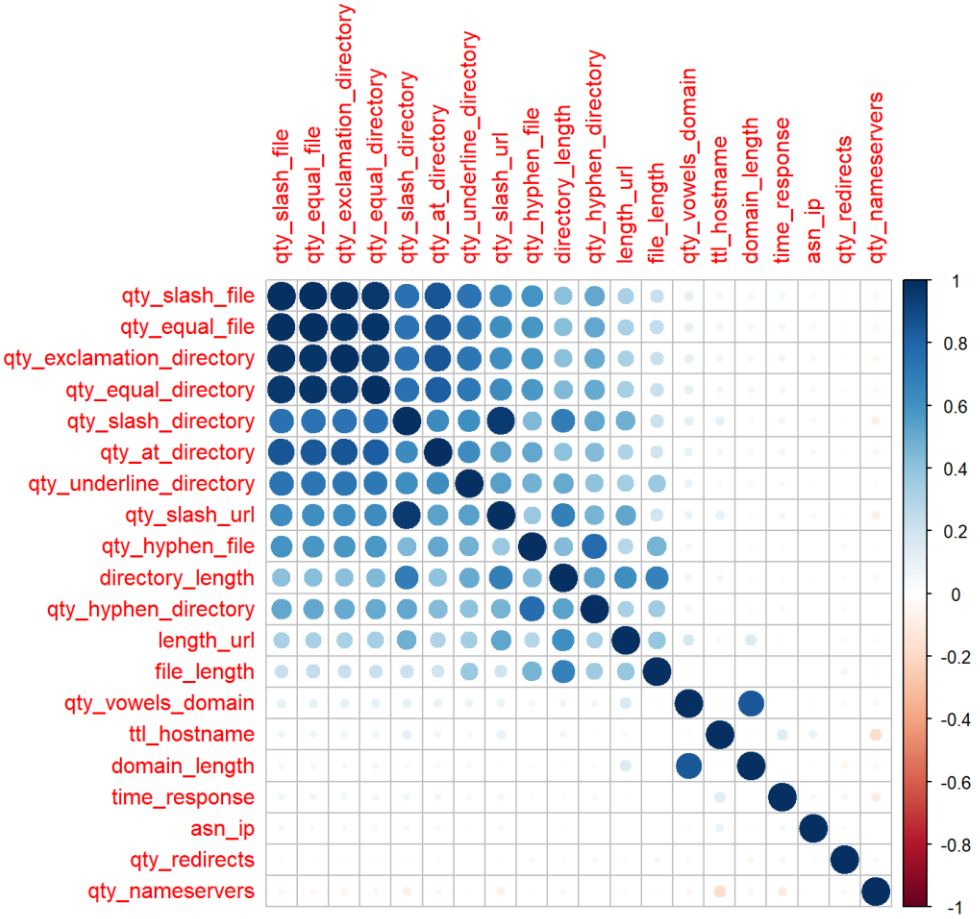

Figure 18: Correlation Matrix Heat Map Example

In summary, the highly correlated features were eliminated because it adds redundancy and complexity to the model. This is especially true if there is a cost (time or money) while training the data. Lastly, all the features that were not correlated were retained to minimize redundancy.

3.3.2 Dimension Reduction using Principal Component Analysis

Another popular method for dimension reduction is principal component analysis (PCA). PCA attempts to reduce the number of features by finding the least number of features that can explain most of the variation in the original features (Kuhn & Johnson,

2013). This method applies linear combinations of the predictors (e.g., principal components) that capture the possible variance.

In machine learning, PCA helps create uncorrelated components (or low correlation) which help improve the model's outcome (Kuhn & Johnson, 2013). However, one must use PCA with caution. For example, most features are on different scales and may have skewed distributions. Thus, it tends to give higher weights to the predictors that have larger variation. To identify the number of components to use, a scree plot helps visualize the number of components and the variability on the y-axis.

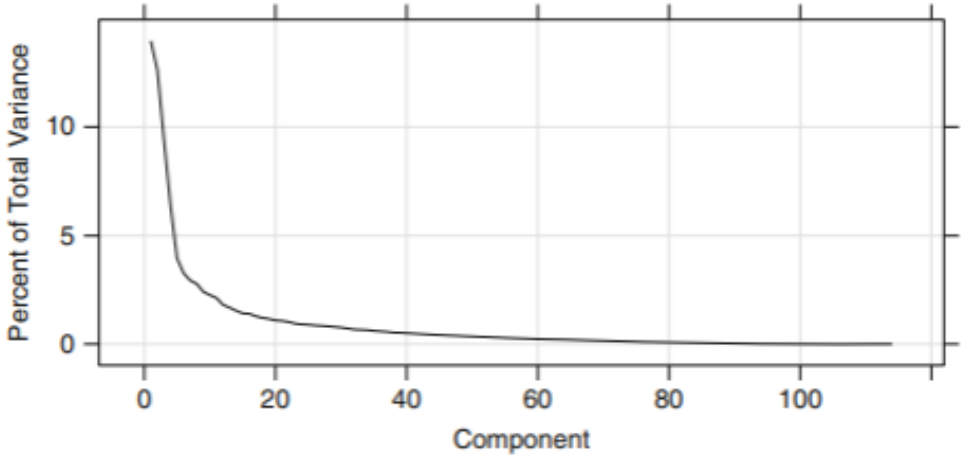

Figure 19: Scree Plot Example (Kuhn & Johnson, 2013)

The PCA results show the variation explained by each component, enabling us to determine which initial predictors were associated with the original data set (Kuhn & Johnson, 2013).

3.3.3 Wrapper Method - Recursive Feature Elimination

The wrapper methods select features by evaluating their combinations using a predictive model. These methods include recursive feature elimination and the Shapley Additive Explanation (SHAP) method.

3.3.3.1 Recursive Feature Elimination

The variable “importance” is computed using the ranking method used for feature selection (Kuhn et al, 2020). For the final subset size, the importance of the models across all resamples are averaged to compute an overall value (Kuhn et al, 2020). Using recursive feature elimination, this example plot shows that the lowest root mean square error (RMSE) is five features. In this case, the top five features are the number of slashes in the directory, the number of slashes in the URL, directory length, the number of name servers (qty_nameservers), and time response from the Mendeley 2020 data set.

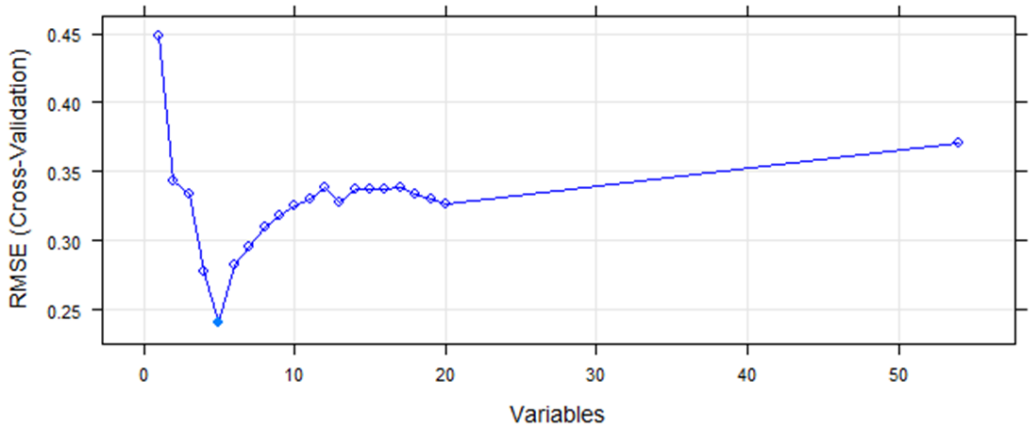

Figure 20: Recursive Feature Elimination Example

3.3.4 Feature Selection using SHAP values

A key challenge of building a predictive machine learning model is understanding the relationship between variables for the target variable (response). Many complex ML models are difficult to explain and not easily interpretable, i.e., Random Forest, XGBoost, Deep Learning, and many others. This has attracted research in explainable AI (XAI) or interpretable machine learning since higher accuracy without understanding is meaningless. Thus, SHAP values can help validate the model and provide consistency and accuracy.

In 2020, Lundberg et al. (2020) proposed a technique to interpret better complex (tree-based) models known as Shapley Additive Explanation (SHAP). Shapley value measures a feature's importance by considering the interaction with other variables using every sequence. SHAP values provide confidence in interpreting the black-box ML model.

The goal is to understand the influence of each feature in the prediction. **Figure 21** provides a conceptual framework to explain how features contribute to the overall model's output.

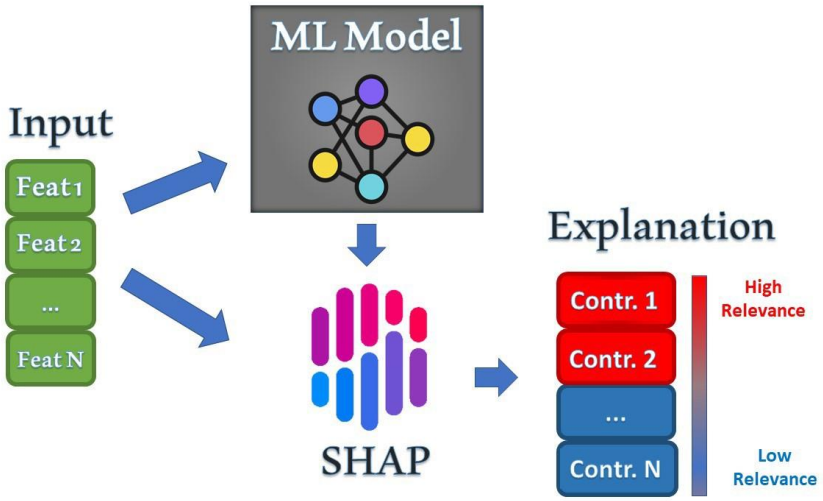

Figure 21: SHAP Feature Selection Example (Lin, 2022)

3.3.5 Feature Selection by Importance Variable (Embedded)

This feature selection method uses the built-in learning algorithm to learn their importance such as Random Forest and XGBoost. These embedded techniques select features as part of the recursive step of tree growth. This step helps examine how important each URL feature is in predicting the response (e.g., phishing or not). Additionally, this feature selection helps address how the predictors influence the

classification of URLs as phishing. **Figure** below shows the relationships between the URL features and phishing classification in a Random Forest model. The goal is to quickly visualize the most important features using the mean decrease in the Gini coefficient. This measures how each variable impacts the homogeneity of the nodes and leaves in the Random Forest model (Martinez-Taboada & Redondo, 2020). The higher the value of the mean decreased Gini, the greater the importance of the variable in the model. The top three important features are **the ratio of domain length over the URL length**, **the number of digits in the URL**, and **directory length**.

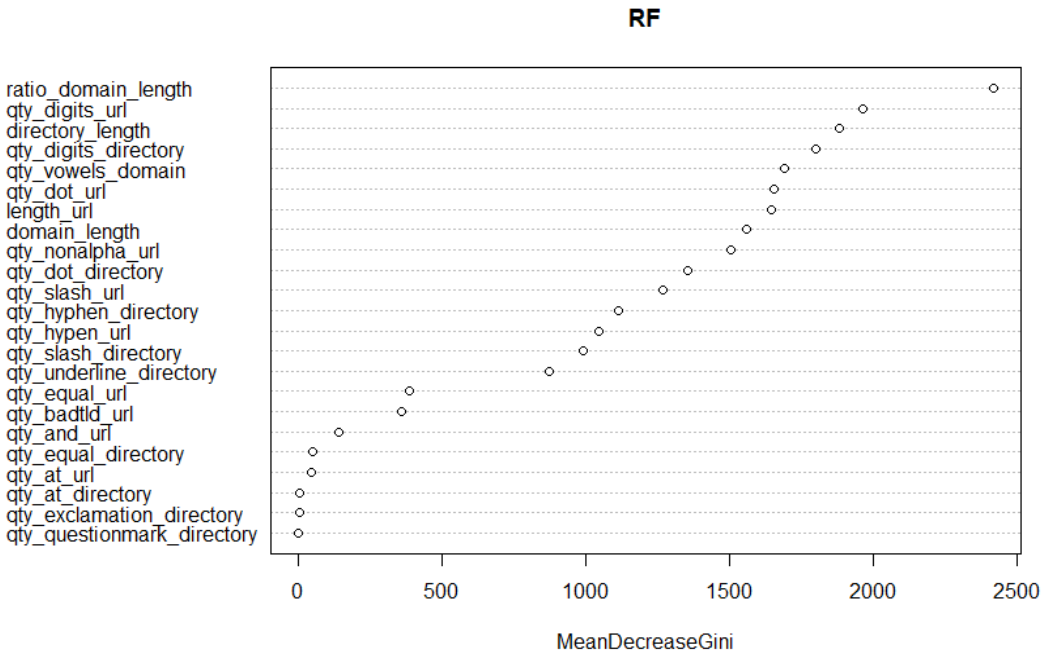

Figure 22: Embedded Feature Selection by Importance

3.3.6 Proposed Ensemble Feature Selection

This praxis combines different feature selection methods to overcome each of their single feature selection limitations. The best feature subsets are obtained by combining filter (i.e., correlation analysis) and embedded method (i.e., Random Forest).

Chen et al., (2020) show the validity of ensemble feature selection; however, the findings and algorithms used mainly applied to the medical domain. The authors used principal component analysis (filter), genetic algorithm (wrapper), and decision tree (embedded) using the C4.5 decision tree algorithm.

3.3.7 Feature Selection Comparison and Analysis

Table 12 provides a high-level summary of the key advantages and limitations of the feature selection methods.

Table 12: Feature Selection Methods Comparison Summary

Feature Selection Methods	Advantages	Limitations	Comments
Correlation (Filter)	Applicable to the high dimensional data set; Statistically tests are widely known and accepted	May be time-consuming to conduct a pair-wise comparison on a large data set; threshold subjectivity	Universally accepted
PCA (feature extraction)	Creates uncorrelated components which are suitable for ML	Possible to generate a PC that is irrelevant to the model outcome.	Must center and scale otherwise some of the predictors will have a higher magnitude
Embedded	Speed; Built-in to model; Accuracy	May lack interpretability	
Recursive Feature Elimination (Wrapper)	Speed; Accuracy; Uses RMSE to provide feature importance score	May be time-consuming; feature importance scores do not quantify the contribution of a feature to the overall accuracy of a model	The feature importance score does not guarantee that is also useful and meaningful
SHAP (Wrapper)	Enhances interpretability; measures how important a feature is to the model's output.	Cannot be used for causal inferences	
Proposed Ensemble Feature Selection	Overcome single feature selection limitations; used ensemble learning to generate feature subsets; improved accuracy by	Computation complex; may be time-consuming	Single-feature selection methods may have similar performance

	reducing the use of irrelevant features		

The following section discusses the procedures and steps for model development and introduces the different machine learning methods for phishing classification.

3.4 Model Development

Once the data is preprocessed, the next step is to split the original data set into a training and testing set. In the training phase, the machine learning classifier trains or "learns" on the training data set to classify the label of the URL. The testing phase uses the remaining data set to evaluate the classifier's performance. This is important to minimize overfitting where the model is narrowly fitted on the training data, leading to poor predictions with new data.

Splitting the data is required to enable the model to understand the relationship between features and the target variable (i.e., response). After learning this relationship, the model is used to predict the target variable in the test using only the feature values. Finally, the target variable values from the test data set are compared to those predicted by the model. This will aid in figuring out whether the model is reliable enough to be extended to unobserved data.

Since the classification model aims to predict if a URL is phishing, this is a classification problem. Once again, the research hypotheses aim to identify the most important features for classifying phishing website attacks and which machine learning classifier can provide the greatest accuracy. A summary of the research hypotheses is listed below:

H1: URL length, domain length, and the number of slashes (‘/’) are the most significant features for classifying phishing website cyberattacks.

H2: A classification model based on machine learning can be developed with at least 80% accuracy to classify phishing attacks.

H3: Random Forest is the most accurate machine learning algorithm for classifying phishing website attacks.

For the phishing classification model development, this research primarily used supervised machine learning algorithms to predict whether a URL is phishing or legitimate.

The classification algorithms are summarized in **Table 13.**

Table 13: Classification Algorithms and Categories

Classification Algorithm	Machine Learning Category
Naïve Bayes	Bayesian
Support Vector Machine (SVM)	General Classifier
K-Nearest Neighbors	Instance-Based
Random Forest	Ensemble
AdaBoost	Ensemble
XGBoost	Ensemble
Neural Net	Artificial Neural Network

The following section examines the different types of supervised machine learning algorithms.

3.4.1 Naïve Bayes

Naive Bayes is a classification algorithm based on Bayes’ Theorem and conditional probability. The Bayes theorem governs how it functions. Due to its easy implementation and minimal training time, it is popular. However, it is not preferred due

to poor performance with fewer data, presuming that the traits are independent. The main limitation of the Naïve Bayes Classifier is the assumption that the predictors are genuinely independent (Wei & Sekiya, 2022).

3.4.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is another commonly used algorithm for supervised machine learning methods (Wei & Sekiya, 2022). One key reason is that the SVM method is simple to use and effective with a variety of independent factors. The SVM algorithm creates a decision boundary called a hyperplane, which helps distinguish the data set into different classes (Wei & Sekiya, 2022). By applying the fundamental approach to nonlinear issues, it can generate useful solutions. However, SMV does not perform well on large and noisy data sets. SVM is appropriate for classification and regression problems.

3.4.3 K-Nearest Neighbors (KNN)

The KNN method generates estimates based on the distance between *k* neighbors quickly and efficiently by assuming the similarity between new and available cases. However, doing so in a huge data setting requires significant memory expenditure. The correct *k* value is crucial to the outcome, both simultaneously.

3.4.4 Random Forest (RF)

Random Forest (RF), created by Leo Brieman and Adele Cutler in 2001, is a tree-based algorithm that combines multiple trees, known as an ensemble method. This algorithm creates many trees and then combines outputs to improve the model's generalization ability. As an ensemble model, it combines numerous weak learners (individual trees) to develop a strong learner. The Random Forest algorithm is a powerful

algorithm that does not require feature scaling, is resistant to separation (overfit), and is less affected by noise. However, it requires a long training time, also necessitation a large amount of memory and processing power. RF is widely applicable to classification and regression problems.

3.4.5 AdaBoost

Adaptive Boosting (AdaBoost) is a statistical machine learning classification algorithm formulated by Yoav Freund and Robert Schapire (Wei & Sekiya, 2022). It belongs to a class of ensemble machine learning algorithms that aggregates the predictions from various weak learners (Wei & Sekiya, 2022). AdaBoost is adaptive in that weak learners are added progressively to the boosting strategy and are then adjusted in favor of those instances misclassified by earlier classifiers (Wei & Sekiya, 2022). It may be less prone to the overfitting issue than other learning algorithms in certain situations.

3.4.6 XGBoost

Extreme Gradient Boosting (XGBoost) is an algorithm based on gradient-boosted decision trees that put speed and performance in the foreground (Wei & Sekiya, 2022). It aims to reduce errors in the previous tree by producing a new tree each time. However, this process can take a long time. It can also be easily overfitted. XGBoost can solve regression, classification, ranking, and user-defined prediction problems (Wei & Sekiya, 2022).

3.5 Model Training and Validation

Model validation ensures that the model performs well with the testing data set. The most common approach for model validation is training and test split. Supervised

machine learning includes selecting a class of models, selecting hyperparameters, fitting the model to the training data, and applying the fitted model to classify labels for new data.

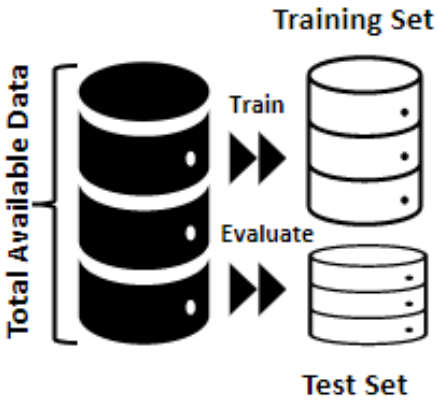

Figure 23: Model Training and Validation Split

To ensure that the machine learning model can be generalized and perform well on unseen data, the k-fold cross-validation helps mitigate overfitting. For example, one could have a 100% accuracy model using training data, but it may lack the capability to generalize for new data. This is the main reason for data splitting using k-fold cross-validation. The phishing classification models were validated using 10-fold cross-validation. This validation method splits the data into an equal number of "k" folds. After performing k-fold cross-validation, the best model is fitted to the whole training set to provide the best model's learning algorithm with more training data.

The following section describes the performance evaluation measures used for model evaluation.

3.6 Model Evaluation

"If you can't measure it, you can't improve it" - Peter Drucker

Phishing can disrupt day-to-day business operations, impact the brand, and compromise sensitive customer information for businesses. There are four classification possibilities for this phishing classification research since there are two classes (e.g., 1 for phishing and 0 for legitimate).

Figure 24 depicts four outputs as a 2x2 confusion matrix.

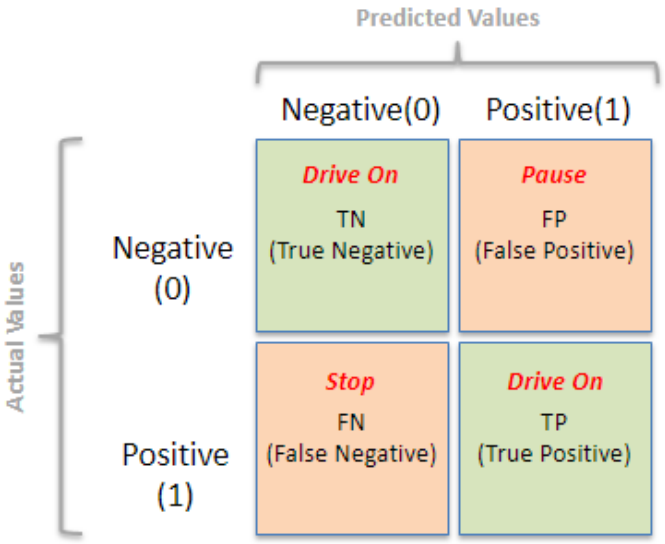

Figure 24: Binary Confusion Matrix

The four basic measures are as follows (Do et al., 2021):

- **True positive (TP**): the number of observations correctly classified as phishing - The model correctly predicted malicious URL.
- **True negative (TN)**: the number of observations correctly classified as legitimate - URL correctly classified as legitimate URL.

- **False positive (FP)**: the number of observations incorrectly classified as phishing - Good URL incorrectly classified as malicious URL.
- **False negative (FN**): the number of observations incorrectly classified as legitimate - Malicious URL incorrectly classified as legitimate URL.

This study evaluated the phishing classification models using precision, recall, F-1, and accuracy. Using the confusion matrix as data inputs, these performance evaluation metrics helped assess the model's performance and identify the best-performing classifier to classify web phishing attacks. The additional performance metrics and formulas are provided below (Do et al., 2021):

- **False Negative Rate (FNR)** shows the number of phishing observations mistakenly identified as legitimate with all phishing incidents (Do et al., 2021):

$$FNR = FN/(FN + TP)$$

- **False Positive Rate (FPR)** demonstrates the number of legitimate observations incorrectly classified as phishing relative to all legitimate instances (Do et al., 2021):

$$FPR = FP/(FP + TN)$$

A false positive rate can also impact SMBs' day-to-day operations since a legitimate website was blocked as phishing.

- **Accuracy** measures the overall rate of correctly detected phishing and legitimate observations to the total number of observations (Do et al., 2021):

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

- **Precision** represents the rate of correctly detected phishing observations to the total number of detected phishing observations (Do et al., 2021). $Precision = \frac{TP}{TP+FP}$
- **Recall** represents the rate of correctly detected phishing instances to the total number of correctly classified phishing and incorrectly classified as legitimate (Do et al., 2021). $Recall = \frac{TP}{TP+FN}$
- **F1-Score** helps understand the overall model's performance. A high f1-score means low false positive and false negatives. $F1\ score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall}$

3.7 Summary and Conclusion

The research methodology chapter discussed the process of how data was collected and preprocessed, provided an in-depth comparison of the feature selection methods, and discussed how the phishing classification model was evaluated, with the overarching goal of building a phishing classification model that can learn and perform well on multiple phishing data sets and unseen data.

In summary, there are two steps for machine learning model development. First, one must choose the cross-validation method to carry out the training. A key consideration for training the model was the balance between optimization and generalization. Therefore, a large amount of data for training helped with overfitting. The model training included splitting the data set into training and testing sets using 80% and 20%, respectively. A 10-fold cross-validation ensured a sound data partition strategy.

Second, the model was "fitted" using the respective methods or classifiers like the Random Forest, decision tree, neural net, etc. This research leveraged supervised

machine learning models to include naive bayes, support vector machines, k-nearest neighbors, Random Forest, AdaBoost, and XGBoost. The next chapter discusses the results of the phishing classification models.

Chapter 4—Results

4.1 Introduction

Chapter four presents the results of this praxis and addresses the following research questions in **Figure 25.**

RQ1: Which features are the most significant in classifying phishing website attacks?
RQ2: Can a machine learning model be built to classify phishing website attacks?
RQ3: Which machine learning algorithm has the highest accuracy in classifying phishing website attacks?

Figure 25: Research Questions

Section 4.2 discusses the feature selection results and answers research question 1. **Section 4.3** summarizes the feature selection results. **Section 4.4** presents the results of the phishing classification models using the first two data sets (e.g., Mendeley 2020 and Phishing URL) and addresses research questions 2 and 3. **Section 4.5** externally validates the phishing classification model and features using the latest publicly available data sets (e.g., Mendeley 2022). **Section 4.6** provides a summary of the model results and conclusion. The last section relates the results to the research questions and hypotheses.

As discussed in the previous section, the literature review identified gaps in using a single feature selection method, a lack of standard features for benchmarking new ML algorithms, and biases in the data set. Feature selection methods were inadequately explained and data access for replication was lacking in most previous research. Therefore, this praxis developed a phishing classification model using open-source tools, addressed the identified limitations of the machine learning model with an ensemble feature selection method, selected a subset of optimal global features, and validated these features on multiple publicly available data sets.

4.2 Feature Selection Results (Research Question 1)

RQ1: *Which features are the most significant in classifying phishing website attacks?*

Section 3.2, Table 7 depicts the most cited and baseline phishing data set comprising 112 features organized into eight groups. The first phishing data set was balanced with 47.7% and 52.3% legitimate and phishing, respectively. Each record (e.g., website) provided the quantity of each feature and its respective label. Of the 112 original features, 35 had zero variance; therefore, they were removed since they did not provide any value to the phishing classification model. **Table 14** provides the URL features filtered due to zero variance, as all the observation values were the same.

Table 14: Features Filter due to Zero Variance

Original Feature #	Features with Zero Variance	Original Feature #	Features with Zero Variance
5	qty_questionmark_url	28	qty_exclamation_domain
7	qty_at_url	29	qty_space_domain
8	qty_and_url	30	qty_tilde_domain
9	qty_exclamation_url	31	qty_comma_domain
10	qty_space_url	32	qty_plus_domain
11	qty_tilde_url	33	qty_asterisk_domain
12	qty_comma_url	34	qty_hashtag_domain
13	qty_plus_url	35	qty_dollar_domain
14	qty_asterisk_url	36	qty_percent_domain
15	qty_hashtag_url	39	domain_in_ip
16	qty_dollar_url	40	server_client_domain
17	qty_percent_url	94	params_length
22	qty_underline_domain	97	email_in_url
23	qty_slash_domain	102	time_domain_expiration
24	qty_questionmark_domain	109	url_google_index
25	qty_equal_domain	110	domain_google_index
26	qty_at_domain	111	url_shortened
27	qty_and_domain	*Note: Original features from Mendeley 2020 (Vrbančič et al., 2020)*	

The remaining original features were analyzed using the **filter-based** method and Pearson's correlation coefficient. The correlation matrix removed highly correlated or redundant features (e.g., greater than 0.75 absolute correlation). This was important because the phishing classification model performed with higher accuracy without highly correlated features. The Caret R package identified the highly correlated features, and statistical significance using a correlation matrix.

Table 15 depicts the features with high correlations.

Table 15: Features with High Correlations

Original Feature #	Features with High Correlation (>0.75)	Original Feature #	Features with High Correlation (>0.75)
5	qty_questionmark_url	36	qty_percent_domain
11	qty_tilde_url	39	domain_in_ip
13	qty_plus_url	40	server_client_domain
14	qty_asterisk_url	42	qty_hyphen_directory
15	qty_hashtag_url	43	qty_underline_directory
16	qty_dollar_url	44	qty_slash_directory
17	qty_percent_url	45	qty_questionmark_directory
18	qty_tld_url	52	qty_comma_directory
22	qty_underline_domain	53	qty_plus_directory
23	qty_slash_domain	54	qty_asterisk_directory
24	qty_questionmark_domain	55	qty_hashtag_directory
26	qty_at_domain	56	qty_dollar_directory
27	qty_and_domain	59	qty_dot_file
28	qty_exclamation_domain	60	qty_hyphen_file
30	qty_tilde_domain	62	qty_slash_file
31	qty_comma_domain	63	qty_questionmark_file
33	qty_asterisk_domain	65	qty_at_file
34	qty_hashtag_domain	66	qty_and_file
35	qty_dollar_domain	*Note: Original features from Mendeley 2020 (Vrbančič et al., 2020)*	

After removing the highly correlated features, **Figure 26** depicts the remaining features, or 27.7% of the original 112 features used to build the phishing classification

model. The correlation matrix depicts features with a positive correlation in blue and a negative correlation in red.

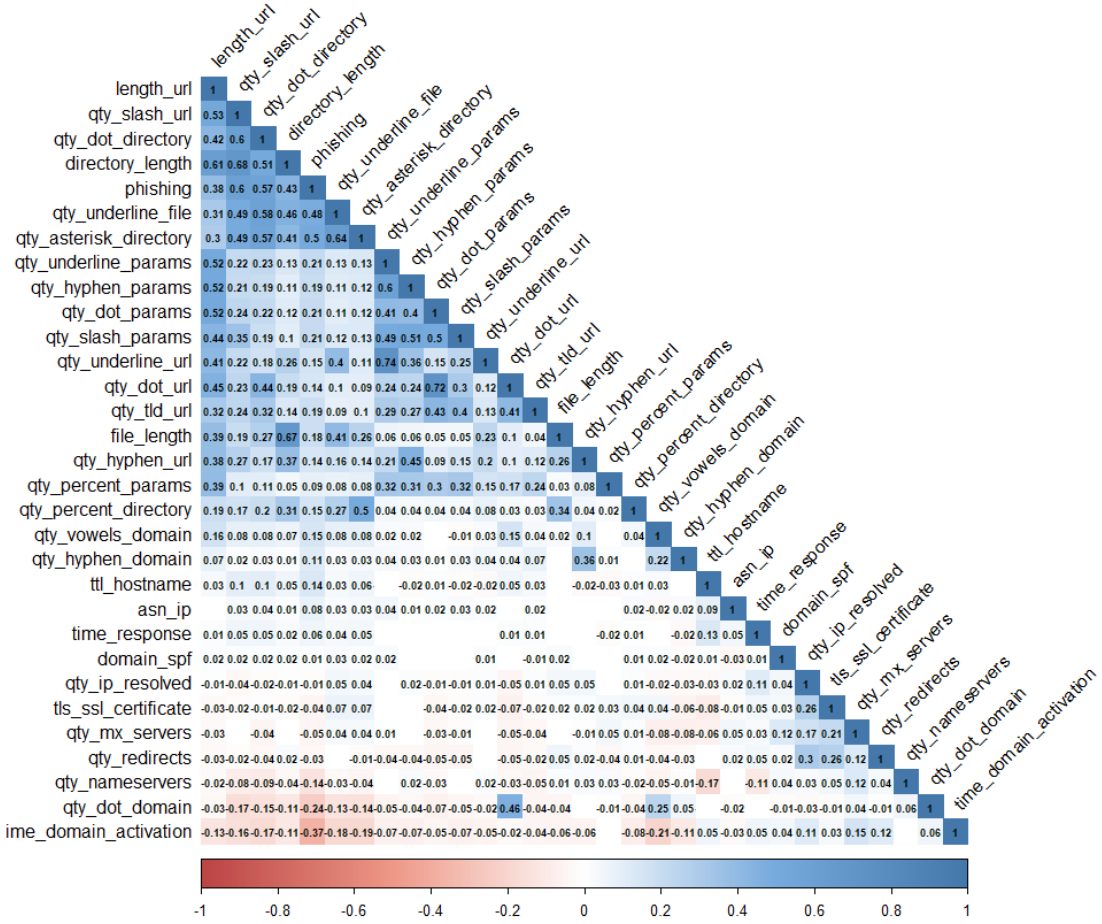

Figure 26: Correlation Matrix

Also, **Figure 26** shows phishing positively correlates with the following features: **the number of slashes in the URL**, the **number of dots in the directory**, **directory length**, and **URL length**, respectively. All the numbers displayed are statistically significant (p-value <0.05), and the insignificant values are blank white cells (reference **Appendix A, Table A-2).**

Using the ensemble feature selection approach, the next step was integrating the **embedded** feature methodology by ranking the features. This method helped identify

features by learning their importance to the model's accuracy. Using the feature selection embedded in the Random Forest classifier, the Caret R package assisted with ranking features by importance.

Figure 27 shows the most important variables of the target (phishing). The more the ML model depends on a given feature to make predictions, the more "important" it is for the model. The variable importance provides a quantitative ranking of the feature's importance to the model. Specifically, the important variable function estimated the variable's importance. It showed that the **directory length, number of slashes in the URL**, **URL length, the number of dots in the directory**, and **file length** were the most important features to classifying phishing websites in the Mendeley 2020 data set.

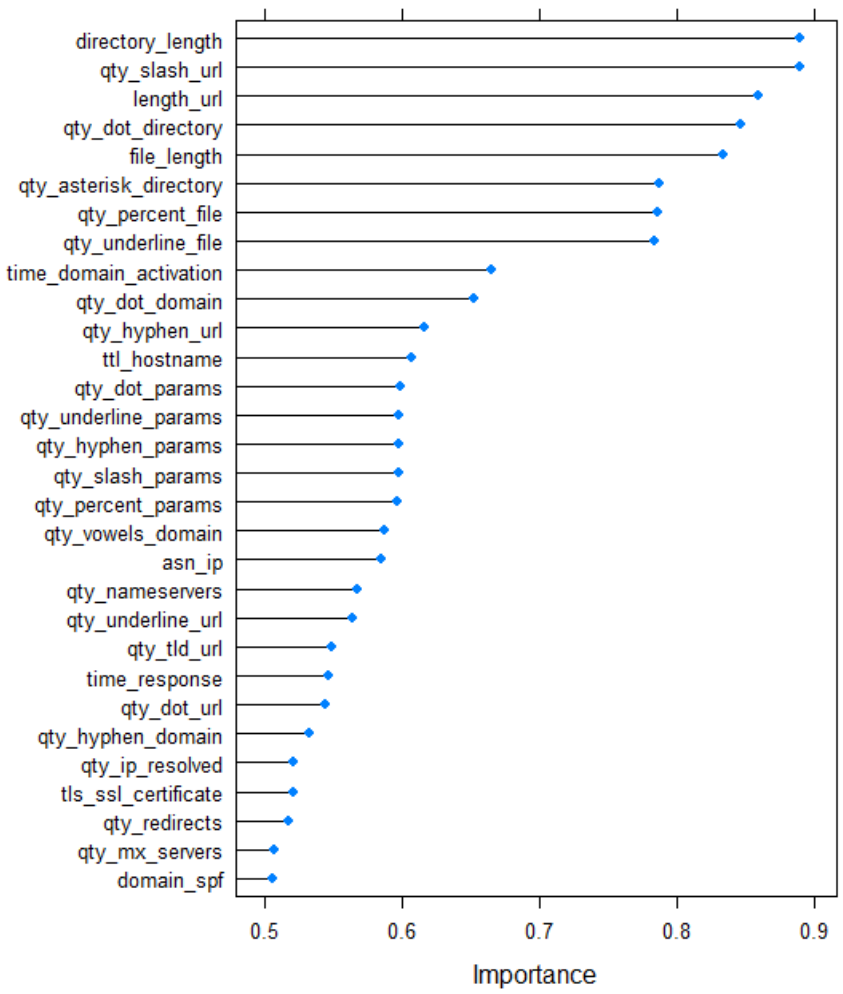

Figure 27: Features by Importance

Based on the critical features identified from data set #1 (Mendeley, 2020), this study used these "learned" URL features to build a parser tool for the second data set. This was critical because the Phishing Site URLs data set (Tiwari, 2022) only provided two features: a URL and its label. **Table 16** below depicts the feature selection results for the phishing classification model with 24 learned features and three features engineered using the first data set.

Table 16: Phishing Classification Model - Baseline Features

#	Feature Name	Description	Significant Or Important

1	url	Website link	
2	qty_slash_url	Total number of '/' in URL	(P)(S)(RF)
3	qty_dot_url	Total number of '.' in URL	(S)
4	qty_equal_url	Total number of '=' in URL	
5	qty_and_url	Total number of '&' in URL	(S)
6	qty_hypen_url	Total number of '-' in URL	
7	qty_at_url	Total number of '@' in URL	
8	qty_digits_url	Total number of [0:9] in URL	
9	qty_nonalpha_url*	Total number of Non-Alphanumeric in URL	
10	length_url	Total number of characters in URL	(P)(RF)
11	qty_badtld_url	Total number of bad top-level domains in URL	
12	domain_name	Name of the domain (string)	
13	domain_length	Total number of characters in Domain	(S)
14	qty_vowels_domain	Total number of 'A, E, I, O, U' in Domain	
15	ratio_domain_length*	Total number of characters in Domain / URL length	(S)
16	ratio_digits_nonalpha*	Number of [0:9] / Number of Non-Alphanumeric	
17	directory_length	Total number of characters in the Directory	(P)(RF)
18	qty_slash_directory	Total number of '/' in the Directory	(S)(RF)
19	qty_hyphen_directory	Total number of '-' in the Directory	
20	qty_underline_directory	Total number of '_' in the Directory	
21	qty_equal_directory	Total number of '=' in Directory	
22	qty_at_directory	Total number of '@' in Directory	
23	qty_exclamation_directory	Total number of '!' in the Directory	
24	qty_dot_directory	Total number of '.' in the Directory	(P)(S)(RF)
25	qty_questionmark_directory	Total number of '?' in the Directory	
26	qty_digits_directory	Total number of [0:9] in Directory	
27	phishing	Label as phishing or legitimate	
Key: Pearson (P); SHAP (S); Random Forest (RF); *Feature Engineered			

Using the Random Forest's embedded feature selection, **Figure 28** shows the important features based on the mean decrease Gini. For example, the higher the value of the mean decrease Gini score, the higher the importance of the variable in the model (Martinez-Taboada & Redondo, 2020). One can see that the **ratio of domain length to URL length** is the most important feature of the phishing classification model. The other key features include the **directory length**, **domain length**, and the **ratio of digits to**

nonalphanumeric characters.

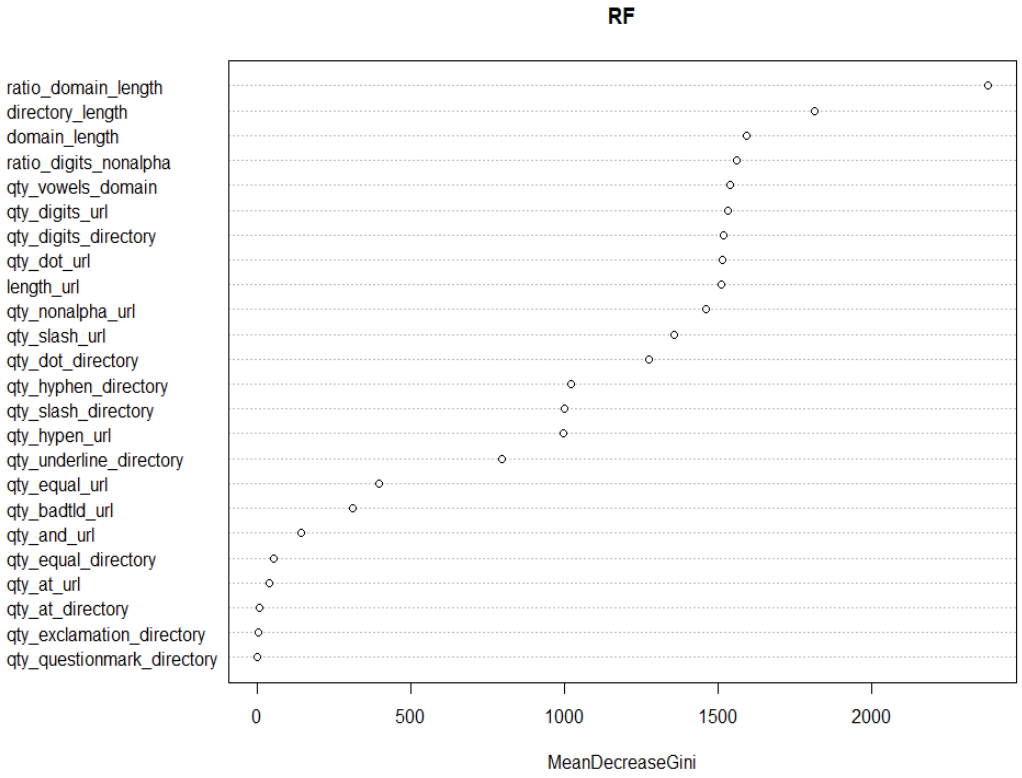

Figure 28: Key Features using Random Forest Classifier

Next, the SHAP method was applied to calculate the feature's impact on the model's output to enhance the phishing classification model's interpretability. **Figure 29 s**hows the SHAP values and features in order of importance with their corresponding mean SHAP value on the y-axis and the x-axis indicates the SHAP value.

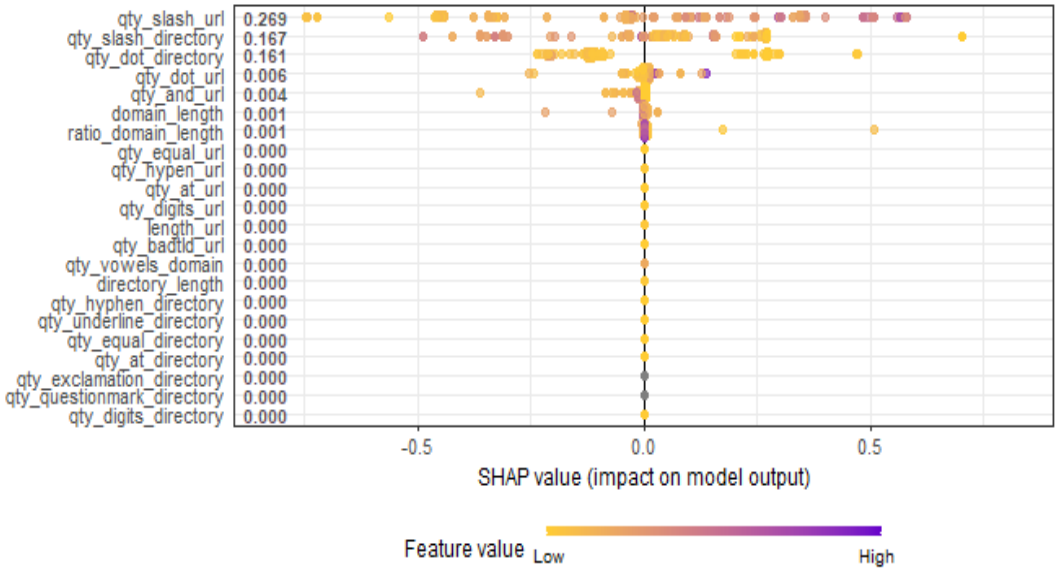

Figure 29: Top Features by SHAP values

Similar to the filter and embedded method, SHAP results show that the **number of slashes in URL**, **number of slashes in the directory**, **number of dots in the directory**, **number of dots in URL**, **number of "&"s in URL**, **domain length**, and the **ratio of domain length to the URL** to be the most important features that directly contribute to the classification model. **Figure 30** also shows that the other features with a zero SHAP value have little impact on predicting phishing URLs.

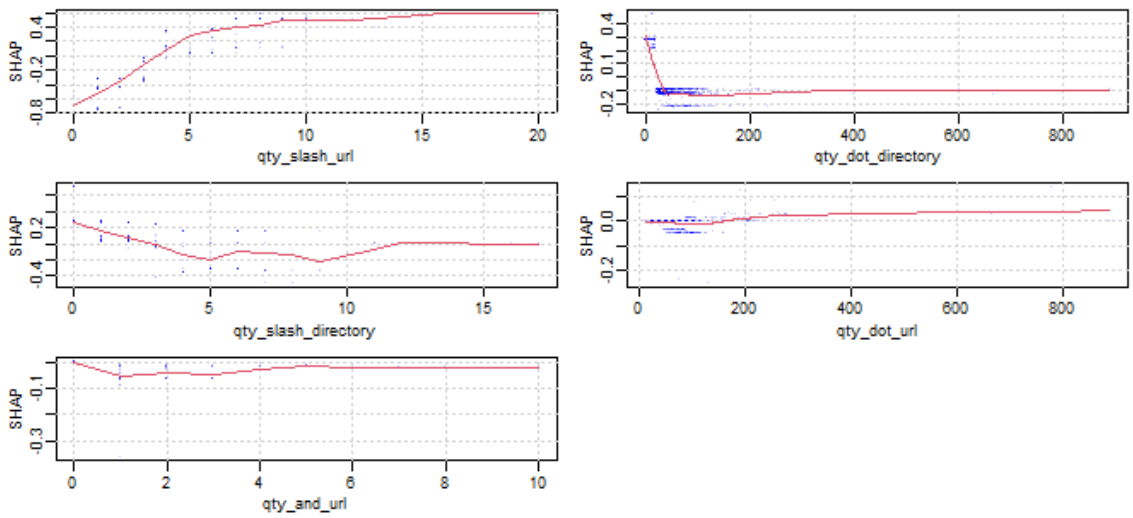

Figure 30: SHAP values and Features

Each blue dot is a row (URL link) from the original data set. Results show that the number of slashes ("/") in the URL and directory have higher positive SHAP values and contribute more towards phishing classification. The SHAP values show that Phishing URLs have more slashes in URLs and directories.

For the qty_slash_url feature, the SHAP values increase with more slashes until around five slashes then the SHAP value increases incrementally. The number of dots in the URL increases the SHAP value almost linearly. Oddly, the number of slashes in the directory decreases SHAP values until about five increases to 6 and decreases to 10.

4.3 Summary of Features Selection

The significant and substantively important features were identified using an ensemble feature selection method. First, the filter method removed features with zero variance and high correlation to increase the effectiveness of the classification model. Second, the embedded method was used to rank features by importance. Last, the SHAP method was applied to enhance the model's interpretability and impact on the model's

output. This ensemble approach aided in selecting the best feature subset for phishing classification.

The goal was to develop an accurate, robust, and more interpretable model for phishing detection that can be applied across multiple phishing data sets. The model results show that it is important to select optimal features not to increase model complexity which could lead to overfitting. This resulted in 31 of the original 112 features remaining for the phishing classification model, representing a reduction of 72.8%. Of the 31 features, only the significant URL-based features were reproduced in data set #2 using the developed URL parser.

The following section presents the phishing classification model results of applying the 27 baseline features (i.e., Table 15).

4.4 Phishing Classification Model Results (Research Questions 2 & 3)

This section provides the results of the phishing classification model. Since the phishing classification model aims to predict if a URL is phishing, this is a binary classification problem. Once again, the research goals were to identify the significant and important features for classifying phishing website attacks and which machine learning classifier can provide the greatest accuracy.

RQ2: Can a machine learning model be built to classify phishing website attacks?

RQ3: Which machine learning algorithm has the highest accuracy in classifying phishing website attacks?

The various machine learning algorithms used to classify phishing were the Random Forest, Ada, KNN, Adaboost, Decision Tree, Logistic Regression, SVM, Neural Net, LDA, and Naïve Bayes. The model's performance was analyzed using accuracy, precision, recall, f1-score, false positive rate (FPR).

4.4.1 Data Set #1 Results (Vrbančič et al., 2020)

As discussed in the research methodology chapter, multiple feature selection techniques were applied and compared to achieve the most accurate phishing classification model. **Table 17** depicts the phishing classification model results using accuracy, precision, recall, f1-score, and false positive rate (FPR). The Random Forest model achieved the highest accuracy using the ensemble feature selection technique and only 31 of the original 112 features. While there were 31 features, this research used only the URL-based features and removed the string features (e.g., domain name, URL name). Of note, the Random Forest model also had the lowest false positive rate, a great performance indicator for detecting phishing websites.

Table 17: Data Set #1 Model Results

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score	FPR
Random Forest	91.93%	90.37%	93.00%	91.67%	9.02%
KNN	89.65%	85.49%	84.60%	85.04%	12.17%
Adaboost	86.34%	86.48%	87.55%	87.01%	14.98%
Decision Tree	85.51%	82.96%	90.96%	83.08%	20.45%
Logistic Regression	83.88%	82.87%	87.19%	84.86%	19.74%
SVM	83.70%	81.49%	89.02%	85.09%	22.13%
Naïve Bayes	80.74%	74.04%	97.24%	84.07%	37.33%

Mendeley 2020: 58,645 rows; 10 cross-validation; 80/20 split

4.4.2 Data Set #2 (Tiwari, 2022)

This phishing URL data set lacked features. This study aims to develop a phishing classification model using an ensemble of phishing data sets to develop the best classification model. While using the key features from the most-cited phishing data set (Vrbančič et al., 2020), the goal was to develop a URL parser with minimal features and apply feature engineering.

Using the first phishing data set, key features were extracted. Using the developed URL parser and engineered features, this study created 27 machine learning model development features. The URL parser helped better determine which features from data set 1 and data set 2 can be ensembled to provide the best phishing classification model across multiple data sets. The URL parser created and extracted 27 features. These include URL length, domain length, and the number of special characters (e.g., '/', '=', '_', '!', '.', '?', '@') and feature engineered the following: **ratio of the domain length to URL length**, **non-alphanumeric**, and **ratio of the number of digits to non-alphanumeric**.

Principal component analysis (PCA) was also applied to validate that 27 features were enough to explain over 90% of the variation. **Figure 31** depicts the PCA results in a scree plot.

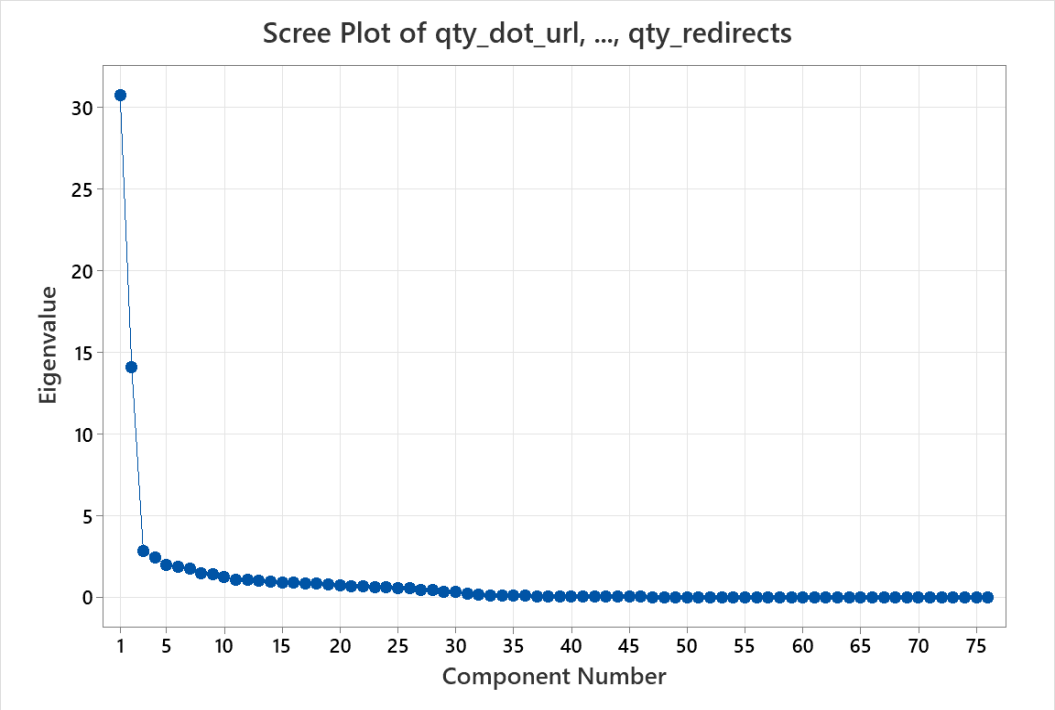

Figure 31: Principal Component Analysis (PCA) on Data Set #1 – Scree Plot

The scree plots show that the first three principal components account for the most significant proportion of the variability (about 63%) and the remaining components explain less variability after about 20 components. Results suggest that researchers should use at least 20 components to account for over 90% of the variability. Additionally, 11 components or features can explain about 80% of the variability, with the most significant gains around five components (68.5%). The full eigenanalysis of the Correlation Matrix is in **Appendix A, Figure A-1**.

After extensive modeling and feature testing, the results show that the Random Forest model performed the best and achieved an accuracy of about 94%. **Figure 32** depicts the top features selected for the phishing classification model and applied to data set #2 (Tiwari, 2022).

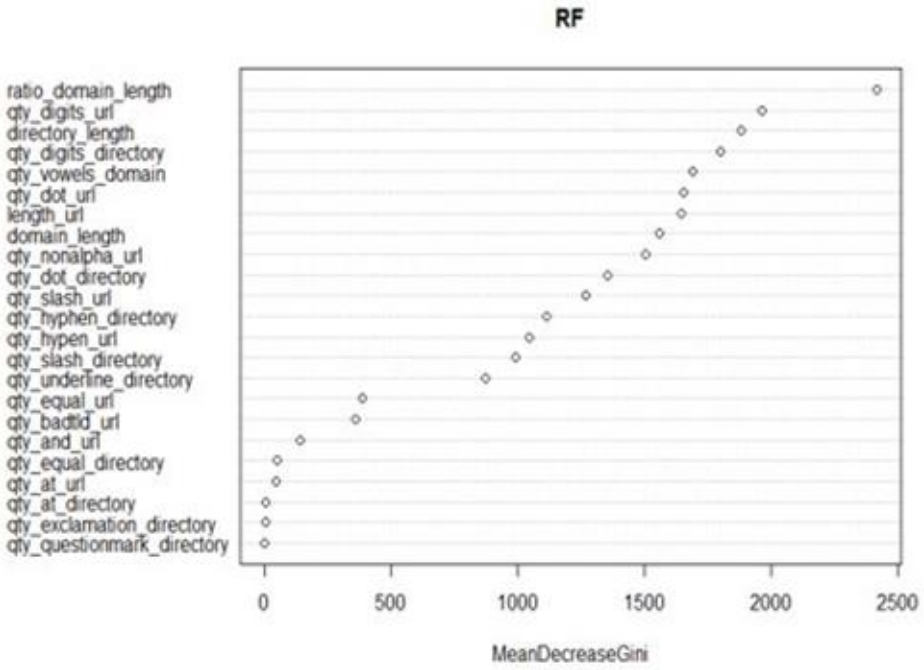

Figure 32: RF Important Features using Data Set #2 (Tiwari, 2022)

Results also show that an URL parser can be developed using key features from an external phishing data set (e.g., DS1). The key features from data set 1 can be combined with another data set to develop a phishing classification model with greater accuracy than 80%. This also demonstrates the transfer learning of features from data set 1 to data set 2.

Table 18 presents all the phishing classification model results using the learned Mendeley (2020) features, the developed URL parser, and data set # 2.

Table 18: Data Set #2 Model Results

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score	FPR
Random Forest	93.73%	91.24%	93.70%	92.54%	8.31%
XGBoost	86.73%	82.31%	86.01%	84.14%	4.04%
KNN	80.19%	81.76%	92.64%	86.86%	18.14%
Adaboost	77.19%	76.19%	98.44%	85.90%	15.42%
Logistic Regression	76.39%	77.15%	95.10%	85.36%	19.75%
Decision Tree	75.69%	75.59%	96.81%	84.89%	18.19%
Support Vector Machine	74.39%	73.87%	99.23%	84.69%	27.56%

Kaggle - Phishing Site URLs (Tiwari, 2022): 507,195 rows; 10 cross-validations; 80/20 split

4.5 External Model Validation

4.5.1 Phishing Website Detection Data Set #3 (Pham, 2022)

Building upon the key features selected and using the URL parser previously developed, this data set served as a validation data set. As noted in the literature review, there was a lack of validation of the findings. For that reason, it was essential to be able to validate a subset of features that can be generalized across all phishing data sets.

As mentioned before, data set #3 (Mendeley 2022) is the most recent publicly available phishing data set, but it only provided URLs and a label for phishing or legitimate. The phishing classification model was further analyzed using the 27 features from the ensemble feature selection results to determine the optimal number of features.

Through systematic testing and extensive modeling of feature subsets, results show that an ML model can be built using five predetermined features. First, the feature selection wrapper Recursive Feature Elimination (RFE) method was used to help identify the optimal number of features or the final subset size. From the Mendeley 2020 data set, the top five features were the number of slashes in the directory, the number of slashes in the URL, directory length, the number of name servers, and time response; however, this was problematic since this included two non-lexical features (reference **Figure 20** in **Section 3.3.3**). Despite the findings, the results validated that five features were sufficient when paired with the Random Forest classifier. Second, the ensemble feature selection helped identify the statistically significant, most important, and explainable features using Pearson's correlation coefficients, SHAP, and embedded methods.

The research identified the following features**: the ratio of domain length to URL length**, the **total number of characters in the domain**, the **ratio of the number of digits to nonalphanumeric characters**, the **number of dots in the directory**, and the **number of slashes in URL** are significant features that can be universally applied across any phishing data set. Of note, these subsets of features reflect three original features and two engineered features (e.g., the ratio of the domain to URL length).

Table 19 shows the phishing classification model results using five features. These results show that these features are global features that could be used for benchmarking other phishing data sets. Using only five predetermined features, the model achieved above 90% accuracy in classifying if a URL link is phishing or not.

Table 19: Data Set #3 Model Results

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score	FPR
Random Forest	94.80%	99.02%	90.50%	94.57%	0.90%

KNN	94.50%	97.30%	92.14%	94.34%	1.90%
XGBoost	94.10%	99.60%	88.20%	94.41%	0.60%
Adaboost	93.75%	99.77%	87.70%	93.35%	0.20%
Naïve Bayes	93.70%	98.34%	88.90%	93.65%	1.50%
Support Vector Machine	93.35%	97.38%	89.10%	93.05%	2.40%
Logistic Regression	92.05%	94.50%	89.30%	91.83%	5.20%

Mendeley: Phishing Website Detection 2022 (Pham, 2022): 2M rows; 10 cross-validations; 80/20 split

Results and key features for the best-performing machine learning classifier, the best-performing Random Forest, are depicted in **Figures 33 and 34.**

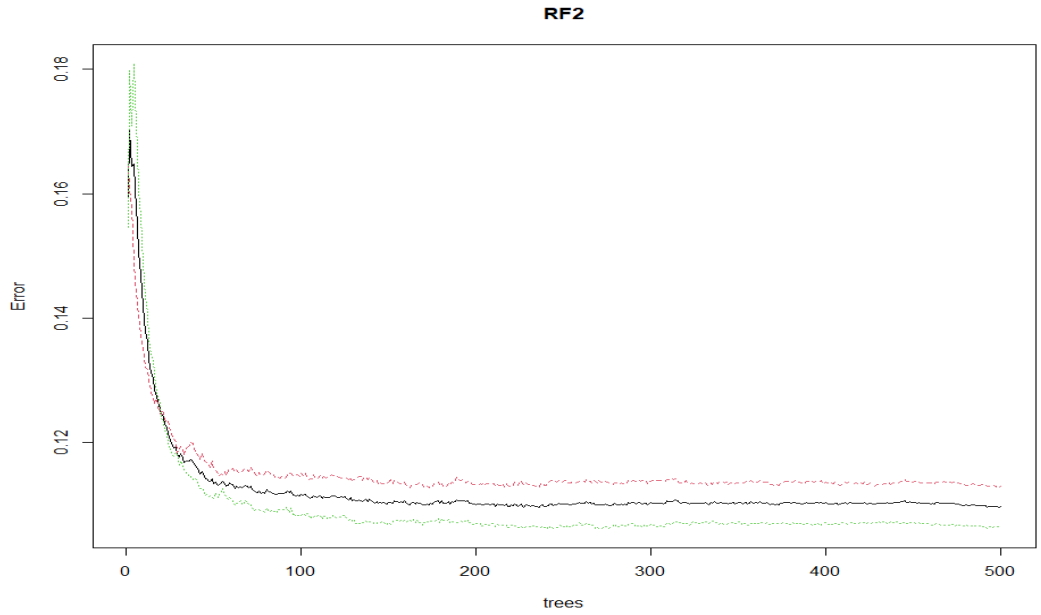

Figure 33: Phishing Classification Model with Random Forest Classifier

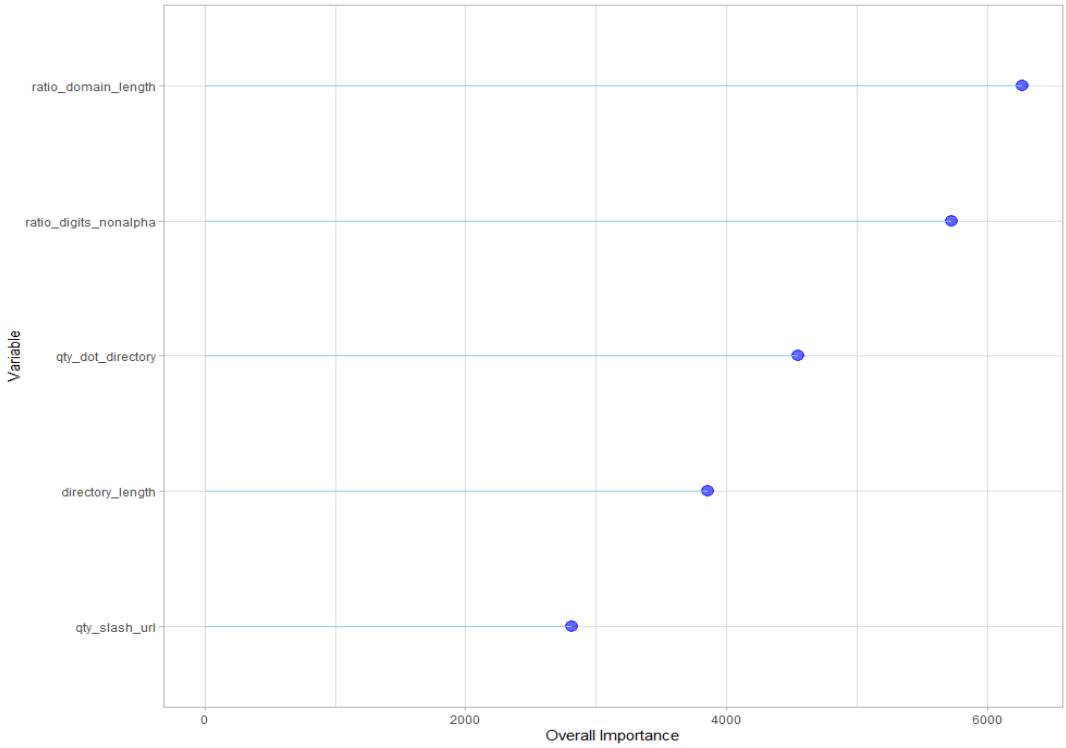

Figure 34: Five Important Features for Phishing Classification

4.5.2 Phishing Validation Data Set #4

The fourth validation data set serves two purposes. First, it is another validation data set on the latest publicly available data using various sources. Secondly, the data set contributes to the phishing community with a data set that can be used for phishing model benchmarking. The results reflect the multi-layered approach used across several phishing data sets, including the ensemble feature selection method from data set 1, featuring engineering, and URL parser applied to data set 2, and the optimal five features from data set #3.

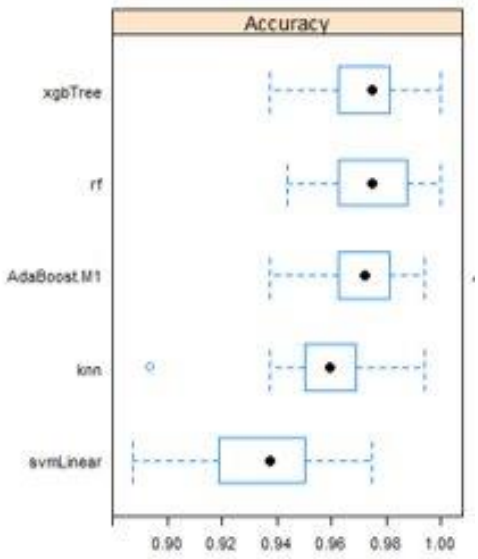

Figure 35: ML Model Results - Accuracy

Figure 35 shows that a phishing classification model using five predetermined features is possible and can achieve an accuracy greater than 90%. The best-performing classifier was the Random Forest, with about 97% accuracy. The Random Forest, XGBoost, and AdaBoost were the top-performing algorithms, but this is not surprising given that all three are ensemble ML models and generally perform better than a single model. The results indicate that the **domain-to-URL length ratio and the number of digits to nonalphanumeric characters** were the top two engineered features for the validation data set. Other researchers should consider adding these features when building a phishing classification model.

The proposed five features provide a baseline set to build a phishing classification model for phishing data sets. Firms can develop a URL parser and feature engineering to obtain these five features: **the ratio of domain length to URL length**, the **total number of characters in the directory**, the **ratio of the number of digits to nonalphanumeric**

characters, the **total number of dots in the directory**, and the **total number of slashes in URL** to classify phishing website with greater than 90% accuracy.

Table 20 shows the best-performing algorithm (Random Forest) results across all four phishing data sets with the proposed five predetermined features for ML-phishing benchmarking.

Table 20: Random Forest Results using five predetermined features

Top-performing algorithm	**Mendeley 2020**	**Phishing URLs**	**Mendeley 2022**	**Song 2023**
Random Forest	91.93%	93.73%	94.80%	97.25%

4.6 Model Summary and Conclusion

The praxis used four publicly available phishing data sets to learn significant features from Mendeley 2020, built a URL parser and, applied it to data set #2, then validated the phishing classification model on the latest publicly available data set (Pham, 2022). Next, the praxis identified an optimal subset of features to benchmark other ML classifiers and data sets. **Table 21** summarizes the data sets, number of records, features, and purpose.

Table 21: Data Set Records and Features

#	Data Set Name	Number of Records	Original Number of Features	Final Number of Features	Purpose
1	**Mendeley 2020**	88,647	112	31	Significant Features
2	**Phishing URLs**	507,195	2	27	URL Parser
3	**Mendeley 2022**	2,000,000	2	27	Validation
4	**Song 2023**	20,0000	2	5	Optimal Number of Features

Firms can leverage the power of open-source tools (e.g., R studio) and machine learning to classify phishing websites and increase their cyber readiness. This chapter analyzed multiple phishing data sets to overcome data set bias and identified an optimal set of features that can be universally applied across any phishing URL or data set with high accuracy (e.g., greater than 80%).

First, ensemble feature selection and techniques were applied to identify the optimal global set of URL-based features. The list of significant features can be found in **Appendix A, Table A-2**. This included running multiple iterations of every subset of features to achieve the highest accuracy. Using the ensemble feature selection approach, **the ratio of domain length to URL length**, the **total number of characters in the directory**, the **ratio of the number of digits to nonalphanumeric characters**, the **total number of dots in the directory**, and the **total number of slashes in URL** were identified and proposed as an optimal set of features that can be applied to any phishing data set with greater than 90% accuracy.

Second, the results demonstrate that a machine learning model can be built to classify phishing websites accurately (e.g., greater than 80%). More specifically, the Random Forest model achieved the highest accuracy across all data sets.

Lastly, a phishing classification model (PCM) can be built with high accuracy with only five URL-based features. Firms can and should use the phishing classification model to reduce risk and detect phishing website threats with cybersecurity analytics, open-source tools, and up-to-date publicly available data sets.

4.7 Research Questions and Hypotheses Results

This section discusses the results related to the research questions and hypotheses while positioning these results within the literature reviewed in Chapter 2. A phishing classification model to detect phishing websites was built using machine learning and multiple phishing data sets to answer the following research questions:

RQ1: Which features are the most significant in classifying phishing website attacks?
RQ2: Can a machine learning model be built to classify phishing website attacks?
RQ3: Which machine learning algorithm has the highest accuracy in classifying phishing website attacks?

This section provides the study results for the three research hypotheses and is summarized in **Table 22**:

Table 22: Summary of Research Hypotheses

Research Hypotheses	Results
H1: URL length, domain length, and the number of slashes (‘/’) are the significant features for classifying phishing website cyberattacks.	No. URL length, directory length, the number of dots in the directory, and the number of slashes are significant features. Additional details are described in **Section 4.2**
H2: A classification model based on machine learning can be developed with at least 80% accuracy to classify phishing attacks.	Yes. A classification model with 97.25% accuracy was developed in this praxis.
H3: The Random Forest is the most accurate machine learning algorithm for classifying phishing website attacks.	Yes. The Random Forest is superior to the other machine learning algorithms, according to analyses conducted.

4.7.1 Hypothesis 1- Significant Features

H1: URL length, domain length, and the number of slashes (‘/’) are the most significant features for classifying phishing website cyberattacks.

The first research hypothesis was tested by applying multiple feature selection techniques to the original 112 features from the well-cited phishing data set (e.g., Mendeley 2020). After extensive modeling and applying the filter-method, Pearson's correlation coefficient shows that the number of slashes in a URL, the number of dots in the directory, directory length, and URL length are positively correlated and statistically significant with phishing.

Table 23 results show the most statistically significant (e.g., p-value <0.05) features for classifying phishing websites. Since the p-values are less than 0.05, one can conclude that there is a statistically significant relationship between these features and the response variable, phishing.

Table 23: Significant Features

Baseline Features	**Pearson Correlation Coefficient**
Number of Slashes ('/') in URL	0.60***
Number of Dots ('.') in URL Directory	0.57***
Total Number of Characters in URL Directory	0.43***
Total Number of Characters in URL	0.38***

*Note: *** denotes a p-value <0.001*

4.7.2 Hypothesis 2 - Phishing Classification Model Evaluation

H2: A classification model based on machine learning can be developed with at least 80% accuracy to classify phishing attacks.

Despite the evolving nature of phishing attacks, a classification model can be built with greater than 80% accuracy. The results show that the developed phishing classification model works well with any phishing data set with or without features. This study demonstrated that a machine learning-based classification model is effective at

classifying phishing websites, from the preprocessed data to data sets with only URL links. This makes sense because malicious links are generated to deceive users, and often, these phishing links are structurally different from registered legitimate links.

Furthermore, results show that the developed features, such as the ratio of domain length to URL length, the number of non-alphanumeric in the URL, and the ratio of numbers to non-alphanumeric, should be considered when developing phishing classification models. The security research community does not commonly use these three features and should be considered when developing future phishing classification models.

4.7.3 Hypothesis 3 - Random Forest Model

H3: Random Forest is the most accurate machine learning algorithm for classifying phishing website attacks.

Lastly, results show that the Random Forest algorithm is the best machine learning algorithm for a multi-phishing data set, as seen in **Section 4.4**. Despite the data sets or feature selection methodology, the Random Forest algorithm achieved the highest accuracy. The XGBoost, another ensemble learning algorithm, performed similarly to the Random Forest model. Key similarities are that it does not require data preprocessing and handles outliers and missing data. It also has a built feature selection and, thus, only includes predictors (e.g., features) that help maximize accuracy.

Chapter 5—Discussion and Conclusions

5.1 Discussion

Despite machine learning and cybersecurity analytics advancements, SMBs are still challenged with phishing detection, which often leads to stolen credentials and other sensitive information. While ML-based phishing detection solutions continue to improve, there is still a gap in feature selection methods that obtain high accuracy and access to quality phishing data sets. Additionally, the literature shows a lack of a standard set of global features for phishing website detection. To that end, this study analyzed multiple feature selection techniques, used several phishing data sets to build a phishing classification model, and proposed an optimal set of global features for phishing classification model benchmarking.

Table 24 summarizes the results related to the research questions and hypotheses while positioning these results within the literature reviewed in Chapter 2.

Table 24: Summary of Research Questions and Hypotheses

Research Question and Hypotheses	Results
RQ1: Which features are the most significant in classifying phishing website attacks? **H1**: URL length, domain length, and the number of slashes (‘/’) are the significant features for classifying phishing website cyberattacks.	URL length, directory length, the number of dots in the directory, and the number of slashes are significant features
RQ2: Can a machine learning model be built to classify phishing website attacks? **H2**: A classification model based on machine learning can be developed with at least 80% accuracy to classify phishing attacks.	Yes. A classification model with 97.25% accuracy was developed in this praxis.

RQ3: Which machine learning algorithm has the highest accuracy in classifying phishing website attacks? **H3**: The Random Forest is the most accurate machine learning algorithm for classifying phishing website attacks.	Yes. Random Forest is superior to the other machine learning algorithms, according to analyses conducted.

5.2 Conclusions and Recommendations

SMBs are often the **easier targets for cybercriminals due to antiquated infrastructure and lax protection plans against cyberattacks**. Cybersecurity analytics and machine learning can enhance SMBs' cybersecurity readiness by using open-source tools to classify phishing web attacks. The sophistication of phishing website attacks continues to grow as threat actors adopt new techniques and technologies. SMBs do not have the financial, human, and technical resources that larger enterprises have, necessitating cost-effective interventions to enhance cybersecurity. Hence, this study tested different methods and demonstrated that SMB phishing challenges could be addressed with open-source tools, historical phishing data, and machine learning. With minimal capital expenditure, SMB owners can use machine learning and open-source tools to detect phishing attacks and prevent future cyberattacks proactively.

Despite the importance of SMBs, the praxis analysis found **limited cybersecurity research focused on SMBs phishing detection using machine learning, insufficient SMB cyber defense resources, and minimal publicly available data sets to empower SMBs to harness the value of machine learning and cybersecurity analytics.** The free tools and resources the government provides are insufficient to protect against phishing websites. SMBs need more than a list of resources and training to deter phishing

cyberattacks. A general cost estimate is provided for developing and implementing a phishing detection tool in **Appendix A, Table A-4.**

After extensive modeling and feature testing, the model results show that **the Random Forest model performed the best and achieved an accuracy of about 97%.** Results show that an URL parser can be developed using key features from an external phishing data set (e.g., DS1). The key features from data set 1 can be combined with another data set to develop a phishing classification model with greater accuracy than 80%. This also demonstrates the transfer learning of features from data set 1 to data set 2.

Using the proposed **ensemble feature selection techniques**, one can identify a **subset of features that can improve model performance and enhance the model's interpretability with fewer features**. This ensemble feature selection technique performed well against several recent publicly available phishing data sets and demonstrated transfer learning of features across multiple data sets. In particular, the following subset of URL-based features performed well across multiple phishing data sets: the ratio of domain length to URL length, the total number of characters in the directory, the ratio of the number of digits to nonalphanumeric characters, the total number of dots in the directory, and the total number of slashes in URL. Experimental results suggest that the optimal subset of features performs best with the Random Forest classifier achieving over 90% accuracy using only (three of the original 112 features and two feature engineered) and an accuracy of 97.25% using the recent sample from openphish.com (Data Set #4)

SMB owners should consider leveraging machine learning with Random Forest to detect and classify phishing websites using the five predetermined features. While it is

impractical for SMBs to check every URL link, the phishing classification model can be further developed and deployed into a cloud environment or web browser and integrated with day-to-day operations. With machine learning, open-source tools, and publicly available data sets, it can be a powerful and free tool for SMBs to stop website phishing attacks. The phishing classification model can help analyze the features and make a prediction on whether a website is a phishing attempt or not.

5.2.1 Key Recommendations

Three key policy recommendations emerge from the research process of creating a novel anti-phishing solution to improve the cyber readiness of SMBs in the U.S.

First, **SMBs should consider cloud solutions to deploy the phishing classification model**. As cloud services become more affordable, SMBs can safely deploy the machine learning-based phishing classification model on the cloud as an API, integrating the model as a browser plugin as a scalable solution versus a more costly on-premises solution (Alani & Tawfik, 2022). More resources and attention are needed in developing, testing, and deploying cloud solutions as a cost-effective cybersecurity intervention.

Second, **cybersecurity data on cyberattacks need to be framed as a public good with commensurate public attention**. This praxis found that there were minimal data and detection tools focused on SMBs and cybersecurity defense. The lack of data is a clear limitation to developing evidence-based solutions for SMBs to mitigate and defend against phishing attacks. Currently, there is no requirement for SMBs to report cyberattacks or phishing attacks, inhibiting transparency around cyber incidents. Therefore, a public database where SMBs are incentivized or mandated to report their

cyberattacks reporting requirements should be designed and implemented in the near future.

Lastly, **government and legislation can help provide subsidy for testing** and **verification of cyber readiness** would enhance SMBs' awareness of and engagement with cybersecurity, ultimately enabling robust system design and encouraging collaborations with the public sector around data-sharing and testing mechanisms.

5.3 Contributions to Body of Knowledge

This praxis and the phishing classification model (PCM) contribute to the body of knowledge by the following:

First, this praxis applies ensemble feature selection with filter, Random Forest embedded, and SHAP methods using multiple publicly available phishing data sets. The extraction of key URL-based features from Data Set #1 was essential for identifying an optimal feature set used for phishing website classification.

Second, this praxis is the ***first*** to build a URL parser using significant and important features from a well-known phishing data set (Mendeley 2020) and apply it to the Phishing URLs data set provided by Kaggle. These key learned features were validated by an RF model while achieving greater than 90% accuracy. The research also tested, validated, and contributed to two new and effective features: the ratio of domain length to URL length, and the number of non-alphanumeric in the URL.

Third, this praxis is the ***first*** to validate a phishing classification model using only five-predetermined features and a combination of filter, SHAP, and embedded methods. This enhanced interpretability and helped identify an optimal subset of features that can be used as performance benchmarking for existing and forthcoming phishing data sets.

Fourth, the praxis provides a new phishing data set that can be used for benchmarking classification models and increase the transparency of this study. Machine learning security researchers and practitioners can investigate new hypotheses and build classification models using the URL parser and the presented data sets to extract the key features for phishing detection.

5.4 Recommendations for Future Research

As threat actors' tactics, techniques, and procedures (TTPs) evolve, phishing attacks increase in volume and sophistication. Thus, feature selection and phishing classification will remain a challenge. Future research should explore other feature selection methods and machine learning methods (e.g., ensemble and deep learning).

Researchers should consider integrating other and larger publicly available data sets to improve the robustness of the phishing classification model. While the phishing data sets have large enough sample sizes, more data may improve the machine learning models' performance and the generalization of the phishing classification model. As cyber adversaries continue to adapt, the research community should consider other data sources for phishing and legitimate links.

While this research focused on URL-based features due to their interoperability and explainable machine learning algorithms (e.g., explainable AI) with other public data sets, researchers can consider improving phishing classification using deep learning and other feature selection methods. Additionally, the research community should look beyond phishing website attacks, such as phishing attacks on mobile devices. As companies rely more on mobile devices, the number of companies that suffered a compromise involving a mobile device has increased by 45% (Verizon, 2022b).

Another area for future work would be applying the methodology in this praxis to phishing-based ransomware attacks. For example, researchers could develop phishing detection solutions to specifically mitigate ransomware's impacts.

References

Aaron, G., Chapin, L., Piscitello, D., & Strutt, C. (2022). Phishing Landscape 2022: An Annual Study of the Scope and Distribution of Phishing. Retrieved June 1, 2023, from Apwg.org website: https://apwg.org/trendsreports/.

Alani, M. M., & Tawfik, H. (2022). PhishNot: A Cloud-Based Machine-Learning Approach to Phishing URL Detection. Computer Networks, 218, 109407.

Alavi, R., Islam, S., & Mouratidis, H. (2016). An information security risk-driven investment model for analyzing human factors. Information & Computer Security, 24(2), 205-227.

Aljabri, M., & Mirza, S. (2022). Phishing attacks detection using machine learning and deep learning models. Paper presented at the 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), 175-180.

APWG. (2022). Retrieved February 28, 2023, from Apwg.org website: https://apwg.org/interisle-study-shows-61-increase-in-phishing-attacks-more-brands-targeted-and-257-increase-in-cryptocurrency-phishing/

APWG. (2023). Phishing Activity Trends Report, 4th Quarter, 2022. Retrieved June 1, 2023, from Apwg.org website: https://apwg.org/trendsreports/.

Berners-Lee, T., Fielding, R., & Masinter, L. (2005). Uniform resource identifier (URI): Generic syntax (No. rfc3986). Retrieved from https://www.rfc-editor.org/rfc/rfc3986?spm=a2c4g.11186623.0.0.549927b5NRn8tS.

Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: A systematic literature review. Knowledge and Information Systems, 64(6), 1457-1500.

Chen, C. W., Tsai, Y. H., Chang, F. R., & Lin, W. C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Systems, 37(5), e12553.

Chidukwani, A., Zander, S., & Koutsakis, P. (2022). A survey on the cybersecurity of small-to-medium businesses: Challenges, research focus and recommendations. IEEE Access, 10, 1. doi:10.1109/ACCESS.2022.3197899

Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S. C., & Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences, 484, 153-166. doi:10.1016/j.ins.2019.01.064

Cloudflare (2023). "What is a phishing attack?" Retrieved February 22, 2023, from Cloudflare website: https://www.cloudflare.com/learning/access-management/phishing-attack/

Do, N. Q., Selamat, A., Krejcar, O., Yokoi, T., & Fujita, H. (2021). Phishing webpage classification via deep learning-based algorithms: An empirical study. Applied Sciences, 11(19), 9210.

FBI. (2023). Internet Crime Report 2022. Retrieved May 1, 2023, from ic3.gov website: https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

Felton, J. (2021). Cyber Resilience of Small Business Owners (Order No. 28318819). Available from ProQuest Dissertations & Theses Global. (2504819573). https://gwlaw.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/cyber-resilience-small-business-owners/docview/2504819573/se-2

Fripp, C. (2022, December 21). AI chatbot ChatGPT is being used for phishing attacks - Here's how. Retrieved January 22, 2023, from Komando.com website: https://www.komando.com/security-privacy/ai-chatbot-phishing/870058/.

GeeksforGeeks. (2021). Types of Phishing Attacks and How to Identify them. GeeksforGeeks. https://www.geeksforgeeks.org/types-of-phishing-attacks-and-how-to-identify-them/

Gupta, A. (2020, October 10). Feature Selection Techniques in Machine Learning (Updated 2023). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

Heidt, M., Gerlach, J. P., & Buxmann, P. (2019). Investigating the security divide between SME and large companies: How SME characteristics influence organizational IT security investments. Information Systems Frontiers, 21, 1285-1305.

Jain, A. K., & Gupta, B. B. (2017). Phishing detection: Analysis of visual similarity based approaches. Security and Communication Networks, 2017

Jakobsson, M., & Myers, S. (2006). Phishing and countermeasures: Understanding the increasing problem of electronic identity theft John Wiley & Sons.

Joshi, A., Lloyd, L., Westin, P., & Seethapathy, S. (2019). Using lexical features for malicious URL detection--a machine learning approach. arXiv Preprint arXiv:1910.06277,

Kalaharsha, P., & Mehtre, B. M. (2021). Detecting phishing sites--an overview. arXiv Preprint arXiv:2103.12739,

Kara, I., Ok, M., & Ozaday, A. (2022). Characteristics of understanding URLs and domain names features: The detection of phishing websites with machine learning methods. IEEE Access, 10, 124420-124428.

Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of statistical software, 28, 1-26.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... & Team, R. C. (2020). Package 'caret'. *The R Journal*, *223*(7).

La Trobe University. (2020, June 16). Datasets for Researchers. Retrieved June 1, 2023, from https://latrobe.libguides.com/datasets.

Lin, R. (2022, September 11). Explainable AI with SHAP — Income Prediction Example. Retrieved February 9, 2023, from Medium website: https://reneelin2019.medium.com/explainable-ai-with-shap-income-prediction-example-3050c19a261b

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S. I. (2020). From local explanations to a global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67.

Martin, J. (2022, March 9). How phishing can ruin the good name of an online brand. Reputation X. https://blog.reputationx.com/guest/whats-phishing

Martinez-Taboada, F., & Redondo, J. I. (2020). The SIESTA (SEAAV Integrated evaluation sedation tool for anesthesia) project: Initial development of a multifactorial sedation assessment tool for dogs. PLoS One, 15(4), e0230799.

Nguyen, L. A. T., To, B. L., Nguyen, H. K., & Nguyen, M. H. (2014). A novel approach for phishing detection using URL-based heuristic. Paper presented at the 2014 International Conference on Computing, Management and Telecommunications (ComManTel), 298-303.

Odeh, A., Keshta, I., Abdelfattah, E. (2021). Machine Learning Techniques for detection of website phishing: A review for promises and challenges. Paper presented at the - 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 813. doi:10.1109/CCWC51732.2021.9375997

Pham, T. (2022). Phishing Website Detection Dataset. [data set]. Mendeley. https://doi.org/10.17632/KVPKC4J658.1

Proofpoint. (2023). Retrieved February 27, 2023, from Twitter website: https://twitter.com/proofpoint/status/1622567081307697155

Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. Neural Computing and Applications, 31, 3851-3873.

Rosenthal, M. (2022). Phishing Statistics. Retrieved February 27, 2023, from Tessian website: https://www.tessian.com/blog/phishing-statistics-2020/#:~:text=In%202021%20Tessian%20research%20found%20that%20employees%20receive,majority%20of%20which%20were%20part%20of%20phishing%20campaigns.

Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. Journal of King Saud University - Computer and Information Sciences, doi:10.1016/j.jksuci.2023.01.004

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. SN Computer Science, 2(3), 160.

SBA. (2023). Strengthen Your Cybersecurity. Retrieved January 13, 2023, from Strengthen your cybersecurity website: https://www.sba.gov/business-guide/manage-your-business/strengthen-your-cybersecurity#section-header-4

SBA Office of Advocacy. (2022). Small Business Facts: Small Business Innovation Measured by Patenting Activity. Retrieved February 21, 2023, from SBA's Office of Advocacy website: https://advocacy.sba.gov/2022/09/13/small-business-facts-small-business-innovation-measured-by-patenting-activity/

Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. arXiv Preprint arXiv:2009.11116,

Shirazi, H., Bezawada, B., & Ray, I. (2018). " Kn0w thy Doma1n name" unbiased phishing detection using domain name based features. Paper presented at the Proceedings of the 23rd ACM on Symposium on Access Control Models and Technologies, 69-75.

Silic, M., & Lowry, P. (2020). Using design-science based gamification to improve organizational security training and compliance. Journal of Management Information Systems, 37(1), 129-161.

Silva, C. M. R. d., Feitosa, E. L., & Garcia, V. C. (2020). Heuristic-based strategy for phishing prediction: A survey of URL-based approach. Computers & Security, 88, 101613. doi:10.1016/j.cose.2019.101613

Sloan, J. (2020). Phishing mitigation for small and medium businesses.

Sumner, A., Yuan, X., Anwar, M., & McBride, M. (2022). Examining factors impacting the effectiveness of anti-phishing trainings. Journal of Computer Information Systems, 62(5), 975-997.

Tang, L., & Mahmoud, Q. H. (2021). A survey of machine learning-based solutions for phishing website detection. Machine Learning and Knowledge Extraction, 3(3), 672-694.

Tiwari, T. (2022). Phishing Site URLs, Version 4. Retrieved August 20, 2022, from https://www.kaggle.com/datasets/taruntiwarihp/phishing-site-urls

U.S. Census Bureau. (2023). 2020 SUSB Annual Data Tables by Establishment Industry. Census.gov. https://www.census.gov/data/tables/2020/econ/susb/2020-susb-annual.html

Verizon. (2022a). Data Breach Investigations Report 2022. Retrieved January 30, 2023, from Verizon Business website: https://www.verizon.com/business/resources/reports/dbir/

Verizon. (2022b). Mobile Security Index 2022. Verizon Business. Retrieved May 4, 2023, from https://www.verizon.com/business/resources/reports/mobile-security-index

Vijayalakshmi, M., Mercy Shalinie, S., Yang, M. H., & U, R. M. (2020). Web phishing detection techniques: A survey on the state-of-the-art, taxonomy and future directions. Iet Networks, 9(5), 235-246.

Vrbančič, G., Fister, I., & Podgorelec, V. (2020). Data sets for phishing websites detection. Data in Brief, 33, 106438. doi:10.1016/j.dib.2020.106438

Wei, Y., & Sekiya, Y. (2022). Feature selection approach for phishing detection based on machine learning. Paper presented at the Proceedings of the International Conference on Applied Cybersecurity (ACS) 2021, 61-70.

Whitten, A., & Tygar, J. D. (1999, August). Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In USENIX security symposium (Vol. 348, pp. 169-184).

Zec, M. (2015). Cyber security measures in SME's: A study of IT professionals' organizational cyber security awareness. Linnaeus University, Kalmar.Zugriff Unter Http://Www.Divaportal.Org/Smash/Get/Diva2, 849211

Zimmermann, V., & Renaud, K. (2019). Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. International Journal of Human-Computer Studies, 131, 169-187. doi:10.1016/j.ijhcs.2019.05.005

Appendix A

Table A-1: Data Set, Description, and Sources

ID	Name	Source	Description	Format	Time Span	Num. of records
1	Phishing website detection	Mendeley 2020	These data consist of a collection of legitimate as well as phishing website instances. Each website is represented by the set of features which denote, whether website is legitimate or not. Data can serve as an input for machine learning process.	CSV	2020	88,647
2	Phishing Site URLs from Kaggle	Kaggle 2022	Dataset classifies URLs as "Good" (not malicious and not Phishing) and "Bad" (malicious and Phishing Site)	CSV	2022	507,195
3	Phishing Website Detection Dataset	Mendeley2 022	Data contains both phishing/malign URL and clean/benign URL. The phishing URLs are crawled from phishtank.org while the clean data comes from commoncrawl.org.	CSV	2022	6,000,000
4	My Validation Dataset	Openphish. com; Majestic Million	Building phishing dataset for validation to test relevance, and effectiveness. Dataset consist of phishing/malign URL from openphish.com and clean/benign URL from Majestic Million.	CSV	2023	20,000

Table A-2: Correlation Coefficients and Significant Values for Phishing Features

Full Correlation Coefficient Matrix with "Phishing" Feature to:	Pearson Correlation Coefficient	Full Correlation Coefficient Matrix with "Phishing" Feature to:	Pearson Correlation Coefficient
qty_slash_directory	0.64***	qty_asterisk_params	0.29***
qty_questionmark_directory	0.63***	qty_dollar_params	0.29***
qty_hashtag_directory	0.63***	qty_exclamation_params	0.29***
qty_slash_file	0.63***	tld_present_params	0.29***
qty_questionmark_file	0.63***	qty_comma_params	0.29***
qty_hashtag_file	0.63***	qty_at_params	0.28***
qty_dollar_file	0.63***	qty_questionmark_params	0.28***
qty_at_file	0.63***	qty_plus_params	0.28***
qty_exclamation_file	0.63***	qty_params	0.25***
qty_and_file	0.63***	qty_equal_params	0.24***
qty_tilde_file	0.63***	qty_equal_url	0.22***
qty_equal_file	0.62***	qty_underline_params	0.21***
qty_space_file	0.62***	qty_and_params	0.21***
qty_comma_file	0.62***	qty_dot_params	0.21***
qty_exclamation_directory	0.62***	qty_slash_params	0.21***
qty_comma_directory	0.62***	qty_hyphen_params	0.19***
qty_space_directory	0.62***	qty_tld_url	0.19***
qty_dot_file	0.62***	file_length	0.18***
qty_tilde_directory	0.62***	qty_percent_directory	0.15***
qty_equal_directory	0.62***	qty_underline_url	0.15***
qty_plus_file	0.61***	qty_percent_file	0.15***
qty_dollar_directory	0.61***	qty_vowels_domain	0.15***
qty_plus_directory	0.61***	qty_hyphen_url	0.14***
qty_slash_url	**0.60*** **	qty_dot_url	0.14***
qty_dot_directory	**0.57*** **	ttl_hostname	0.14***
qty_and_directory	0.57***	qty_hyphen_domain	0.11***
qty_asterisk_file	0.54***	qty_percent_params	0.09***
qty_at_directory	0.54***	domain_length	0.08***
qty_asterisk_directory	0.50***	asn_ip	0.08***
qty_underline_file	0.48***	time_response	0.06***
qty_underline_directory	0.47***	domain_spf	0.01***
directory_length	**0.43*** **	qty_ip_resolved	-0.01***
length_url	**0.38*** **	qty_redirects	-0.03***
qty_hyphen_file	0.33***	tls_ssl_certificate	-0.04***
qty_hyphen_directory	0.30***	qty_mx_servers	-0.05***
qty_hashtag_params	0.29***	qty_nameservers	-0.14***
qty_tilde_params	0.29***	qty_dot_domain	-0.24***
qty_space_params	0.29***	time_domain_activation	-0.37***

Note: Features above are significant with p-values <0.001

code	p-value
***	[0, 0.001]

Eigenanalysis of the Correlation Matrix

Eigenvalue	30.729	14.078	2.851	2.438	1.975	1.850	1.736	1.499	1.404	1.235	1.100
Proportion	0.404	0.185	0.038	0.032	0.026	0.024	0.023	0.020	0.018	0.016	0.014
Cumulative	0.404	0.590	0.627	0.659	0.685	0.709	0.732	0.752	0.771	0.787	0.801

Eigenvalue	1.055	1.022	0.978	0.933	0.883	0.838	0.831	0.803	0.747	0.697	0.681
Proportion	0.014	0.013	0.013	0.012	0.012	0.011	0.011	0.011	0.010	0.009	0.009
Cumulative	0.815	0.829	0.841	0.854	0.865	0.876	0.887	0.898	0.908	0.917	0.926

Eigenvalue	0.644	0.634	0.563	0.552	0.480	0.439	0.356	0.333	0.234	0.188	0.132
Proportion	0.008	0.008	0.007	0.007	0.006	0.006	0.005	0.004	0.003	0.002	0.002
Cumulative	0.934	0.943	0.950	0.957	0.964	0.969	0.974	0.978	0.982	0.984	0.986

Eigenvalue	0.101	0.093	0.093	0.085	0.080	0.070	0.067	0.066	0.048	0.044	0.041
Proportion	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Cumulative	0.987	0.988	0.990	0.991	0.992	0.993	0.993	0.994	0.995	0.996	0.996

Eigenvalue	0.039	0.030	0.029	0.025	0.024	0.021	0.020	0.018	0.015	0.013	0.011
Proportion	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	0.997	0.997	0.997	0.998	0.998	0.998	0.999	0.999	0.999	0.999	0.999

Eigenvalue	0.010	0.008	0.006	0.005	0.004	0.003	0.003	0.003	0.002	0.002	0.002
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Eigenvalue	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure A-1: Eigenanalysis of the Correlation Matrix

Table A-3: Phishing URLs Data Set – 27 Features Parsed

URL	qty_slash	qty_dot_u	qty_equal	qty_and_	qty_hyper	qty_at_ur	qty_digits	qty_nonal	length_ur	qty_badtl	domain_n	domain_le	qty_vowe	ratio_dom	ratio_digi	directory_	directory_name	qty_slash	qty_hyphi	qty_unde	qty_equal	qty_at_dir	qty_excla	qty_dot_d	qty_quest	qty_digits	phishing
youtube.com/watch?v=Qfx3hmr	1	31	1	0	0	0	1	5	31	0	youtube.c	11	5	0.354839	0.2	7	youtube.com/watch	1	0	0	0	0	0	17	0	0	0
thecanadianencyclopedia.com/i	1	67	2	1	0	0	8	7	67	0	thecanadi	27	11	0.402985	1.142857	23	thecanadianencyclopedia.com	1	0	0	0	0	0	37	0	0	0
phideltaphi.org/?page=About	1	27	1	0	0	0	0	4	27	0	phideltap	15	5	0.555556	0	11	phideltaphi.org/	1	0	0	0	0	0	16	0	0	0
www.acm.org/crossroads/doc/ir	4	59	0	0	0	0	8	59	0	www.acm	11	2	0.186441	0	3	www.acm.org/crossroads/doc		4	0	0	0	0	0	48	0	0	0
apsva.us/longbranch/	2	20	0	0	0	0	0	3	20	0	apsva.us	8	3	0.4	0	5	apsva.us/longbranch/	2	0	0	0	0	0	20	0	0	0
www.getcruising.com/cruising/c	2	37	0	0	0	0	0	5	37	2	www.getc	19	5	0.513514	0	3	www.getcruising.com/cruising	2	0	0	0	0	0	37	0	0	0
www.linux-mag.com/id/3308	2	25	0	0	1	0	4	5	25	0	www.linu	17	4	0.68	0.8	3	www.linux-mag.com/id/3308	2	1	0	0	0	0	25	0	4	0
autographwarehouse.com/Quek	1	65	0	0	0	0	0	3	65	0	autograph	22	10	0.338462	0	18	autographwarehouse.com/Qu	1	0	0	0	0	0	65	0	0	0
porneskimo.com/dino%20bravo	1	31	0	0	0	0	2	4	31	0	porneskin	14	5	0.451613	0.5	10	porneskimo.com/dino%20bra	1	0	0	0	0	0	31	0	2	0
umsl.edu/~whmc/guides/whm0	3	33	0	0	0	0	4	6	33	0	umsl.edu	8	3	0.242424	0.666667	4	umsl.edu/~whmc/guides/whr	3	0	0	0	0	0	33	0	4	0
hollyeats.com/ArthurBryants.ht	1	31	0	0	0	0	0	3	31	0	hollyeats.	13	4	0.419355	0	9	hollyeats.com/ArthurBryants.l	1	0	0	0	0	0	31	0	0	0
lead411.com/James_McGill_149!	1	36	0	0	0	0	9	5	36	0	lead411.c	11	3	0.305556	1.8	7	lead411.com/James_McGill_1	1	0	2	0	0	0	36	0	9	0
cmhg.gc.ca/cmh/page-84-eng.as	2	30	0	0	2	0	2	7	30	0	cmhg.gc.c	10	1	0.333333	0.285714	4	cmhg.gc.ca/cmh/page-84-eng.	2	2	0	0	0	0	30	0	2	0
spokeo.com/Nicole+Brunner	1	25	0	0	0	0	0	3	25	1	spokeo.cc	10	4	0.4	0	6	spokeo.com/Nicole+Brunner	1	0	0	0	0	0	25	0	0	0
section1fieldhockey.com/archi	1	38	0	0	0	0	1	3	38	0	section1fi	23	8	0.605263	0.333333	19	section1fieldhockey.com/arc	1	0	0	0	0	0	38	0	1	0
artcyclopedia.com/artists/tousig	2	48	0	0	0	0	0	5	48	0	artcyclope	17	6	0.354167	0	13	artcyclopedia.com/artists/tou	2	0	1	0	0	0	48	0	0	0
legacy.com/obituaries/telegram	3	108	3	2	4	0	13	15	108	0	legacy.cor	10	3	0.092593	0.866667	6	legacy.com/obituaries/telegr	3	1	0	0	0	0	52	0	0	0
freespace.virgin.net/gerard.mdi	2	50	0	0	0	0	1	6	50	0	freespace.	20	7	0.4	0.166667	9	freespace.virgin.net/gerard.m	2	0	0	0	0	0	50	0	1	0
www.initiative.ru/English/index	2	35	0	0	0	0	0	5	35	1	www.initi	17	7	0.485714	0	3	www.initiative.ru/English/ind	2	0	0	0	0	0	35	0	0	0

Table A-4: Estimated Cost for SMB Phishing Detection System

	Option 1: In-House using Open Source System	Option 2: Managed Service Provider
Phishing Detection System Development	Contract 1x Security Architect (80 hours – 2 weeks) = $200 / hr x 80 = $1,600 one-time fee	$1,200 - $2,280 per year or $100- $200 per month (pending the number of devices and features needed)

	Pros	Cons
Option 1: In-House using open-source tools	Full control of infrastructure, security, fast access to data, and can contract as needed	Need upfront capital for resources; require expertise; deployment and maintenance costs
Option 2: Managed Service Provider	Manage service, provide expertise for support, and reoccurring costs	Reliant on the vendor, responsiveness, less control of data
***Contracting Support**	Hire as Needed	One-time expense

Average Costs for MSP Pricing Models (cpcyber.com)

Pricing Model Type	Cost Per User/Device/Month	Average Monthly Expense (Estimated for 100 employees)
Per- Data Usage Pricing	$10 to $500	$1,000 to $50,000
Per- Device Pricing	$75 to $250	$7,500 to $25,000
Per- User Pricing	$75 to $250	$7,500 to $25,000
Cloud-Based Pricing	$125 to $300	$12,500 to $30,000
Tiered Pricing	$30 to $150	$3,000 to $15,000

Costs	Google Workspace with Phishing Protection	Average Monthly Expense (Estimated for 100 employees)
Per- User Pricing	$6 to $18	$600 to $1,800