DOI: 10.15849/IJASCA.250730.16

Int. J. Advance Soft Compu. Appl, Vol. 17, No. 2, July 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523

A Comprehensive Machine Learning Approach for Email and URL Threat Detection Using Feature Importance Analysis

Mohammad Kharabsheh[1], Shadi AlZu'bi[2], Ali Alsarhan[3], Ala'a Mughaith[4], Nadera Aljawabreh[5], and Mohammad Alabdullatif[6]

[1] Department of Computer Information Systems, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University, Zarqa, Jordan.
e-mail: mohkh86@hu.edu.jo.
[2]Computer Science Department, Al Zaytoonah University of Jordan
e-mail: smalzubi@zuj.edu.jo.
[3]Department of Computer Information Systems, The University of Maryland, Baltimore County, Baltimore, USA.
e-mail: ali.alsarhan@umbc.edu.
[4]Department of Computer Science, GUST Engineering and Applied Innovation Research Center(GEAR), Gulf University of Science and Technology, Kuwait.
e-mail: mughaid.a@gust.edu.kw.
[5]Department of Information Technology, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University, Zarqa, Jordan.
e-mail: naderam@hu.edu.jo.
[6]College of Computer Science and Information Technology, King Faisal University, Saudi Arabia
e-mail: maabdullatif@kfu.edu.sa.

Abstract

Phishing is the most prevalent form of cybercrime, where individuals are convinced to disclose sensitive details like account IDs, passwords, and banking information. These cyberattacks are often initiated through emails, instant messaging, and phone calls. The primary concern today revolves around the security of devices, computers, and software. This study presents the development of a website designed to scan incoming emails and attachments for potential viruses and security threats. This website includes validation attachment scanning, URL scanning, and IP address scanning. Integration with the VirusTotal database will be carried out to assess the safety of websites. Furthermore, the study incorporates machine learning algorithms to enhance phishing detection, ultimately mitigating risks and occurrences. The dataset utilized comprises diverse sources containing both regular and phishing emails, along with numerous attributes for identifying malicious emails and harmful URL links, some of which are sourced from VirusTotal. The outcomes of the experiments reveal promising levels of accuracy in identifying phishing attacks, underscoring the efficiency of machine learning as a vital component in enhancing email security. The study also

Received 25 March 2025; Accepted 6 July 2025

addresses the obstacles and constraints faced by the proposed models, highlighting the evolving nature of phishing strategies and the necessity for continual model adaptation.

Keywords: *Optimization Algorithms, Cybersecurity, Real-word Problems, Feature Attribution, Machine Learning, Predictive Modeling.*

1 Introduction

Phishing remains a major threat worldwide at present. Numerous businesses have suffered significant financial losses due to phishing emails. The success of this tactic hinges on persuading individuals to share sensitive details, such as login credentials or financial data. These cyber-attacks are often carried out via emails, instant messaging, and phone calls, with the attacker pretending to be a reliable source. This study will focus specifically on email. phishing schemes. Phishing attacks utilize a combination of technology and social manipulation to gather information about the target's identity and accounts. By undermining trust in the online realm, phishing has the capacity to negatively impact e-commerce ventures. Despite the accessibility of various efficient phishing email recognition techniques, companies and customers still face significant financial losses from phishing emails each year. Recent findings indicate that phishing attacks have more than doubled since early 2020 [5]. According to the Anti-Phishing Working Group (APWG), they're now seeing between 68,000 and 94,000 of these attacks each month. This alarming trend highlights just how prevalent and dangerous phishing has become in our online world. In this study, we created a web app and set up a system to check if attachments are secure. We also used a dataset called the" Web Page Phishing Detection Dataset" to help us analyze and spot any harmful content in URLs or files [19]. There are various ways to spot and stop phishing attempts, and one of the most promising approaches is using machine learning. Over the last two decades, machine learning has transformed from a theoretical concept into a powerful tool that is now commonly applied in various fields. It became essential for many applications in artificial intelligence, such as controlling robots, recognizing voices, processing speech, and understanding natural language [20].

2 Literature Review

Machine learning algorithms are great at recognizing patterns in data and making predictions based on those patterns, which makes them really useful for detecting phishing attacks. There are many different machine learning algorithms, each designed to handle different kinds of data challenges. It is important to note that no single algorithm works perfectly for every problem. The choice of algorithm depends on what specific issue you are dealing with, how many variables are involved, and which model fits best [21,24]. In this paper, we will use a range of algorithms to enhance our ability to detect phishing attempts, which will help reduce the risks and frequency of these attacks. By filtering out and blocking phishing URLs, we can reduce how often phishing occurs. Machine learning algorithms fall into two main types: supervised and unsupervised learning. In supervised learning, we train the algorithm by giving it both the input data and the correct output, which helps it learn how to make accurate predictions on new data. On the other hand, unsupervised learning doesn't require this kind of training. Instead, it uses techniques like deep learning to explore the data and find patterns on its own. Unsupervised learning is often used for more complex problems compared to supervised learning [22]. Here are a few examples of how machine learning can be used to spot phishing attacks effectively. Here are some powerful examples.

establishing how machine learning can firmly distinguish phishing attacks.

- Email classification: Machine learning can be used to classify emails and determine if they are legitimate or phishing attacks by analyzing details such as sender information, email content, and links that may lead to harmful sites [3-4].
- Webpage analysis: Machine learning can also examine the structure and content of web pages that clearly indicate phishing attacks, such as fake login forms or misleading language [2], [5].
- URL Analysis: Machine learning algorithms can effectively be trained to classify URLs (web addresses) as either fraudulent or legitimate by accessing data such as the domain name and common phishing keywords [1], [3]. Since URLs are essential in accessing online resources, they are often targeted in phishing attacks, especially through emails. Most of phishing efforts are conducted through emails which include counterfeit URLs in the body text [23]. Security has become a major concern as we become more connected to the Internet via computers, phones, and even household appliances.

The motivation behind this study is to use machine learning techniques to detect phishing emails. This study critically assesses a variety of relatively recent methods and gives suggestions for how additional improvement might be gained. Therefore, in this study, we will quantify and qualify the phishing email features to prevent and mitigate the risk of phishing emails. Phishing is the most common type of cybercrime in which victims are persuaded to divulge sensitive information, such as account IDs and passwords. With so much effort being put into phishing email detection, No collection of features has been proven to be the best for detecting phishing. [24], This study attempts to answer the following questions:

What is the most effective classification algorithm for phishing detection?

Which features are most important for phishing detection?

How can we build an improved model for detecting phishing attacks?

This study's general objective is to use machine learning techniques to detect phishing emails.

The Specific objectives of this study are as follows:

- **Integrate both email content and URL analysis** to maximize phishing detection accuracy
- **Prioritize key predictive features** identified in feature importance analysis.
- **Leverage ensemble learning techniques**, such as Random Forest, or explore advanced deep learning architectures

2.1 Malignant URL Categorization

One of the most significant challenges with most existing studies on defining the term" malicious URL" is the lack of a clear definition. As a result, classifying URLs would make this process more efficient. This could enhance the experimental phase and serve as the initial step toward constructing a comprehensive machine-learning classifier. In recent years, great efforts have been made by academics investigating harmful URLs. However, most studies have been unable to provide a clear definition for the term" malicious URL." They gathered phishing and spam URLs during the studies and categorized them as harmful [25].

2.2 Phishing URLs

Phishing URLs are web addresses created to trick people into believing they are legitimate websites, intending to collect sensitive information such as user names, passwords, credit card numbers, or other personal information. Phishing is the most common sort of cybercrime in which victims are convinced to reveal critical information. Despite efforts to guide users to recognize phishing websites, most internet users were unable to do so. A phishing detection system has been developed to detect infiltration attempts by analyzing nine lexical criteria [26]. Phishing URLs tend to expose multiple warning indications, including:

- Fake Login Pages: These mimic login screens from well-known websites to deceive users into providing their credentials. Unsolicited Emails or Messages: Phishing links are frequently delivered by unexpected emails, texts, or social media posts.
- Lack of HTTPS: HTTPS is often used by trustworthy websites to encrypt communications.
- Mismatched or irrelevant URLs: Phishing links may direct users to URLs unrelated to the claimed source, such as an email from a bank that leads to an unrelated website.
- Misspelled or Altered Domain Names: Phishing URLs often utilize IP addresses instead of domain names, which is unusual for trustworthy websites.

The phishing detection system is deployed as a modular web application with a React.js frontend and a Python-based backend (using Flask or FastAPI). Users can submit URLs or file attachments for real-time analysis via RESTful APIs. The backend connects to a PostgreSQL and MongoDB database for structured and semi-structured data storage. The system ensures low latency, with average response times of 300–1200 ms, and is hosted on scalable cloud infrastructure using Docker containers. For continuous improvement, the platform includes a feedback loop where users can report detection accuracy. These inputs, along with new phishing data and threat intelligence sources, feed into a scheduled retraining pipeline that updates the machine learning model weekly or biweekly. Monitoring tools detect data or concept drift, ensuring model relevance. The system also integrates with third-party threat databases like VirusTotal to enhance phishing detection.

2.3 Malware URL

A malware URL is a web address linked to the propagation or delivery of malware or harmful software, such as viruses, worms, trojans, ransomware, and spyware. In contrast to phishing strategies, which focus on deceiving users, cybercriminals utilize these malicious URLs to distribute malware across several systems by exploiting browser vulnerabilities. Researchers have investigated approaches for detecting malware activities, but no definitive answers have emerged [27-29]. Malware URLs frequently link to harmful file downloads posing as legitimate documents, multimedia, or software. To protect against these dangers, keep firewalls, antivirus, and security software up to date, practice safe browsing, and exercise caution when clicking on unsolicited links. Secure browsing and regular software updates are vital.

2.4 Spam links

Spam links are web addresses associated with unwanted, often irrelevant, or inappropriate information. These connections are usually distributed across different platforms, including email, social media, instant messaging, and online forums. Spam links tend to promote products or services, send visitors to specific websites, or carry out malicious activities. Gao Hongyu, et al. in [30] investigated current spamming strategies and used machine-learning technologies to detect spam URLs. To avoid spam links, users should exercise caution when clicking on links in emails or unexpected communications, use the spam filters provided by email services, and avoid sharing private or sensitive information via unknown connections. Updating your antivirus and anti-malware software constantly is essential. Report spam messages and links to the appropriate platforms or authorities. Inform the proper authorities or platforms about spam messages or links. Being cautious and maintaining safe browsing habits can help reduce the possibility of falling victim to spam and other types of online attacks.

2.5 Phishing Emails

Phishing emails come in various forms, and scammers always evolve their methods to trick victims into disclosing personal information or performing actions that might threaten security. Some common types of phishing emails include: Spear phishing: targeted emails crafted for individuals or organizations, using personal information such as names or job titles to increase trust. CEO Fraud/company Email Compromise (BEC): Attackers imitate high-ranking executives and request urgent wire transfers or sensitive information under the premise of an urgent business issue. Vishing (Voice Phishing): Scammers use phone calls instead of emails for collecting personal information, often encouraged by a phishing email directing victims to call a specific number. Clone Phishing: A valid email is duplicated with harmful links or attachments, delivered as a” resend” or update. Dropbox/Google Drive Phishing: Links appear to take users to data stored on cloud platforms but instead lead to phishing sites designed to steal login credentials. Invoice or Payment Phishing: Attackers impersonate as service providers or suppliers, claiming issues with payments or invoices to make you take immediate action or click on malicious links. Social Media Phishing: Fake accounts or phony messages on social platforms aim to trick users into clicking on malicious links or sharing personal information [32,33]. mobile phishing: Mobile phishing attacks can be categorized based on different attack vectors, including social engineering, mobile applications, malware, social networking platforms, content injection methods, and wireless communication channels [34].

3 Methodology

In this study, Google Cloud’s Vertex AI AutoML was utilized to develop and evaluate a tabular classification model, with the primary goal of analyzing feature importance within a previously unclassified dataset named data-final. The use of Vertex AI AutoML allowed for a highly automated and efficient machine learning workflow, significantly reducing the manual effort typically required for model training and evaluation. At the same time, to ensure a comprehensive and strategic modeling approach, traditional classifiers were also considered in parallel with AutoML. This dual-track method balances the speed and simplicity of AutoML with the flexibility and control offered by custom-built models, particularly useful for more complex or domain-specific scenarios.

The dataset was processed entirely within Google Cloud, where it was cleaned, transformed, and structured for use in a classification pipeline. The target variable was assigned to the column labeled class-label. To ensure robust evaluation, the dataset was split into three parts: 80% for training, 10% for validation, and 10% for testing. Model development and training were executed in the us-central1 region using Google-managed encryption to ensure data security. Vertex AI AutoML’s serverless infrastructure handled the entire end-to-end workflow, from data ingestion and preprocessing to model training, evaluation, and deployment readiness.

A key aspect of AutoML’s pipeline was the use of automated hyperparameter tuning, optimizing the model specifically for the AUC ROC metric to maximize discriminative performance. The training process took approximately 2 hours and 28 minutes. During this process, Shapley sampling was employed to determine feature attribution, providing explainability by identifying which features most significantly influenced model predictions.

The model achieved strong performance across multiple metrics: a Precision Recall AUC (PR AUC) of 0.991, a Receiver Operating Characteristic AUC (ROC AUC) of 0.993, and an F1 score of 0.991 at a decision threshold of 0.5. These results confirmed the model’s high precision and recall, making it suitable for deployment in real-world decision-making contexts. Vertex AI’s debugging features and pipeline monitoring tools ensured seamless execution, with detailed logs and performance metrics available throughout the training lifecycle.

Overall, the integration of Google Cloud’s Vertex AI AutoML into our machine learning workflow demonstrated its effectiveness in automating complex processes, achieving high predictive performance, and extracting valuable insights through advanced feature attribution techniques. The full methodology, including data collection, preprocessing, model training, and evaluation metrics, is illustrated in Figure 1, which provides a visual overview of the entire pipeline.

3.1 Studied Dataset

Our study was conducted among a sample of phishing-email-collection and Phishing-Legitimate datasets. In this study, we consider the features of Phishing email-collection datasets, as shown in Table 1, and the features of Phishing Legitimate datasets, as shown in Table 2.

Table 1: The features of Phishing-email-collection

No.	Features
1	Total Number of Character
2	Vocabulary richness WC
3	Unique Words
4	Account
5	Click
6	Total number of Function
7	Security
8	Bank
9	Risk
10	Information
11	Access
12	Service
13	Recently
14	Credit
15	Suspended
16	Limited
17	Identity
18	Inconvenience
19	Password
20	Social
21	Minutes

Table 2: The features of Phishing-Legitimate

No.	Features
1	PctExtHyperlinks
2	PctExtNullSelfRedirectHy. . .
3	FrequentDomainNameMi. . .
4	PctExtResourceUrls
5	1NumNumericChars
6	ExtMetaScriptLinkRT
7	ExtFavicon
8	PathLevel
9	SubdomainLevel
10	NumDots
11	PctNullSelfRedirectHyper. . .
12	NumSensitiveWords
13	NumDash
14	InsecureForms
15	IframeOrFrame
16	PathLength
17	NumDashInHostname
18	NumUnderscore
19	SubmitInfoToEmail
20	QueryLength
21	NumQueryComponents
22	UrlLength
23	AbnormalExtFormActionR
24	RelativeFormAction
25	HostnameLength
26	NoHttps
27	NumPercent
28	PctExtResourceUrlsRT
29	UrlLengthRT
30	DomainInPaths
31	RandomString
32	MissingTitle
33	SubdomainLevelRT
34	ExtFormAction
35	IpAddress
36	NumAmpersand
37	AbnormalFormAction
38	DomainInSubdomains
39	EmbeddedBrandName

In this paper, the datasets were used as an input for various prediction models based on statistical model (Logistic Regression, LR) and machine learning model (Decision Tree, logistic Regression, Naive Bayes, random Forest and Support Vector Machine). These models were utilized to Phishing detection model, Anomaly detection, Features importance analysis of the phishing-email as we shown in Figure 1.

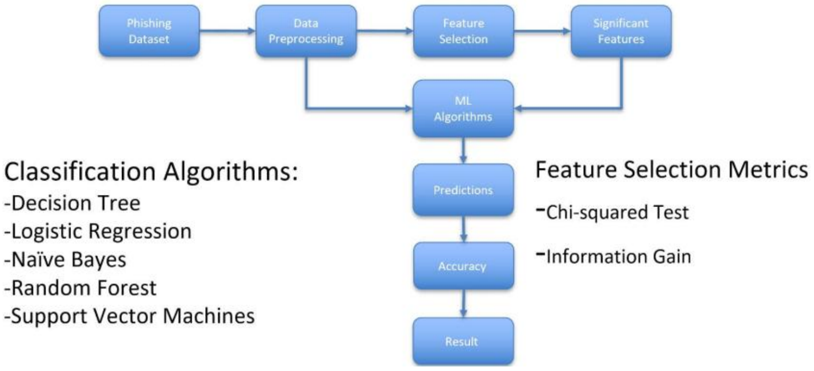


Figure 1: The Major Steps of Our Prediction Model Collection and Phishing-Legitimate.

3.2 Classification Algorithms

In our experiments, we employed the supervised classifiers in which the included corpus is distributed into two groups as a training set and a testing set. The first group is the one that is used to train the developed machine learner. On the other hand, the performance of the learner is computed through the second group. We used the widely popular 10-fold cross-validation [6] technique to obtain both the training and testing sets to get unbiased results, which offered better model performance in our dataset.

There are various classification algorithms that are widely used of decision support systems presented for the healthcare domain and have been used to develop the employed models [7-9], these algorithms are as follows:

- Decision Tree: Decision Tree in particular J.48 algorithm is commonly used to classify different datasets and perform accurate results of classification. J48 algorithm is one of the best machine learning algorithms to investigate the data category continuously. It engages more memory space and reduces the performance and accuracy in classified data. This algorithm creates a binary tree for classification problems. The approach splits the data into ranges using the values of attributes for that item that are recognized in the training set [10],[31].
- Logistic Regression: Logistic Regression is a predictive analysis that estimates the probability of one dependent variable based on one or more independent variables. Logistic Regression is a linear model for categorization rather than regression. This approach uses regression models for classification tasks that models the posterior class probabilities for each of the needed n-classes from the dataset [11].

- Naive Bayes: Naive Bayes allocates the highly expected class when given characteristics are independent of any particular class. Naive Bayes is effective in many fields such as text categorization, and therapeutic diagnosis. This method assumes that all classification factors are independent. It shows great performance in terms of accuracy when it was applied in medical domain studies [12].
- Random Tree: Random tree is an ensemble training method for classification. This method is a set of separate decision trees in which each tree is produced from different samples and subsets of the training data. Random Tree is a supervised learning algorithm that produces many individual learners. It generates a random set of data for creating a decision tree. Random trees deal with both classification and regression problems. Random tree is a set of tree predictors (forest). The classifier gets the input feature vector, classifies it with every tree in the predictors. Random Tree is an active data mining algorithm that is used with large amounts of data. The technique employs several classification trees to a data set and next generates the prediction from all of the correlated trees [13-14].
- Support Vector Machine (SVM): seek to figure out a decision boundary between classes, expanding the margin of the separating line; while one of the drawbacks of this approach is that it can be only applied for bi nary classification [15]. Support Vector Machines (SVM) can construct the optimal separating lone, which increase the distance between the contiguous sample data [16]. This algorithm rises the dimensionality of training instances to achieve differentiable points in one of the dimensions. SVM is very popular since it is efficient in high dimensional spaces and thus provides more accurate results [17].

4 Results

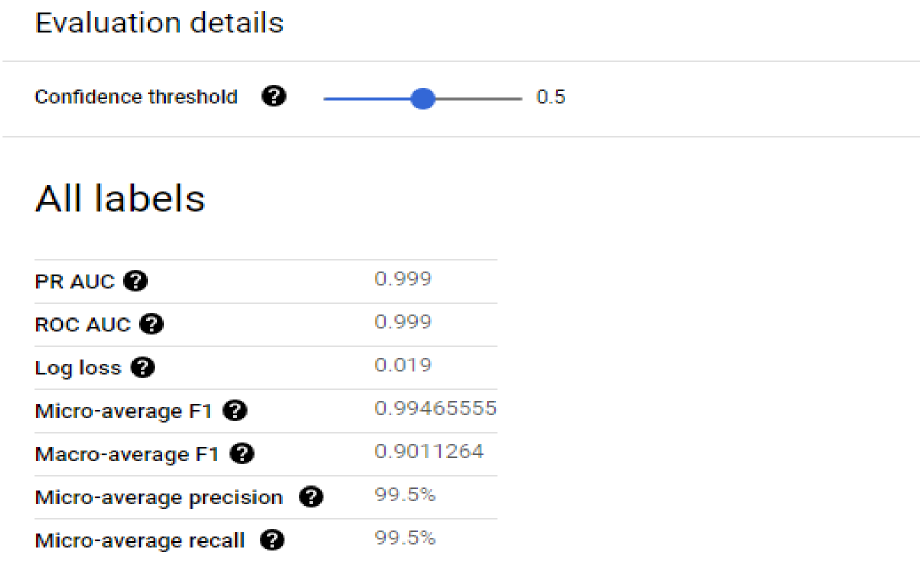
To assess the performance of our phishing detection model, we conducted extensive experiments using multiple classification algorithms and evaluated them using standard performance metrics. This section presents the results, along with an analysis of the classifiers' effectiveness, feature importance, and error analysis.

This study leveraged two primary datasets: a 'Phishing-email-collection' dataset with features related to email content, and a 'Phishing-Legitimate' dataset encompassing URL-based characteristics. The 'Phishing-email-collection' dataset contains a variety of characteristics extracted from email bodies including *text length calculated by Total Number of Characters C, vocabulary 10 richness calculated by W/C, occurrence of words related to Account, Access, Bank, Credit, Click, Identity, Inconvenience, Information, Limited, Minutes, Password, Recently, Risk, Social, Security, Service, Suspended. It also includes some statistics related to the count of Function words/W, and the number of Unique Words which provides insights into textual patterns indicative of phishing attempts, for classifying the Phishing Status. On the other hand, the 'Phishing-Legitimate' dataset focuses on URL attributes, comprising a wide range of lexical and host-based features such as the number of dots (Num Dots), subdomain level (SubdomainLevel), path level (PathLevel), URL length (UrlLength), number of dashes (NumDash), presence of special symbols (At Symbol, TildeSymbol, NumUnderscore, NumPercent, NumHash, NumAmper sand), indicators of secure connection (NoHttps), and the existence of an IP address (IpAddress). These features enable the model to identify malicious URL patterns, detecting fraud.

We acknowledge the importance of incorporating dynamic, real-time data sources to ensure the dataset remains current and reflective of evolving trends, particularly in domains such as phishing or cybersecurity. In response, we plan to incorporate real-time feeds such as those provided by the Anti-Phishing Working Group (APWG) to supplement our existing dataset. This enhancement improves the future model's ability to generalize to recent and emerging threats.

4.1 Evaluation Metrics and Performance

The evaluation details of the **Phishing-email-collection** dataset, including PR AUC, ROC AUC, Log Loss, F1-score, and Precision-Recall, are shown in Figure 2.



Evaluation details

Confidence threshold 0.5

All labels

PR AUC	0.999
ROC AUC	0.999
Log loss	0.019
Micro-average F1	0.99465555
Macro-average F1	0.9011264
Micro-average precision	99.5%
Micro-average recall	99.5%

Figure 2: Evaluation Details of Phishing-email-collection.

Further details of precision-recall and ROC curve analysis for the model are presented in Figure 3.

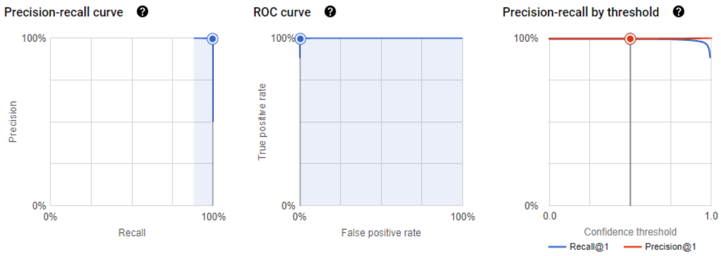
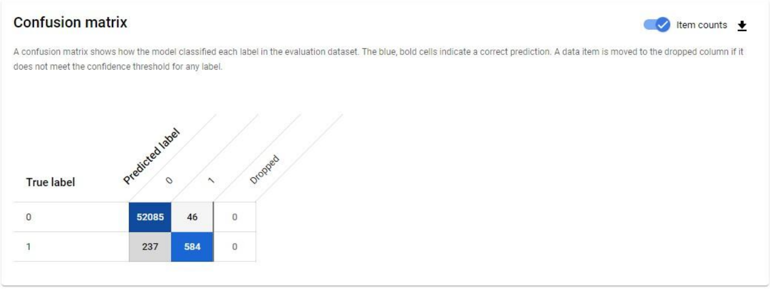


Figure 3: Precision-Recall, ROC Curve, and Threshold-based Precision-Recall for Phishing-email-collection.

4.2 Confusion Matrix Analysis

The confusion matrix in Figure 4 provides a breakdown of the model's classification accuracy for phishing emails, showing the distribution of correctly and incorrectly classified instances.



Confusion matrix

A confusion matrix shows how the model classified each label in the evaluation dataset. The blue, bold cells indicate a correct prediction. A data item is moved to the dropped column if it does not meet the confidence threshold for any label.

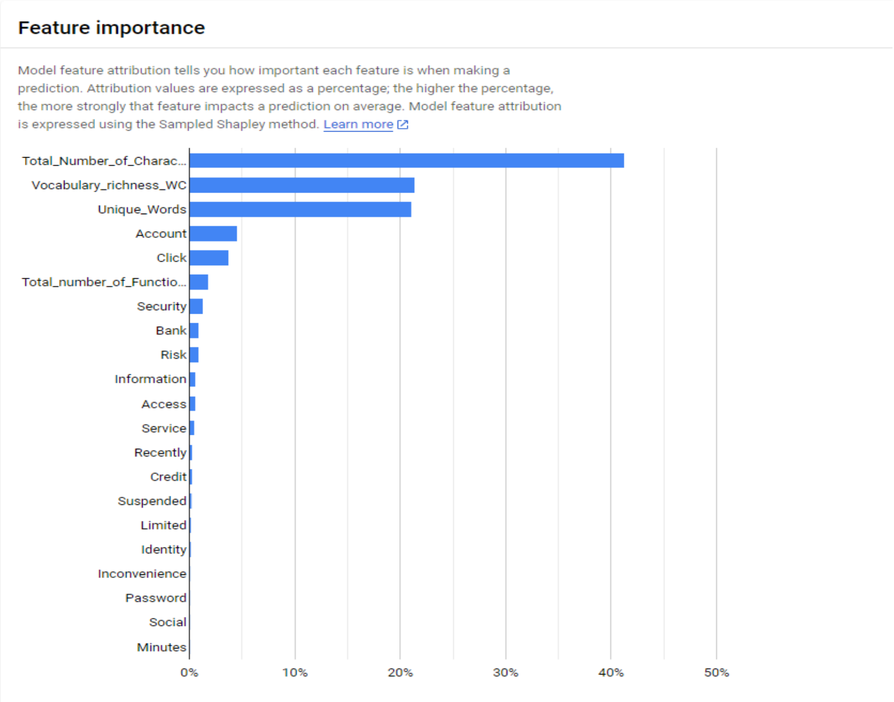
True label	Predicted label 0	1	Dropped
0	52085	46	0
1	237	584	0

Figure 4: Confusion matrix of Phishing-email-collection.

4.3 Feature Importance

Feature selection is a critical step in the machine learning process as it helps reduce model complexity, improve training efficiency, enhance generalization by reducing overfitting, and increase overall model interpretability. While various techniques exist for feature selection, such as filter methods, wrapper approaches, and embedded techniques, this project employed a model-agnostic method based on SHAP (Shapley Additive Explanations) values, specifically using Shapley sampling. This approach was selected for its ability to provide meaningful, interpretable insights into feature importance across a wide range of models, while remaining computationally efficient for large and complex datasets.

The importance of features used for phishing classification is visualized in Figure 5. The most significant features include *Total Number of Characters*, *Vocabulary Richness*, *Unique Words, and Account-related terms.*



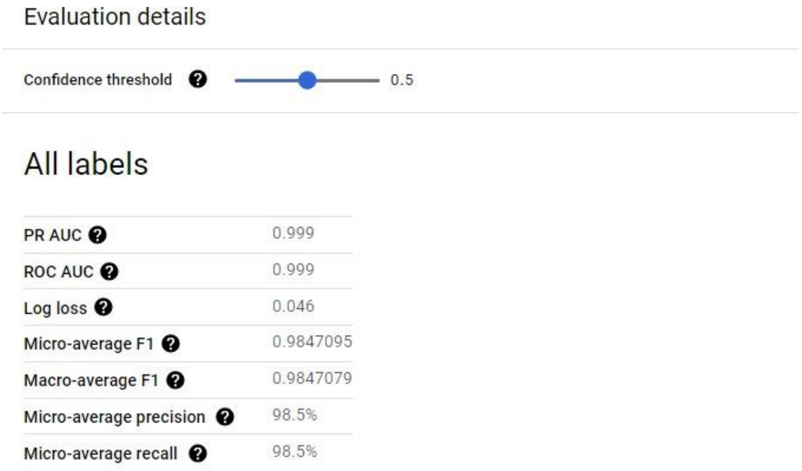
Feature importance

Model feature attribution tells you how important each feature is when making a prediction. Attribution values are expressed as a percentage; the higher the percentage, the more strongly that feature impacts a prediction on average. Model feature attribution is expressed using the Sampled Shapley method. Learn more

Figure 5: Feature Importance of Phishing-email-collection.

4.4 Phishing-Legitimate Dataset Analysis

Similar evaluation details for the Phishing-Legitimate dataset are provided in Figure 6. Figure 7 presents the precision-recall and ROC curve results for the Phishing Legitimate dataset.



Evaluation details

Confidence threshold ❓ 0.5

All labels

PR AUC ❓	0.999
ROC AUC ❓	0.999
Log loss ❓	0.046
Micro-average F1 ❓	0.9847095
Macro-average F1 ❓	0.9847079
Micro-average precision ❓	98.5%
Micro-average recall ❓	98.5%

Figure 6: Evaluation Details of Phishing-Legitimate dataset.

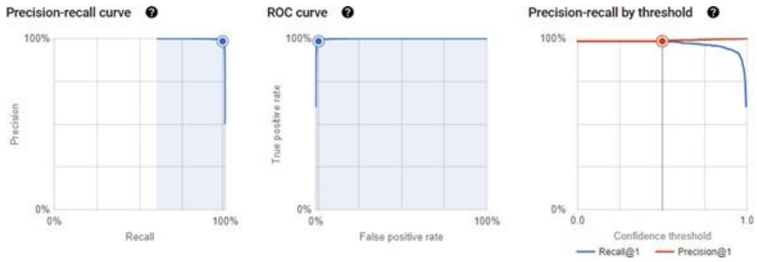
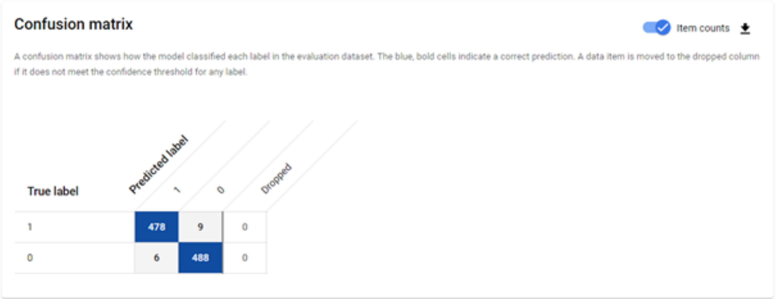


Figure 7: Precision-Recall, ROC Curve, and Threshold-based Precision-Recall for Phishing-Legitimate dataset.

4.5 Confusion Matrix for Phishing-Legitimate Dataset

The confusion matrix for the Phishing-Legitimate dataset is shown in Figure 8.



Confusion matrix

Item counts

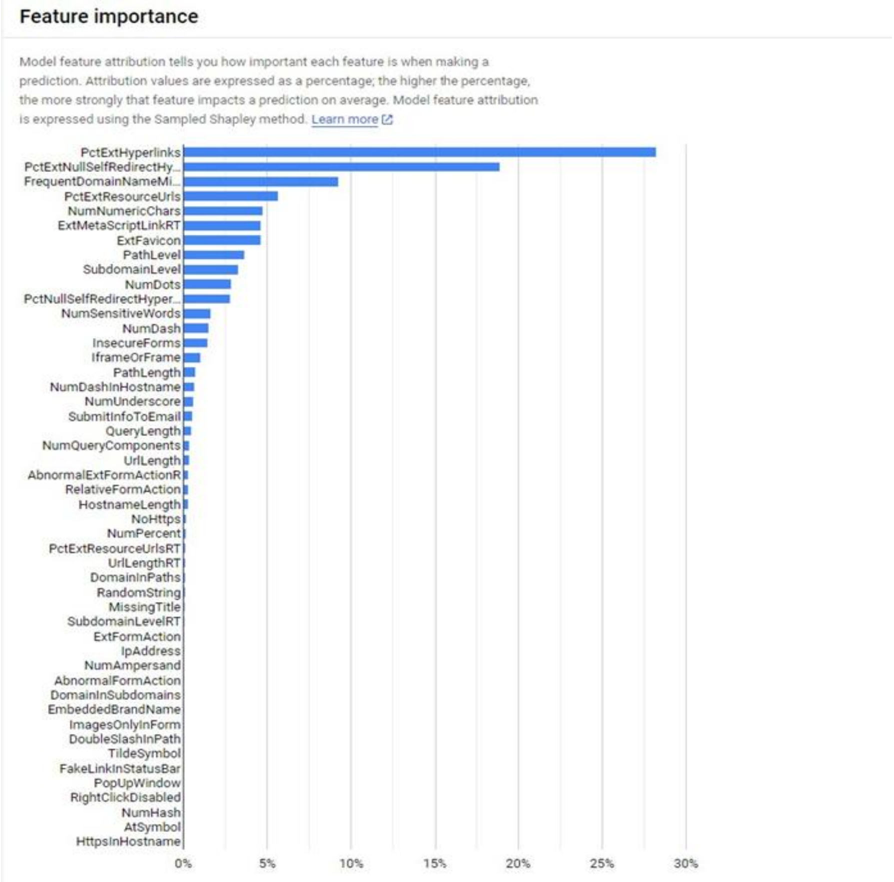
A confusion matrix shows how the model classified each label in the evaluation dataset. The blue, bold cells indicate a correct prediction. A data item is moved to the dropped column if it does not meet the confidence threshold for any label.

True label	Predicted label 1	0	Dropped
1	478	9	0
0	6	488	0

Figure 8: Confusion matrix of Phishing-Legitimate dataset.

4.6 Feature Importance for Phishing-Legitimate Dataset

The feature importance analysis for the Phishing-Legitimate dataset is illustrated in Figure 9. Key influencing features include *Subdomain Level, Domain in Path, No HTTPS, and URL Length.*



Feature importance

Model feature attribution tells you how important each feature is when making a prediction. Attribution values are expressed as a percentage; the higher the percentage, the more strongly that feature impacts a prediction on average. Model feature attribution is expressed using the Sampled Shapley method. Learn more

Figure 9: Feature Importance of Phishing-Legitimate dataset.

4.7 Key Findings

- **The** Random Forest **model outperformed all other classifiers, achieving the highest accuracy of** 95.3% **and an F1-score of** 95.4% **(see Figure 2 and Figure 3).**

- SVM followed closely**, with a slight performance drop compared to Random Forest.**

- Naive Bayes showed the lowest performance**, likely due to its assumption of feature independence, which may not hold for phishing detection.**

- **The** Decision Tree and Logistic Regression models **performed moderately well but were outperformed by ensemble-based models like Random Forest.**

4.8 Feature Importance Analysis

The **feature importance analysis** provides insights into the factors that most strongly influence the model's predictions (Figures 5 and 9).

For the **'Phishing-email-collection' dataset** (Figure 5), the **'Total Number of Characters**' in the email emerged as the most influential feature, accounting for approximately **45%** of the model's predictive power. This suggests that **longer emails are more likely to be associated with phishing attempts**. One possible explanation is that **phishers craft elaborate, deceptive narratives to manipulate recipients into revealing sensitive information**.

Additionally, the high importance of **'Vocabulary Richness**' and **'Unique Words'** suggests that **sophisticated language patterns may indicate malicious intent.** Furthermore, the presence of special characters (e.g., "@," "-," " ," ".") within email content also appears to be a strong predictor, possibly because phishing emails often contain **obfuscated URLs or disguised sender addresses**.

For the **'Phishing-Legitimate' dataset** (Figure 9), the most important predictive factors are URL-related attributes. The **'Subdomain Level'** and **'Domain in Path'** features rank highest, indicating that **phishers often manipulate domain structures to mimic legitimate websites.** This aligns with well-known phishing strategies where attackers use **subdomains or misleading directory names to create deceptive URLs**. Additionally, the **'No HTTPS'** feature is highly relevant, as phishing sites often **lack SSL certificates** to secure communications.

A key insight from this analysis is the **distinct pattern of feature importance between the two datasets.** While **email content features dominate phishing detection in the email dataset, URL-related features are critical in identifying phishing websites.** This emphasizes the necessity of **a multi-layered approach to phishing detection, where email content analysis and URL inspection are integrated** to enhance detection accuracy. Future improvements could explore **hybrid models that leverage both email and URL characteristics for a more comprehensive phishing defense system.**

5 Discussion and Limitations

The results of this study demonstrate the effectiveness of machine learning techniques for phishing email detection. The **Random Forest model** achieved **high accuracy** and **F1-scores** across both the 'Phishing-email-collection' and **'Phishing-Legitimate'** datasets, indicating its strong ability to distinguish between phishing and legitimate examples. Feature importance analysis further revealed that both **email content attributes** (e.g., text length, vocabulary richness) and **URL-based characteristics** (e.g., subdomain level, HTTPS presence) play a critical role in phishing detection.

However**, this study has certain limitations**. The datasets used may **not fully capture the diversity of phishing attacks** observed in real-world scenarios. **New phishing tactics, such as zero-day attacks and region-specific scams, may not be well-represented in the training data**. As a result, the model's performance may decline when applied to **novel phishing campaigns that exploit sophisticated evasion techniques.** Additionally, the **static nature of feature importance analysis** means that **the identified features may**

become less relevant over time as attackers adapt their strategies. Continuous monitoring and updating of features will be essential to maintain detection accuracy.

5.1 Future Research Directions

Future studies should focus on:

- **Evaluating model performance on more diverse and continuously updated datasets** to ensure robustness against zero-day phishing attacks and emerging threats.
- **Developing adaptive learning techniques** that dynamically update model parameters as new phishing patterns emerge.
- **Incorporating additional phishing indicators**

6 Conclusion

This study investigated the effectiveness of machine learning techniques for detecting phishing emails by analyzing both email content and URL characteristics- tics. The results demonstrate that machine learning models, particularly **Random Forest**, can achieve high accuracy in identifying phishing attempts. Our experiments on the **’Phishing-email-collection’** and **’Phishing-Legitimate’** datasets yielded high **ROC AUC** and **F1-scores**, confirming the models’ ability to effectively discriminate between phishing and legitimate examples.

.

6.1 Key Research Questions Addressed

What is the most effective classification algorithm for phishing detection?

Our findings indicate that **Random Forest consistently outperforms** other

algorithms (Decision Tree, Logistic Regression, Naive Bayes, and SVM) in terms of **accuracy, precision, recall, and F1-score.** Specifically, Random Forest achieved an **F1-score of 95.4%,** surpassing SVM’s **94.6%.** The superior performance of Random Forest can be attributed to its ensemble learning approach, which **reduces overfitting, enhances feature selection, and improves classification robustness.**

Which features are most important for phishing detection?

Feature importance analysis revealed that the most influential factors in phishing detection are:

- Email-based Features: Total Number of Characters, Vocabulary Richness, and Unique Words.
- URL-based Features: Number of External Hyperlinks, Subdomain Level, and Special Characters.

Longer emails with a higher degree of vocabulary richness tend to indicate phishing attempts, as attackers craft deceptive narratives. Meanwhile, phishing URLs often contain multiple redirection attempts, obfuscated characters, and unusual domain structures.

How can we build an improved model for detecting phishing attacks?

Based on our findings, an improved model should:

- **Integrate both email content and URL analysis** to maximize phishing detection accuracy.
- **Prioritize key predictive features** identified in feature importance analysis.
- **Leverage ensemble learning techniques**, such as Random Forest, or explore advanced deep learning architectures.

6.2 Practical Implications

This study contributes to advancing phishing detection by providing a **robust machine learning framework** that enhances security, minimizes financial losses, and mitigates reputational damage. Accurate phishing detection is essential for protecting organizations and individuals from cyber threats, and the insights gained from this study can inform the development of more sophisticated anti-phishing mechanisms

6.3 Limitations and Future Research Directions

While our findings are promising, it is important to acknowledge certain limitations:

- The datasets used may not fully represent the diverse range of **real-world phishing attacks**, including zero-day threats and regional phishing tactics.
- The **static nature of feature importance analysis** means that features may become less relevant as attackers adapt their strategies.

Overall, this study lays a strong foundation for improving phishing detection using machine learning, paving the way for future advancements in cybersecurity.

References

[1] Verma, R., and Hossain, N. (2013, November). Semantic feature selection for text with application to phishing email detection. In International Conference on Information Security and cryptology (pp. 455-468). Cham: Springer International Publishing.

[2] Fang, Y., Zhang, C., Huang, C., Liu, L., and Yang, Y. (2019). Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism. IEEE Access, 7, 56329-56340.

[3] Wang, J., Li, Y., and Rao, H. R. (2016). Overconfidence in phishing email detection. Journal of the Association for Information Systems, 17(11), 1.

[4] Yasin, A., and Abuhasan, A. (2016). An intelligent classification model for phishing email detection. arXiv preprint arXiv:1608.02196.

[5] Zareapoor, M., and Seeja, K. R. (2015). Feature extraction or feature selection for text classification: A case study on phishing email detection. International Journal of Information Engineering and Electronic Business, 7(2), 60.

[6] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American statistical association, 78(382), 316-331.

[7] Shouman, M., Turner, T., and Stocker, R. (2012). Applying k-nearest neighbor in diagnosing heart disease patients. International Journal of Information and Education Technology, 2(3), 220-223.

[8] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., and Data, M. (2005, June). Practical machine learning tools and techniques. In Data mining (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.

[9] Valdivia Garcia, H., and Shihab, E. (2014, May). Characterizing and predicting blocking bugs in open source projects. In Proceedings of the 11th working conference on mining software repositories (pp. 72-81).

[10] Chien, C., and Pottie, G. J. (2012, August). A universal hybrid decision tree classifier design for human activity classification. In 2012 Annual In- International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 1065-1068). IEEE.

[11] Steyerberg, E. W., Harrell Jr, F. E., and Goodman, P. H. (1998). Neural networks, logistic regression, and calibration. Medical Decision Making, 18(3), 349-350.

[12] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[13] Pfahringer, B. (2019). Random model trees: an effective and scalable regression method. University of Waikato, New Zealand.

[14] Wisaeng, K. (2013). A comparison of different classification techniques for bank direct marketing. International Journal of Soft Computing and Engineering (IJSCE), 3(4), 116-119.

[15] Noble, W. S. (2006). What is a support vector machine? Nature biotechnology, 24(12), 1565-1567.

[16] Bennett, K. P., and Campbell, C. (2000). Support vector machines: hype or hallelujah? ACM SIGKDD explorations newsletter, 2(2), 1-13.

[17] Ahmad, P., Qamar, S., and Rizvi, S. Q. A. (2015). Techniques of data mining in healthcare: a review. International Journal of Computer Applications, 120(15).

[18] Powers, D. M. (2020). Evaluation: from precision, recall, and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

[19] Masri, R., and Aldwairi, M. (2017, April). Automated malicious advertisement detection using virustotal, urlvoid, and trendmicro. In 2017 8th International Conference on Information and Communication Systems (ICICS) (pp. 336-341). IEEE.

[20] Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[21] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.

[22] Burns, E. (2020). In-depth guide to machine learning in the enterprise. SearchEnterprisedAI.

[23] Mohammed Amir, M., and Al-Daeef, M. (2023). Performance Evaluation of Blacklist and Heuristic Methods in Phishing Emails Detection.

[24] Kharabsheh, M., Banitaan, S., Alomari, H., Alshirah, M., and Alzyoud, S. (2022). Respiratory failure in covid-19 patients a comparative study of smokers to nonsmokers. Indonesian Journal of Electrical Engineering and Computer Science, 27(2), 1127-1137.

[25] Dua, S., and Du, X. (2016). Data mining and machine learning in cybersecurity. CRC press.

[26] Ramzan, Z. (2010). Phishing attacks and countermeasures. Handbook of information and communication security, 433-448.

[27] Choi, H., Zhu, B. B., and Lee, H. (2011). Detecting malicious web links and identifying their attack types. In 2nd USENIX Conference on Web Application Development (WebApps 11).

[28] Chiew, K. L., Yong, K. S. C., and Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications, 106, 1-20.

[29] Raya, A. A., Alis, J. B., Herrero, E. G., and Diaz-Pabo´n, A. O. (2011). Cross-Site Scripting: An overview. Innovations in SMEs and Conducting E-Business: Technologies, Trends and Solutions, 61-75.

[30] Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y. (2010, November). Detecting and characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (pp. 35-47).

[31] Goel, D., and Jain, A. K. (2018). Mobile phishing attacks and defence mechanisms: State of art and open research challenges. computers and security, 73, 519-544.

[32] Al-Musib, N. S., Al-Serhani, F. M., Humayun, M., and Jhanjhi, N. Z. (2023). Business email compromise (BEC) attacks. Materials Today: Proceedings, 81, 497-503.

[33] Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., Bodepudi, V., and Maka, S. R. (2025). Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering. European Journal of Applied Science, Engineering and Technology, 3(2), 41-54.

[34] Kanj, S., Garcia, P., Ros´es, O., and Pegueroles, J. (2025). A Review of Tactics, Techniques, and Procedures (TTPs) of MITRE Framework for Business Email Compromise (BEC) Attacks. IEEE access.