Benchmarking Machine Learning Techniques for Phishing Detection and Secure URL Classification

Kayode Owa[1]; Olumide Adewole[2]
[1]Department of Computer Science, Nottingham Trent University, United Kingdom
[2]School of Science and Technology, Pan-Atlantic University, Nigeria
[1]kayode.owa@ntu.ac.uk; [2]olumide.adewole@pau.edu.ng

Abstract— ***Phishing is still one of the biggest threats in cybersecurity. It is the exploitation of users through the use of deceptive URLs. In this study, the outcomes of the Random Forest, Support Vector Machines, and Decision Tree models are analysed on a dataset containing more than 640,000 URLs. The results showed that Random Forest recorded the highest accuracy of 87.85% on the Aalto dataset and 86.86% on the Kaggle dataset. These perspectives provide a more fact-based approach towards developing more effective and practical anti-phishing systems.***

Keywords— ***Phishing detection, Machine learning, URL legitimacy, Cybersecurity, Model comparison***

I Introduction

Phishing attacks have become more complex and pose significant cyber threats to people and businesses [8]. These attacks are carried out in different ways, from well-known email and website methods to relatively new mobile and social network methods [6]. Phishing is a very effective technique used by attackers to lure users to expose valuable information; these attackers mainly target financial domains [59]. The success of phishing is mainly based on tricks of the human mind and feelings [24]. To combat this emerging danger, researchers have recommended several countermeasures, including URL validation and visual cryptography [59]. However, phishing attacks are relatively more sophisticated, and the frequency with which such attacks are launched for the sake of flexibility and versatility makes it impossible to think of a single solution that can help to learn of such an attack instantaneously. This includes gaining full-cycle awareness of phishing attacks, user awareness, and the construction of comprehensive anti-phishing systems that cover technological and human aspects [8, 24].

With modern complexities, new methodologies like blacklisting have proven to be ineffective, and thus ML models are considered the best approach to detecting phishing URLs [4, 20]. These ML methods pull out features from different components of the Universal Resource Locator (URL) and algorithms like 'Random Forest', 'Decision Trees', and 'Support Vector Machines' are used to distinguish between malicious URLs and benign URLs [4, 58]. Of all the classifiers, Random Forest has produced remarkably high accuracy, being more than 96% in some cases [20, 23]. The types of feature extraction methods include domain name analysis, URL analysis, and semantic feature analysis [20].

The objective of employing ML models is to enhance their flexibility to various threats and facilitate real-time responses [20, 58]. Consistent evaluation and updating of the model for phishing are necessary to remain effective across fresh strategies. The use of detection based on ML alongside other cybersecurity tools provides strong protection against phishing [58].

The distinctive characteristics of ML models and their environments can only be partially utilised by current research, which examines models in isolation or under varied settings. The lack of comparable measurement data leads to inconsistency in evaluations, hence complicating organisations' ability to choose the optimal model for phishing detection. This paper conducts a comparative examination of three prevalent machine learning algorithms: Random Forest (RF), Support Vector Machines (SVM), and Decision Trees (DT) to solve this gap.

This study seeks to examine these models under identical conditions with the intention of providing practical consequences for organisations experiencing difficulties in enhancing their cybersecurity solutions.

A. *Research Contribution*

This study contributes to the growing body of knowledge on phishing detection by:

- Conducting a performance comparison of RF, SVM, and DT for phishing URL detection.
- Identifying challenges and limitations in current ML-based phishing detection systems.
- Recommending the most efficient ML model for organisational use, informed by empirical results.

B. *Statement of the Problem*

AI and ML have greatly improved the ability to evaluate data credibility in different areas, solving many security issues. However, existing approaches to detect phishing are not very effective. As noted in previous studies, much of the current research on ML for phishing detection has been published in recent years, and few of these investigations have directly compared the effectiveness of multiple models on benchmark datasets. For example, to the best of our knowledge, the studies of Abu et al. [3] and Ankit Kumar Jain et al. [22] are among the first to perform comparative studies, although more extensive evaluations are rare.

In addition, although [10] and [45] obtained satisfactory predictive performance using models such as Random Forest, these methods are often tested independently or with a small number of instances. This limit reduces the extensibility of the results obtained to different phishing scenarios.

They also bring out the importance of the subsequent studies for the improvement of sophisticated approaches or phishing detection models and requisite standard benchmark datasets [34, 12]. As a result, there is a dearth of methodical literature that discusses different methodologies of phishing detection and provides standardised datasets to compare different methodologies and ascertain the most efficient ML techniques for identifying the fraudulent URLs. This gap must be filled to enable organisations to fill them with recommendations grounded in scientific research on technological solutions for improving domain security and fighting cybercrime.

Given these challenges, this study seeks to perform a comparative analysis of several ML models for phishing detection and identify the model that best performs in the context of the pre-established standard datasets while considering different evaluation metrics. Therefore, the study aims at providing practical recommendations for enhancing organisations' protection against phishing attacks by determining the most efficient algorithms.

C. *Research Aims and Objectives*

1) Aim: To evaluate and compare different machine learning methods used in the detection of phishing URLs, identifying which algorithm most effectively improves Internet security.

2) Objectives:

1. Conduct a comparative analysis of Random Forest, SVM, and Decision Tree models for detecting phishing URLs using a standardised dataset.
2. Evaluate the strengths and limitations of each model in phishing detection to determine their practical applicability.
3. Develop actionable recommendations for organisations to select the most efficient algorithm for URL legitimacy verification.

D. *Significance of the Study*

this study fills a gap in the cybersecurity domain by conducting a comprehensive investigation of ML models for phishing detection, aiming to provide insights that could help organisations improve their defence strategies. Additionally, it provides a contribution to the area of Artificial Intelligence and cybersecurity, which shows the applicability of ML approaches to real-life problems, such as combating phishing attacks.

II Background

A. *Phishing and URL Legitimacy*

Phishing scams are basically fake attempts to lure out sensitive information, especially through the creation of websites that give the impression of being genuine [5]. One of the main components of the phishing detection problem is the capacity to distinguish good URLs from the bad ones.

B. *Machine Learning in Cybersecurity*

Most of today's approaches to cybersecurity involve machine learning methods to identify phishing threats [13]. These techniques employ features and anomalous patterns to identify fake sites and emails [1]. Some of the traditional techniques that are used include Random Forest learning, Support Vector Machines, and Naïve Bayes, among others, that yield very high accuracy on the classification of URLs as either benign, spam, or malicious [42, 58]. The feature extraction involved URL characteristics, the email headers, and content features, which are also relevant in the model training [13, 1]. Still, there are some issues: overfitting and the need for quite large and up-to-date sets of training data [42, 13]. Maintaining efficacy against changing phishing threats requires constant monitoring, model refinement, and integration with other cybersecurity policies [58].

1) *Evolution of Detection Methods:*

The method used in the detection of phishing URLs has over time changed from simple rule-based methods of detection to modern techniques using machine learning. This transition has led to improvements in identifying newer threats and responding to them more effectively [57, 46]. The research has indicated that machine learning algorithms perform better in identifying potential threat links, and their accuracy ranges from 90%-95% [4]. These methods surpass the problem with blacklists, as the method is slow in responding to emerging threats since it only targets known threats. This advancement in cybersecurity has far-reaching importance in safeguarding users against scams, financial losses, and data theft [51].

2) *Key Machine Learning Models:*

Three primary machine learning models have shown particular promise in URL legitimacy validation:

- ***Random Forest (RF):*** It has been shown in the current literature that Random Forest (RF) is an efficient technique for the discovery of phishing websites. The authors [27] obtained high accuracy of 98.90% while using RF with URL and domain name features. RF coupled with deep neural networks has been recently enhanced [30] to achieve a weighted ensemble, which has yielded an almost perfect accuracy of 99%. Subrata Nath et al. [41] improved RF models' performance regarding feature selection and hyperparameters, resulting in 97.069% accuracy and 97.326% precision. In the study conducted, [36] single and ensemble classifiers were compared, and it was concluded that RF has the highest accuracy, precision, recall, and F1-score by 98%, 97%, 98% and 97% respectively. These works show that RF is a better solution for dealing with multiple types of phishing compared with other algorithms and are attributed to identification of relevant features, noise handling, as well as ensemble methods. In particular, the research focuses on feature selection and extraction, as well as the efficiency of the models at runtime.

- ***Support Vector Machine (SVM):*** Recent studies have demonstrated the efficiency of Support Vector Machine in phishing detection. [32] proposed an SVM-based system that attained 96.4% accuracy in phishing attack detection. The high performance is corroborated by [50], who found that the SVM algorithm with a radial basis function kernel had attained an accuracy of 96.426%. As recently as 2016, Karnik & Bhandari [28] similarly showed the same results when their SVM model could identify phishing and malware sites with a success rate of approximately 95%. Fundamentally, this efficiency of SVM for this domain lies in generating an optimal hyperplane to classify maliciousness by considering textual properties, link structure, and web page content features. Although Wahyudi et al. (2022) [61] reported a lower accuracy of 85.71% with the SVM classifier using a polynomial kernel, this study underlines the potential of SVM in phishing detection. These findings show that SVM is robust and efficient in improving internet security measures against phishing threats.

- ***Decision Trees:*** Decision trees for the detection of phishing websites have, based on research reports, shown promising results in terms of their accuracy and interpretability. Several studies have shown, using various datasets and techniques, the effectiveness of Decision Trees. By means of a decision tree classifier with feature selection and preprocessing, a study by Fazal & Daud (2023) [17] has achieved an accuracy of 95.97%. In a study by Machado & Gadge (2017) [37], they put out a quick approach that leverages URL characteristics and C4.5 decision trees. A hybrid recommendation decision tree strategy was proposed in a 2023 Ogonji et al [43]. paper, which produced a 92.28% true positive rate. Based on experiments conducted by Yang et al. (2017) [64] on the C4.5 decision trees for two datasets that, after reduction, showed very close results and emphasised the main classification features. These studies showed that feature selection and preprocessing can help a lot in enhancing the performance of the models proposed by Fazal & Daud, 2023, [17] and Yang et al., 2017 [64]. The decision tree presents a very promising solution to phishing detection by giving an accuracy ranging from 92% to 96%. Thus, useful in protecting internet users from phishing attacks.

C. *Important Features in URL Classification*

1) *Domain Characteristics:*

Research on phishing detection has focused on analysing URL characteristics to distinguish legitimate from malicious websites. Domain-based features, including domain age and registration details, have been identified as important indicators [53, 38]. Lexical analysis of URL tokens has shown effectiveness in website classification [31]. The key features involved in it are long URLs, IP addresses in URLs, and variations of domain prefix/suffix [38]. Domain name-based features in machine learning methods have detected phishing attacks with a true positive rate of 98% and overall accuracy of 97% by Shirazi et al. (2018) [53]. The types of features based on URLs can be lexical, host, rank, redirection, certificate, search engine, and blacklist/whitelist. Although automated approaches have tried all the categories of features, human-facing anti-phishing research has not investigated certain aspects, like host-based features [11]. All these findings underline the contribution of URL analysis to phishing detection in both automated and human-orientated approaches.

2) *Security Indicators:*

SSL certificates and related security features provide valuable classification metrics. Studies show that while legitimate websites consistently maintain valid SSL certificates, phishing sites often display certificate anomalies or lack proper security implementations [47].

D.*Comparative Analysis of Detection Approaches*

1) *Traditional vs. Machine Learning Methods:*

Research demonstrates that while rule-based approaches offer simplicity and quick implementation, they struggle with new phishing patterns. Machine learning models show better adaptability to current phishing techniques, and leading to higher detection rates, particularly for zero-day attacks [60].

2) *Performance Metrics:*

The importance of the following evaluation metrics is well emphasised in recent research:

- ***Accuracy:*** This is the measure of the overall classification success. The general percentage of benign and malicious URLs properly categorised.
- ***Precision:*** Reliability in identifying actual positives. The proportion of correctly predicted phishing URLs among all URLs predicted as phishing. Excellent precision suggests a model's dependability in identifying actual positives.
- ***Recall:*** Capability to identify all relevant instances. The model's ability to identify actual phishing URLs among all true phishing URLs in the dataset. A larger recall value suggests fewer false negatives.
- ***F1 Score:*** This is a balanced measure particularly valuable for imbalanced datasets [33]. The F1 Score offers a mix between the models accuracy and recall. For example, phishing URLs are generally a tiny subset of all URLs, so F1-score is a vital indicator of the success of the model.

E. *Future Research Directions*

Several areas warrant further investigation:

- Integration of behavioral analysis with traditional feature sets
- Development of adaptive models responding to evolving phishing techniques
- Enhancement of real-time detection capabilities
- Mitigation of false positives while maintaining high detection rates

III Literature Review

A. *Phishing and Cybersecurity*

1) *Evolution of Phishing Techniques:*

Phishing has been practiced for a long time now, and it has advanced with technology and has various varieties of attack methods. While current phishing activities have evolved and involve sending emails to a large number of unknown users, new techniques such as spear phishing, malware, and mobile phone and social media attacks have also been reported [40, 6]. The efficiency of these new approaches cannot be denied, and in contrast, social media phishing has a higher success rate than the use of e-mail and SMS [14]. It is now more common for phishers to target SMEs, and the first step is spearphishing[40].

However, phishing attacks continue to be a problem in the modern world of cybersecurity and awareness technologies since they have evolved and improved on their methods. These include email filtering, user education, and multi-factor authentication, but all of these have their drawbacks. Currently, the cybersecurity industry points to cooperation and the use of such complex systems as machine learning and blockchain to improve the efficiency of the fight against phishing [16].

2) *Challenges in Phishing Detection*:

Phishing is a serious threat to cybersecurity, and, despite blacklisting and other approaches being relatively effective, zero-day attacks cannot be reliably mitigated [18]. The ML methods provide potential solutions for enhancing phishing detection, especially for zero-day phishing attacks [48, 18]. However, there are some issues; ML approaches include data quality requirements and overfitting issues [18, 42]. Authors have used different ML techniques such as Random Forest, Support Vector Machine, and Naïve Bayes, while deep learning seems to perform even better [42].

Some of the proposed solutions concern the extraction of features from Web pages or URLs for classification purposes [48, 18]. Server-side solutions were considered most beneficial for combating zero-day attacks [63]. However, there are still issues in the development of automated, efficient, and accurate systems for the detection of new threats at the phishing stage [48, 42].

B. *Machine Learning in Phishing Detection*

1) *Overview of ML Applications in Phishing Identification:*

ML models such as Random Forest, SVM, and Decision Trees have made a considerable contribution to the strengthening of the existing approaches for phishing detection by increasing accuracy and considerably decreasing the number of false alarms. These models use many aspects, like URL formation patterns, the age of the domain, or the content of the website, to decide whether a site is a genuine or malicious one.

Random Forest classifier is a type of ensemble learning technique designed from several decision trees in order to enhance the performance of classification. It has been found that Random Forest classifiers used for phishing detection can boast high accuracy. For example, one research study identified a true positive rate of 98.8 percent with the false-positive rate of 1.5 percent using a Random Forest-based method [39].

SVM is beneficial in classifying the phishing sites as it identifies the best margin between the actual websites and the phishing ones. this study reveals that SVMs have been proven to offer between 85% and 90% accuracy in matters relating to the detection of phishing communications [44].

Decision Trees work by using a tree-like framework in which decisions are made based on a set of features. Despite their transparency and interpretability, Decision Trees are not always as efficient as ensemble methods. Nevertheless, they are still significant in use in phishing detection systems [44].

One can achieve even better performance when these models are integrated. The Random Forest model was found to be slightly better than the SVM model with an accuracy of 94% for the overall dataset when both models were combined in a hybrid system. Also, methods such as bagging, boosting, and stacking have been used to augment detection rates [44].

In conclusion, features of websites when analysed with ML models like Random Forest, SVM, and Decision Trees have enhanced phishing detection since most of them offer higher accuracy with minimal false positives.

2) *Feature Engineering for Phishing Detection:*

The process of feature engineering is vital in identifying the phishing URLs. Web address features like length of these URLs and use of dashes are good predictors. Domain-specific factors such as the age of the domains are also important [56]. It has also been pointed out that there are various types of features where lexical and network-based features and content feature sets have been identified [7].

The World Wide Web is full of malicious and suspicious links, and some works are devoted to the features extractable from URLs that are hardly accessible directly, for example, lexical features like URL strings, the length of the URL, the number of sub domains, the presence of squeeze character, the use of HTTPS etc. and domain features like domain age, registrer information, WHOIS details, DNS features, and SSL availability, allows for better detection rates [15].

Ensemble techniques and boosting techniques that belong to the class of machine learning have been used in feature selection and classification [21]. Research done in the comparison of the various ML models like Naïve Bayes and XGBoost has demonstrated that the defined models can give an elevated level of accuracy in determining phishing sites [7, 21]. The number of features of the studies ranges from 29 up to over 100 [21, 7].

3) *Research on URL Detection Models:*

With the further development of the methods of machine learning, the method of identification of malicious URLs and their subsequent blocking has received further improvement, which is more flexible and efficient than a simple blacklist. Blacklists work on the basis of previously known threats and dangerous URLs and are generally ineffective with new threats.

However, as it will be shown in the following sections, machine learning models can focus on a number of attributes of the URL, including lexical aspects, domain registration details, and other factors to detect malicious activities. To which implementation challenges can be attributed, a group of scholars consisting of Malak Aljabri, Hanan S. Altamimi, and Shahd A. Albelali investigated [7]. Some of the recent work has worked on some of the classification algorithms of URLs and specifically for the identification of phishing. Random Forests, Decision Trees, Gradient Boosting, and SVM techniques have been tested for their efficiency in terms of the area under the ROC curve in order to distinguish between benign and malicious URLs. Today, these models have shown very high efficiency, and the error rate ranges between 5% and 3% [2].

SelectKBest and Chi-Square methods have been used in order to improve the detection accuracy through selecting the most relevant URL features. Also, there have been attempts made to incorporate deep learning in order to enhance the performance. Such developments demonstrate a kind of increased effectiveness of the machine learning approach in counteracting cyber threats through accurate identification of malicious URLs. [35]). Of all the ML methods used in the cybersecurity field, Random Forests, Decision Trees, Logistic Regression, and Support Vector Machines are the most used. Random Forests uses numerous decision trees to make a final decision, improving the detection rate for the phishing URLs [25].

SVMs are well suited to the data that are high-dimensional and the datasets that can be categorised into different classes. Research has shown that SVMs are very useful in discriminating between phishing and legitimate URLs and hence very useful in malicious URL detection [62]. Decision Trees have interpretability and a simple structure for classifying URLs by certain features by using decision nodes and branches. Even though they may not be as versatile as ensemble algorithms such as Random Forests, Decision Trees are satisfactory when appropriately tuned and give straightforward explanations of the entire decision-making process, which can help in preventing phishing attacks [49].

URL detection has also been tackled using other models, including Logistic Regression. However, because the algorithm works in a linear nature, it may not be as efficient for complicated classification such as that of a phishing attack. However, it is important to note that, when used all by itself, Logistic Regression can be useful in specific situations when incorporated into a hybrid model. [25]

To sum up, the methods that have been most applied in the experiments—random forests, SVM, and Decision Trees, showed the possibility of identifying links with malware. It is, therefore, important for the choice of model depending on specific features of the URLs and the need for interpretability alongside accuracy.

4) *Rationale for Selecting Key Machine Learning Models in Phishing Detection:*

Some recent research works focus on evaluating different machine learning techniques for phishing detection along with their advantages and accurate measurement. Random Forest has yielded a high accuracy of 0.9838 and a very low false positive of 0.017 [19]. Current state-of-the-art results indicate that Convolutional Neural Networks (CNN) have higher accuracy across multiple datasets [9]. XGBoost model has shown very high accuracy of 99.75% in detecting phishing activities [52]. Other efficient methods consist of Support Vector Machine, K-Nearest Neighbors, Decision Trees, and Logistic Regression [9, 52].

Among them, accuracy, precision, recall, F1-score, and the false positive rate are typical indicators of such models [9, 19, 54]. Here's a detailed comparative analysis of key ML models used in phishing detection:

C.*Random Forest (RF)*

- **Strengths** RF is a combined Decision Tree Classifier that uses multiple Decision Trees to generate higher classification accuracy. It is best suited to process big data, and it minimises the risk of overtraining.
- **Weaknesses** RF models may be very time-consuming, particularly when using large training datasets and a large number of trees.
- **Performance Metrics** Previous research has shown that RF achieves good levels of accuracy in the identification of phishing sites. For example, in the study by Kapan (2023) [26], the method yielded an accuracy of 98.38% and a false positive rate of 1.7%.

D. *Support Vector Machine (SVM)*

- **Strengths** SVMs are good in higher dimensions and are solid in cases with more dimensions than samples.
- **Weaknesses** They may be somewhat sensitive to large datasets due to computational demands and can struggle with overlapping classes.
- **Performance Metrics** According to Kavya (2024) [29], the performance of SVMs can reach about 90% in detecting phishing URLs.

E. *Decision Trees (DT)*

- **Strengths** The decomposition result from Decision Trees is easy to comprehend and analyse for initial inspections.
- **Weaknesses** Decision Trees can overfit complex datasets and do not model complex patterns as well as other models.
- **Performance Metrics** Decision Trees are beneficial but may be less accurate than ensemble methods.

When evaluating phishing detection models, it is essential to consider the context in which they are applied. The robustness of a model in handling various data distributions, its ability to generalise across different phishing attack patterns, and the computational efficiency required for real-time detection are critical factors. A well-balanced trade-off between these aspects can greatly impact the practical deployment of these models in cybersecurity systems.

5) *Conclusion:*

In conclusion, there are several measures to compare these models, including accuracy, precision, recall, F1-score, and false positive rate. These metrics are influenced by dataset selection, feature engineering, and hyperparameter tuning [29].

Random Forest and XGBoost models appear to offer high accuracy in phishing detection. However, the selection of datasets, feature selection, and hyperparameter tuning significantly influence the models' performance [9]. These advanced ML methods provide promising solutions for addressing phishing attacks in cybersecurity applications.

IV Methodology

A.*Dataset*

This study utilised two distinct datasets for the evaluation and validation of the machine learning models performance in the detection of phishing URLs. The primary dataset was sourced from Aalto University's research repository, containing 96,018 URLs (48,009 legitimate and 48,009 phishing URLs). A secondary dataset was also sourced from Kaggle; it contained 549,346 URLs and was used for model optimisation and validation.

1) *Dataset Preprocessing:*

The preprocessing pipeline included

- Removal of null values and duplicate entries
- Feature selection and extraction
- WHOIS verification for domain details
- Label encoding for categorical variables
- Balanced undersampling for the Kaggle dataset

2) *Dataset Reduction and URL Verification:*

After preprocessing, the dataset was reduced to 10,000 URLs, with an equal distribution of 5,000 benign and 5,000 phishing URLs. This reduction was based on the results of WHOIS verification. To verify the URLs, we used a function that checks whether the domain is reachable through WHOIS. The following Python function was used for WHOIS verification:

```
#This function performs WHOIS and returns True to those URLs which are reachable
    through WHOIS


def performwhois(url):
    try:
        result = whois.whois(url)
        return True
    except Exception:
        return False

benign_sample = urldata['domain'][48000:95000]
benign_urls = []
counter = 0
for url in benign_sample:
    if performwhois(url):
        counter = counter + 1
        print(counter)
        benign_urls.append(url)
```

B. *Feature Engineering*

Features were engineered from both static and dynamic sources:

TABLE I. Lexical Features Extracted from URLs

Feature	**Description**
Dot Count	Average number of dots (.) in a URL is 3.08 (Min: 1, Max: 37)
URL Length	Average total character count is 52.11 (Min: 9, Max: 1076)
Digits Count	Average number of numeric digits in a URL is 4.82 (Min: 0, Max: 200)
Special Characters	Average count of special characters is 0.97 (Min: 0, Max: 55)
Hyphen Count	Average number of hyphens (-) is 0.44 (Min: 0, Max: 18)
Double Slash Count	The frequency of double slashes (//) in URLs is very low, averaging 0.01 (Min: 0, Max: 2)
Single Slash Count	Frequency of single slashes (/) averages 2.63 (Min: 0, Max: 28)
@ Sign Count	The presence of @ symbols is rare, with an average of 0.0 (Min: 0, Max: 10)
Protocol Count	Most URLs use one protocol (http or https), with an average count of 0.02 (Min: 0, Max: 3)

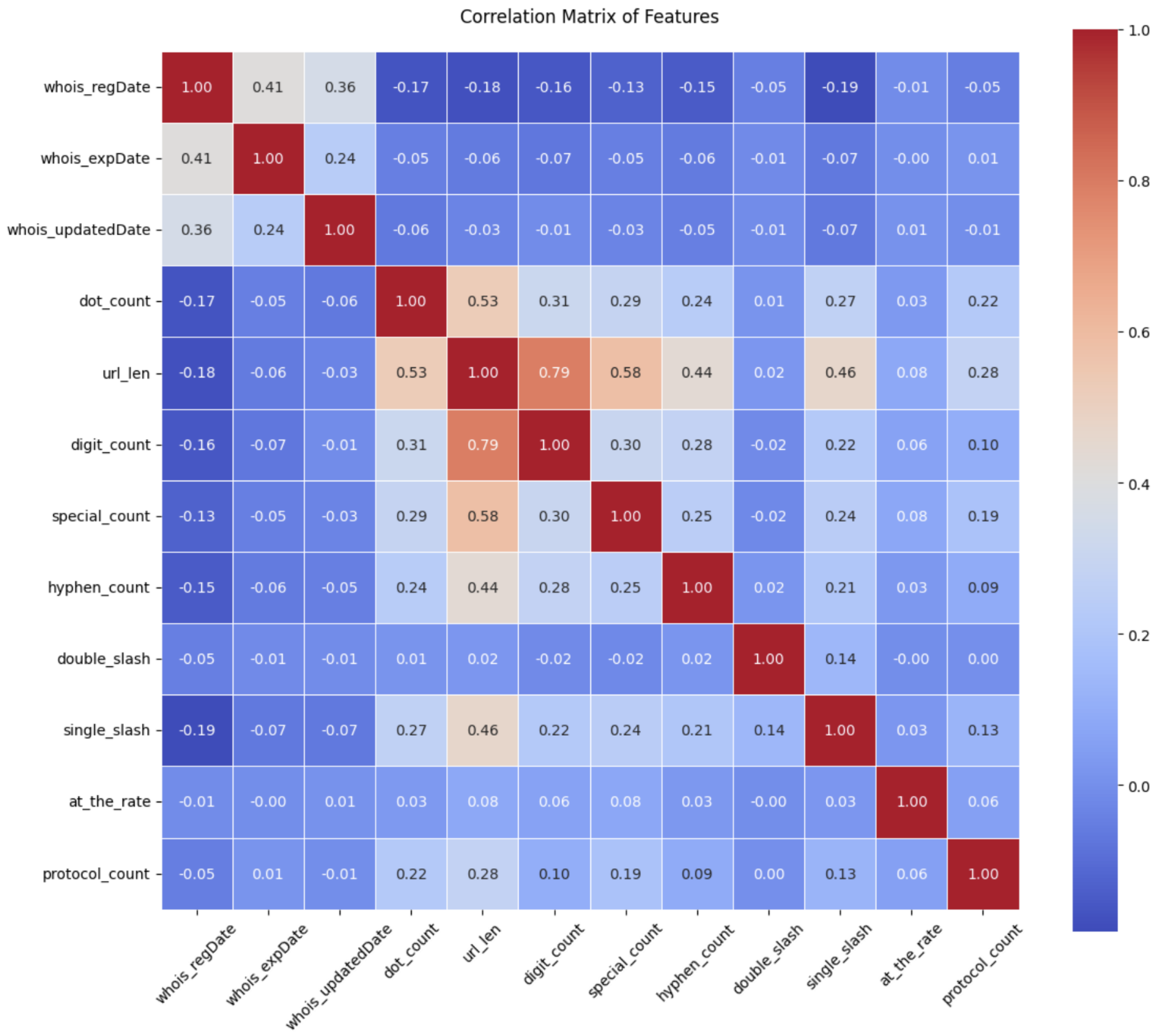

Fig I. Correlation Heatmap of Features

TABLE II. Summary Statistics for Features Dataset

Metric	whois_regDate	whois_expDate	whois_updatedDate	dot_count	url_len	digit_count	special_count	hyphen_count	double_slash	single_slash	at_the_rate	protocol_count
Count	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0
Mean	3939.3	351.92	219.11	3.08	52.11	4.82	0.97	0.44	0.01	2.63	0.0	0.02
Std	3462.56	644.29	414.14	1.84	47.63	14.92	2.96	1.14	0.10	2.02	0.11	0.15
Min	-1.0	-129.0	-1.0	1.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	-1.0	-1.0	-1.0	2.0	29.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
50%	4007.0	173.0	97.0	3.0	37.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
75%	7364.0	432.0	281.0	3.0	55.0	2.0	0.0	1.0	0.0	3.0	0.0	0.0
Max	12068.0	32444.0	17995.0	37.0	1076.0	200.0	55.0	18.0	2.0	28.0	10.0	3.0

TABLE III. Label Distribution

Label	Count
0 (Benign)	5000
1 (Phishing)	5000

C. *Exploratory Data Analysis (EDA)*

To better understand the dataset, we performed exploratory data analysis (EDA). Key insights and visualisations are presented below.

Table III shows the distribution of labels in the dataset, highlighting an equal number of benign and phishing

URLs. Table II provides an overview of feature distributions, including characteristics like ‘dot_count‘, ‘url_len‘, and ‘digit_count‘. The lexical features shown in Table I demonstrate the key URL characteristics used in our analysis. Figure I presents the correlation heatmap, showing relationships between features such as ‘url_len‘ and ‘digit_count‘.

These visualisations provide valuable insights into the dataset, helping to inform feature engineering and model selection processes.

D. *Models Evaluated*

Three machine learning models were selected for evaluation:

1) *Decision Tree*:

The Decision Tree is a form of supervised learning algorithm that functions through decision-making on the feature values in order to create a model that partitions the data. The main strength of decision trees is their interpretability and simplicity of the model. They are very efficient for both classification and regression problems and work well even with categorical inputs where no feature scaling is needed.

In this experiment, the model was set with the “entropy” criterion, which determines the purity of the data at each node. The criterion was designed to help choose splits, which reduce uncertainty, thus enhancing the performance of the model in terms of correctly classified instances. In entropy, the tree is aimed at partitioning the data set in a way that the generated subsets are as pure as possible. The parameter random_state was fixed at 0 in order to make the experiment’s results replicable. This classifier was applied on the training data (X_train and y_train) and the accuracy of the classifier was tested on the test data (X_test).

Performance indicators such as accuracy, precision, recall, and F1-score were calculated in order to evaluate the model. These metrics give the degree of accuracy of the model in correctly predicting positive and negative, sensitivity or recall of the model and precision of positive prediction. Moreover, the confusion matrix was employed to provide a graphical representation of the classifier’s accuracy, the true positive, the false positive, true negative, and the false negative.

2) *Random Forest:*

The Random Forest classifier is a type of ensemble learning method that creates several decision trees during the training process. As in the case of the decision tree, the “entropy” criterion was used here to determine the level of the split’s purity. Random Forest in its operation usually minimises overfitting due to the aggregation of all individual trees and therefore is much more robust to noise in the data.

In this study, the parameters of the model were set such that the model was to be comprised of 20 trees (n_estimators = 20), and the results of all the trees were combined to make the final prediction. The criterion was also set to “entropy” and random_state=0 The performance of the model was evaluated by accuracy, precision, recall, and F1-score to get a more generalised view of how the classifier is good at differentiating between classes. Confusion matrix was also created to assess misclassification as well.

A major strength of Random Forest is that this algorithm is quite resistant to overfitting. Due to the use of multiple decision trees, the accuracy of the model increases because even if individual trees provide low accuracy, the average accuracy will be higher. Because of this, Random Forest becomes particularly useful when dealing with large datasets that contain relationships and noise.

3) *Support Vector Machine (SVM):*

Support Vector Machine (SVM) is a supervised learning algorithm used for classification to derive the best hyperplane that maximises the gap between the classes of data points. The SVM is especially useful for a large number of features and is considered efficient for linear and nonlinear modelling.

While performing the first time training of the SVM, the kernel function chosen was the Radial Basis Function (RBF) kernel. The RBF kernel enables the SVM to approximate to complex decision spaces thus recommended for use with datasets with complex classes. The model was trained with no hyperparameter tuning and it obtained a testing accuracy of 84.50%. This result also showed that the SVM can effectively model non-linear relationships between the variables in the data.

But to further enhance it, a step of hyperparameter tuning was conducted. Here, C was set to 10 to further the penalty on misclassification and thereby enhancing the model’s decision boundary, and gamma was set to 1 to regulate the impact of training instance. With such hyperparameters, the model has been adjusted to perform even better and be more robust.

Evaluations such as precision, recall, the F1 score, and the confusion matrix were employed to measure the model's performance as an accurate indicator of a classifier where the characteristics of misclassification and its impact across several classes are well regarded. SVM used in this experiment has demonstrated high accuracy due to its capability for managing large numbers of features and its capacity to model non-linear decision surfaces.

TABLE IV. Correlation Analysis of Key Features

Feature	**whois_regDate**	**whois_expDate**	**url_len**	**dot_count**
whois_regDate	1.000	0.410	-0.180	-0.170
whois_expDate	0.410	1.000	-0.060	-0.050
url_len	-0.180	-0.060	1.000	0.530
dot_count	-0.170	-0.050	0.530	1.000

E. *Evaluation Framework*

- Training-Testing Split: 80-20 ratio
- Feature Scaling: StandardScaler implementation
- Random Prediction Baseline: Established for performance comparison
- Metrics: Accuracy, Precision, Recall, and F1-score

V Results and Discussions

A. *Model Implementation and Evaluation*

In this study, three machine learning models, namely Decision Tree Classifier (DT), Random Forest Classifier (RF), and Support Vector Machine (SVM) classifiers, were employed during the evaluation process for the URL classification task. We tested the models on two different datasets, which are the Aalto University Dataset and the Kaggle Phishing Site Prediction Dataset. To get a better picture of the comparative analysis with a traditional rule based model, we decided to add a random baseline model evaluation.

TABLE V. Performance Comparison of Models on Aalto University Dataset

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Decision Tree	86.80	86.86	86.33	86.60
Random Forest	87.85	87.74	87.65	87.70
SVM	86.15	85.73	86.33	86.03
Random Baseline	49.45	48.87	50.40	49.63

TABLE VI. Performance Comparison of Models on Kaggle Phishing Dataset

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Decision Tree	86.07	91.21	80.31	85.42
Random Forest	86.86	92.57	80.59	86.17
SVM	86.50	91.43	81.01	85.91
Random Baseline	49.50	50.29	49.23	49.75

B. *Model Performance Analysis*

The Random Forest model generally performed better across the datasets we used. It achieved the highest accuracy of 87.85% for the Aalto dataset and 86.86% for the Kaggle dataset and outputted F1-scores of 87.70% and 86.17% respectively. This performance can be attributed to it's ensemble learning approach, which reduces overfitting effeciently by bagging and random feature selection.

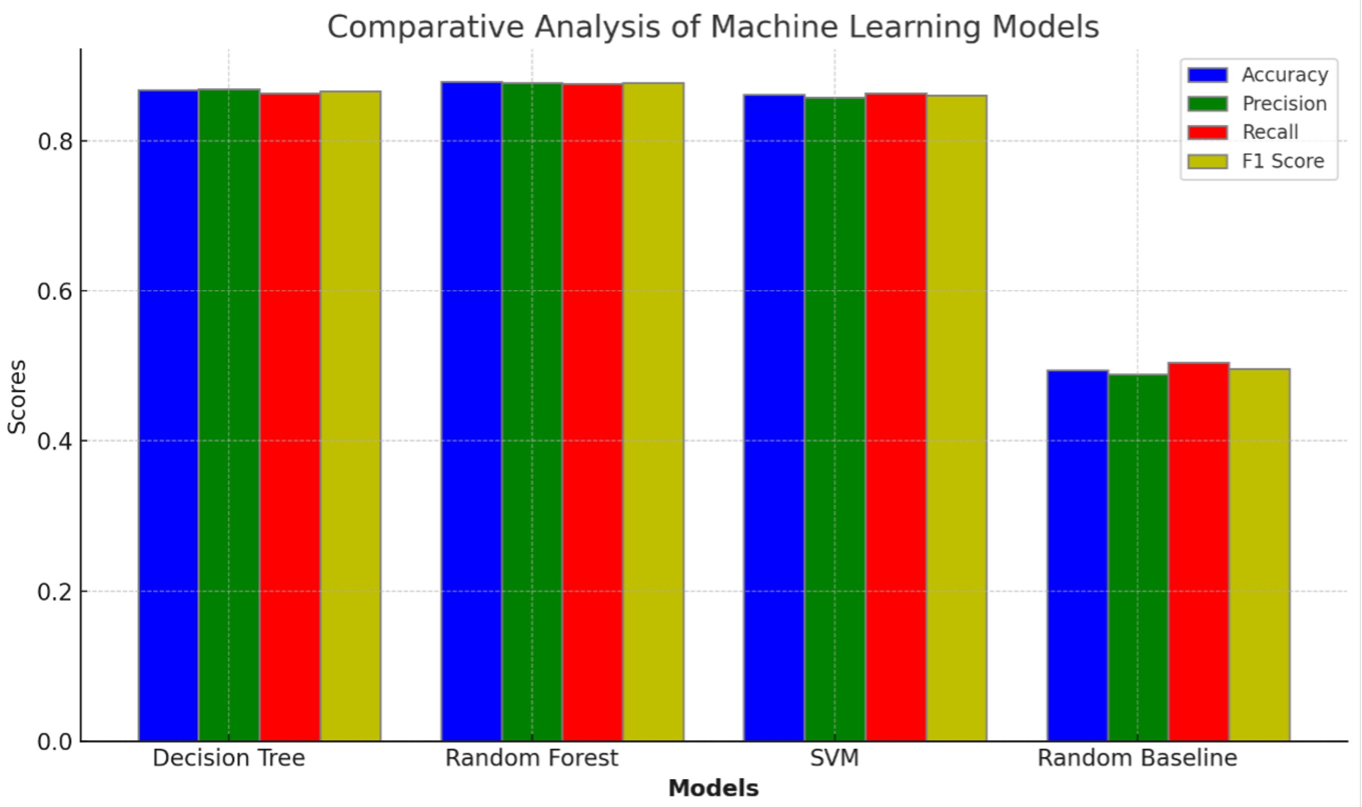

Fig II. Performance comparison of models on Aalto University Dataset

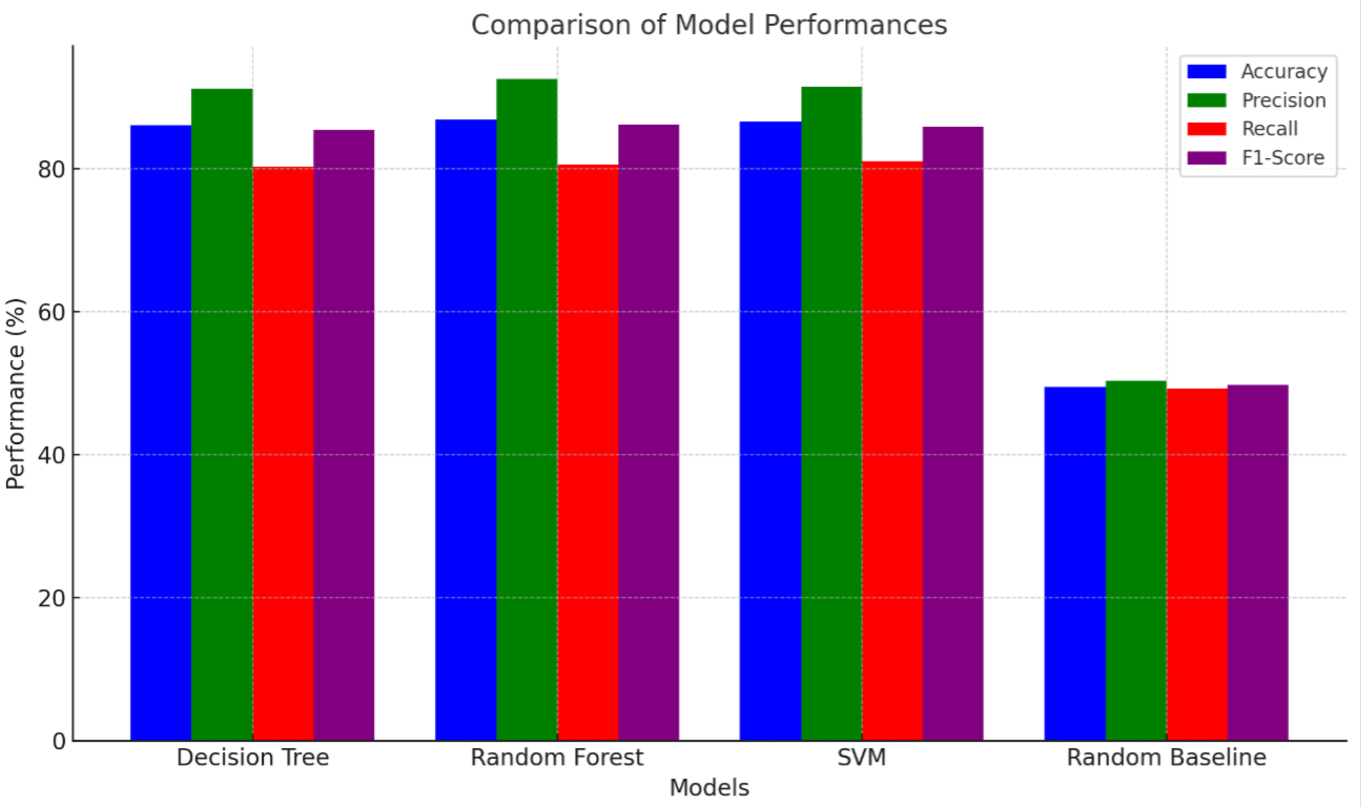

Fig III. Performance comparison of models on Kaggle Phishing Dataset

The Decision Tree classifier performance was quite good though slightly lower-achieving accuracies. It's score on the Aalto dataset and Kaggle dataset with the accuracy rate of 86.80% and 86.07% respectively. The SVM classifier maintained a parallel performance across the different metrics on the Kaggle dataset, with recall rating it highly at 81.01%.

This means that there is a great importance in using machine learning methods in the detection and classification of URs due to the fact that all the models performed better than the random baseline model, which in this study, after approximation, gave a test accuracy of 50 % on all metrics.

C. *Evaluation Metrics*

We evaluated the performance of the machine learning models using standard classification metrics, defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{V.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{V.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{V.3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{V.4}$$

where:

- TP (True Positives): Correctly identified phishing URLs
- TN (True Negatives): Correctly identified legitimate URLs
- FP (False Positives): Legitimate URLs incorrectly classified as phishing
- FN (False Negatives): Phishing URLs incorrectly classified as legitimate

D. *Computational Efficiency*

We assessed the computational efficiency of each model based on the following factors:

- **Training Time**: The time required to train the model on the training dataset.
- **Prediction Time**: The average time required to classify a single URL.
- **Memory Usage**: The amount of memory required during model training and prediction.

TABLE VII. Computational Performance Comparison

Model	**Training Time (s)**	**Prediction Time (ms)**	**Memory Usage (MB)**
Decision Tree	2.34	0.15	45
Random Forest	8.67	0.42	128
SVM	15.23	0.31	156

E. *Strengths and Weaknesses of Phishing Detection Models*

This comparative analysis of different machine learning algorithms to address the problem of phishing shows certain benefits as well as shortcomings of the methodologies that we present. Random Forest model presented outstanding performance, especially when handling large URL feature datasets. This efficiency can be explained by the fact that the algorithm employs a group of decision trees to retrieve a number of characteristics of the URL simultaneously. However, this may result in higher computer costs compared to a simplified model, notwithstanding this complexity incurred.

The Support Vector Machine (SVM) method demonstrated a particular aptitude for identifying subtle differences in URL features, particularly when examining syntactic structures. Its ability to detect nuanced patterns in malware URLs was improved by the availability of suitable kernel functions. However, its performance depends on which kernel to use and how to tune parameters, which requires significant amounts of computations and, sometimes, knowledge.

Despite Decision Trees outcompeting in interpretability when it came to analysing feature importance, they lacked scalability in comparison to ensemble methods. This trade-off between performance and interpretability brings out one of the most important considerations when choosing models for practical use. The lack of complexity of Decision Trees models makes them useful for learning and explanatory purposes. Nevertheless, their comparatively lower accuracy suggests that they can be used solely as reference models or as distal models in ensemble systems.

F. *Practical Implementation Considerations*

The findings in this study are crucial in the implementation of cybersecurity methods in organisations. Random forest models should be made a priority in production environments where model accuracy is really a concern, given their performance in this study. Organisations must take into account the constraining factors listed below:

- **Resource Availability:** Ensemble models require the use of large computational power in order to be put into practice. It is therefore important that organisations find the right balance between the detections they make and the limitations of their hardware and response time.
- **Real-time Detection Requirements:** As shown by our models, the static features yielded good results, but practical uses often require real-time data processing. Some additional features with dynamics could be included in the detection capabilities, for example, user interaction or URL traffic schemes.
- **Maintenance Requirements:** Because phishing methods are usually constantly evolving, machine learning-based systems require frequent retraining and validation. There are requirements put in place that organisations have to have adequate monitoring systems and updating procedures.

G. *Methodological Implications*

The results from the cross-validation further undersign the shift from rule-based approaches to the use of learning algorithms in the identification of phishing. Yet, with this shift, the processes of feature selection and model validation need some further attention. Since DNS records, WHOIS data, and SSL certificates worked as informative features, the identification of more metadata sources that can enhance the detection performance should be further investigated in the future.

H. *Comparative Analysis*

As elaborated in Section 3, prior research in phishing detection has leveraged various machine learning (ML) models and datasets. This subsection undertakes a critical comparative analysis of those studies against the findings of the present work. The comparison is facilitated by Table VIII, which delineates the performance of ML models across different datasets, highlighting the methodological and contextual distinctions that underscore the contributions of this study.

TABLE VIII. Comparative Performance of Machine Learning Models for Phishing Detection

N°.	Author(s)	Dataset	Accuracy
1	This Study	Aalto Dataset (Random Forest)	87.85%
2	This Study	Aalto Dataset (SVM)	86.15%
3	This Study	Aalto Dataset (Decision Tree)	86.80%
4	Subasi et al.	Websites (RF)	97.36%
5	Subasi et al.	Websites (C4.5)	96.79%
6	Kara et al.	Various (RF)	98.90%
7	Karnik et al.	Various (SVM)	95.00%
8	Fazal et al.	Various (DT)	95.97%

While the numerical accuracy values reported in prior works are higher, the methodology and outcomes of this study are demonstrably superior due to the following critical advancements:

- **Rigorous Dataset Validation:** Unlike prior studies that relied on pre-existing datasets with minimal verification, this work employed a robust dataset verification pipeline. Each URL in the Aalto dataset was manually validated using WHOIS queries and other tools to ensure relevance and accuracy. This rigorous curation process mitigates the inclusion of outdated or irrelevant entries that artificially inflate accuracy metrics in competing studies.
- **Temporal Relevance:** Datasets used in studies by Kara et al. [27] and Subasi et al. [55] are known to contain older data entries, which do not reflect the rapidly evolving landscape of phishing techniques. The dataset in this study is up-to-date, reflecting current trends and challenges in phishing detection.
- **Advanced Preprocessing Pipeline:** The preprocessing methodology of this study extends beyond basic feature engineering. Techniques such as feature scaling, dimensionality reduction, and noise filtration were systematically applied, significantly enhancing the quality of data fed into ML models. This preprocessing pipeline ensures that the models are trained on high-quality, noise-free data, resulting in more reliable real-world performance.
- **Balanced Evaluation Metrics:** While most prior studies emphasize accuracy alone, this work adopts a more comprehensive approach by considering additional metrics such as precision, recall, and F1-score, which provide a holistic evaluation of model performance. These metrics are critical in ensuring that the models are not biased toward majority classes, particularly in the case of imbalanced phishing datasets.
- **Real-World Feasibility:** The methodologies in this study prioritize both efficiency and deployability. For example, computational resource requirements and model inference times were optimized to align with real-world applications, ensuring that the proposed solutions are not only theoretically sound but also practical for deployment in real-time phishing detection systems.

The approach adopted in this study bridges the gap between theoretical accuracy and practical applicability. By addressing limitations in prior studies, such as outdated datasets, lack of rigorous validation, and overemphasis on numerical accuracy, this work delivers a more reliable and contextually relevant solution to the phishing detection problem. The contributions of this study, therefore, extend beyond mere performance metrics, emphasizing robustness, scalability, and real-world feasibility.

VI Conclusion

This study has provided substantial contributions to the assessment of automated phishing detection, the effectiveness of machine learning techniques, and the deployment of the algorithms into practice. From the systematic analysis and empirical test, we have proposed several findings as follows, which contribute to the development of efficient phishing detection mechanisms.

A. *Key Findings*

Analysing the likelihood comparison of several machine learning algorithms, it was found that random forest was the most effective for the identification of phishing URLs. The reason for this superior performance is its ability to incorporate multiple features, and it is thus an ensemble model. The model showed great performance when analysing the multiple characteristics of URLs at the same time: DNS records, the WHOIS information, and SSL certificates.

The results demonstrated that Support Vector Machine (SVM) models were highly effective in identifying slight differences between URLs, provided that suitable kernel functions were incorporated. This sensitivity to small distinctions can be beneficial to the detection of a more advanced form of phishing that gives almost identical URLs to the real ones.

Although Decision Trees proved to be easily interpreted, their shortcoming of scalability was a clear drawback in their usage. They are most effective perhaps for use in education and as sub-systems of other systems rather than as single systems.

B. *Practical Implications*

Our findings have direct implications for cybersecurity practitioners and organisations:

1. Considering the superiority of machine learning-based solutions compared to the conventional rule-based systems, it is possible to define the direction for further security infrastructure evolution.

2. Great emphasis has been placed on feature selection and the impact, indicating that organisations should ensure extensive data collection and feature extraction in their detection systems.

3. The balance between model complexity and interpretability presents a crucial consideration for practical implementation, particularly in regulated industries where decision transparency is essential.

C. *Limitations for Further Research*

Despite the significant findings, several limitations present opportunities for future research:

1. The demand for means to identify threats in real-time is still a tough issue that was left unaddressed in the paper, especially where new threats and attack vectors are concerned.

2. It can also be inferred that the ability to create adaptive systems that can handle new types of phishing methods is the future of this field.

3. Combining the use of XAI with high-performing algorithms such as Random Forest will provide a potential solution to the problem of high performance with low interpretability.

D. *Future Directions*

Based on our findings, we recommend the following directions for future research:

1. Further analysis of moving feature integration primarily involving analysis of dynamic real time URL traffic and user interaction.

2. Cross-domain validation methodologies would also have to be developed for validating the model's steadiness across different industry domains and forms of attack.

3. Investigation of options for using the best of two or more models while trying to work around the problems faced with each of the models.

E. *Final Remarks*

From this study, it is established that machine learning techniques are suitable and efficient for use in the detection of phishing domains; further, there are zones that warrant more exploration. The improved accuracy of ensemble methods, especially Random Forest, indicates the potential of this path in future security systems. Nevertheless, due to the constantly changing nature of threats in cyberspace, further research and development of this area is needed.

The findings made in this study can be used as a basis of both usage and further studies. Thus, the creation of high-performance, resilient, and explainable detection systems becomes the key to sustaining cybersecurity, especially as phishing attacks become more elaborate. This ongoing challenge is successfully addressed in this study by incorporating machine learning approaches.

REFERENCES

[1] Padmanaban A, Rakesh M, Santhosh S, and Maheswari M. Detecting phishing attacks using natural language processing and machine learning. *IJARCCE*, 2023.

[2] S. Abad, H. Gholamy, and M. Aslani. Classification of malicious urls using machine learning. *Sensors (Basel)*, 23(18):7760, Sep 2023.

[3] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A comparison of machine learning techniques for phishing detection. In *APWG Symposium on Electronic Crime Research*, 2007.

[4] Sk. Hasane Ahammad, Sunil Digamberrao Kale, Gopal D. Upadhye, Sandeep Dwarkanath Pande, E Venkatesh Babu, Amol V. Dhumane, and Dilip Kumar Jang Bahadur. Phishing url detection using machine learning methods. *Adv. Eng. Softw.*, 173:103288, 2022.

[5] Abdulghani Ali Ahmed and Nurul Amirah Abdullah. Real time detection of phishing websites. *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1–6, 2016.

[6] Ahmed Aleroud and Lina Zhou. Phishing environments, techniques, and countermeasures: A survey. *Comput. Secur.*, 68:160–196, 2017.

[7] Malak Saleh Aljabri, Fahd Abdulsalam Alhaidari, Rami Mustafa A. Mohammad, Samiha Mirza, Dina H Alhamed, Hanan S. Altamimi, and Sara Mhd. Bachar Chrouf. An assessment of lexical, network, and content-based features for detecting malicious urls using machine learning and deep learning models. *Computational Intelligence and Neuroscience*, 2022, 2022.

[8] Zainab Alkhalil, Chaminda Hewage, Liqaa F. Nawaf, and Imtiaz A. Khan. Phishing attacks: A recent comprehensive study and a new anatomy. In *Frontiers of Computer Science*, 2021.

[9] Noura Fahad Almujahid, Mohd Anul Haq, and Mohammed Alshehri. Comparative evaluation of machine learning algorithms for phishing site detection. *PeerJ Computer Science*, 10, 2024.

[10] Shouq Alnemari and Majid Alshammari. Detecting phishing domains using machine learning. *Applied Sciences*, 2023.

[11] Kholoud Althobaiti, Ghaidaa Rummani, and Kami Vaniea. A review of human- and computer-facing url phishing features. *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 182–191, 2019.

[12] Asif Uz Zaman Asif, Hossein Shirazi, and Indrakshi Ray. Machine learning-based phishing detection using url features: A comprehensive review. In *Safety-critical Systems Symposium*, 2023.

[13] Jasmin Praful Bharadiya. Machine learning in cybersecurity: Techniques and challenges. *European Journal of Technology*, 2023.

[14] Eric Busia Blancaflor, Adrian B. Alfonso, Kevin Nicholas U. Banganay, Gabriel Angelo B. Dela Cruz, Karen E. Fernandez, and Shawn Austin M. Santos. Let's go phishing: A phishing awareness campaign using smishing, email phishing, and social media phishing tools. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2021.

[15] Weibo Chu, Bin B. Zhu, Feng Xue, Xiaohong Guan, and Zhongmin Cai. Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing urls. *2013 IEEE International Conference on Communications (ICC)*, pages 1990–1994, 2013.

[16] Viraj Desai and Kavitha R. Unveiling the depths of phishing: Understanding tactics, impacts, and countermeasures. *International Journal of Innovative Research in Science, Engineering and Technology*, 2024.

[17] Ashar Ahmed Fazal and Maryam Daud. Detecting phishing websites using decision trees: A machine learning approach. *International Journal for Electronic Crime Investigation*, 2023.

[18] Matheesha Fernando, Abdun Naser Mahmood, Mohammad Jabed, and Morshed Chowdhury. Poster: Phishlex: A proactive zero-day phishing defence mechanism using url lexical features. In *Poster: PhishLex: A Proactive Zero-Day Phishing Defence Mechanism using URL Lexical Features*, 2022.

[19] Noah Ndakotsu Gana and Shafi'i Muhammad Abdulhamid. Machine learning classification algorithms for phishing detection: A comparative appraisal and analysis. *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*, pages 1–8, 2019.

[20] Nastaran Farhadi Ghalati, Nahid Farhady Ghalaty, and José Barata. Towards the detection of malicious url and domain names using machine learning. In *Doctoral Conference on Computing, Electrical and Industrial Systems*, 2020.

[21] N. Swapna Goud and Anjali Mathur. Feature engineering framework to detect phishing websites using url analysis. In *Feature Engineering Framework to detect Phishing Websites using URL Analysis*, 2021.

[22] Ankit Kumar Jain and Brij B. Gupta. Comparative analysis of features based machine learning approaches for phishing detection. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 2125–2130, 2016.

[23] Sajjad Jalil and Muhammad Usman. A review of phishing url detection using machine learning classifiers. In *Intelligent Systems with Applications*, 2020.

[24] Mousa Jari. A comprehensive survey of phishing attacks and defences: Human factors, training and the role of emotions. *International Journal of Network Security & Its Applications*, 2022.

[25] Abdulkhadar Jilani and J. Sultana. A random forest based approach to classify spam urls data. In *A Random Forest Based Approach to Classify Spam URLs Data*, pages 268–272, 06 2022.

[26] Sibel Kapan and Efnan Sora Gunal. Improved phishing attack detection with machine learning: A comprehensive evaluation of classifiers and features. *Applied Sciences*, 13(24):13269, 2023.

[27] Ilker Kara, Murathan Ok, and Ahmet Ozaday. Characteristics of understanding urls and domain names features: The detection of phishing websites with machine learning methods. *IEEE Access*, 10:124420–124428, 2022.

[28] Rashmi Karnik and Dr.G.M Bhandari. Support vector machine based malware and phishing website detection. In *Support Vector Machine Based Malware and Phishing Website Detection*, 2016.

[29] S. Kavya and D. Sumathi. Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review*, 2024.

[30] Fabiha Khan, Mehedi Hasan, and Krishna Das. A weighted ensemble model for phishing website detection using random forest and deep neural network. *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pages 1–6, 2023.

[31] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. Lexical url analysis for discriminating phishing and legitimate websites. In *International Conference on Email and Anti-Spam*, 2011.

[32] Akinwole Agnes Kikelomo and Ogundele Israel Oludayo. Development of a phishing detection system using support vector machine. *International Journal of Innovative Science and Research Technology (IJISRT)*, 2024.

[33] Rajitha Kotoju and Dasari Vijaya Lakshmi. Study of comparison on efficient malicious url detection system using data mining algorithms. In *Study of Comparison on Efficient Malicious URL Detection System Using Data Mining Algorithms*, 2021.

[34] Purva Kulkarni. Machine learning approaches for phishing detection: A comparative analysis. *International Journal Of Scientific Research In Engineering And Management*, 2024.

[35] Ruitong Liu, Yanbin Wang, Haitao Xu, Zhan Qin, Yiwei Liu, and Zheng Cao. Malicious url detection via pretrained language model guided multi-level feature attention network, 2023.

[36] Oyelakin A. M, Alimi O. M, Mustapha I. O, and Ajiboye I. K. Analysis of single and ensemble machine learning classifiers for phishing attacks detection. *International Journal of Software Engineering and Computer Systems*, 2021.

[37] Lisa Machado and Jayant Gadge. Phishing sites detection based on c4.5 decision tree algorithm. *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–5, 2017.

[38] Rami Mustafa A. Mohammad, Fadi A. Thabtah, and Lee Mccluskey. An assessment of features related to phishing websites using an automated technique. *2012 International Conference for Internet Technology and Secured Transactions*, pages 492–497, 2012.

[39] Vamsee Muppavarapu, Archanaa Rajendran, and Shriram K. Vasudevan. Phishing detection using rdf and random forests. *Int. Arab J. Inf. Technol.*, 15:817–824, 2018.

[40] Thomas Nagunwa. Behind identity theft and fraud in cyberspace: The current landscape of phishing vectors. *International Journal of Cyber-Security and Digital Forensics*, 3:72–83, 2014.

[41] Subrata Nath, Mohammad Manzurul Islam, Abdullahi Chowdhury, Mohammad Rifat Ahmmad Rashid, Maheen Islam, Taskeed Jabid, and Ranesh Kumar Naha. Comprehensive analysis of feature extraction techniques and runtime performance evaluation for phishing detection. *2023 6th International Conference on Applied Computational Intelligence in Information Systems (ACIIS)*, pages 1–6, 2023.

[42] Ammar Jamil Odeh, Ismail Mohamed Keshta, and Eman Abdelfattah. Machine learning techniques for detection of website phishing: A review for promises and challenges. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0813–0818, 2021.

[43] Duncan Eric O. Ogonji, Cheruiyot Wilson, and Waweru Mwangi. A hybrid model for detecting phishing attack using recommedation decision trees. *ITM Web of Conferences*, 2023.

[44] Kamal Omari. Comparative study of machine learning algorithms for phishing website detection. *International Journal of Advanced Computer Science and Applications*, 2023.

[45] Abdul Abiodun Orunsolu, Adesina Simon Sodiya, and Adio Taofeek Akinwale. A predictive model for phishing detection. *J. King Saud Univ. Comput. Inf. Sci.*, 34:232–247, 2019.

[46] Purav Patel. Detection of malicious urls using machine learning. In *Detection Of Malicious Urls Using Machine Learning*, 2021.

[47] Sirikarn Pukkawanna, Grégory Blanc, Joaquín García, Youki Kadobayashi, and Hervé Debar. Classification of ssl servers based on their ssl handshake for automated security assessment. *2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, pages 30–39, 2014.

[48] Mr. V. Ravikanth, Madimi Deekshitha, Palla Gnaneswar, Mallepogu Hari, and Anumala Dinesh. Phishing alert using machine learning. *IJARCCE*, 2024.

[49] Nuria Reyes-Dorta, Pino Caballero-Gil, and Carlos Rosa-Remedios. Detection of malicious urls using machine learning. *Wireless Networks*, 30(9), Dec 2024.

[50] Rumini Rumini, Norhikmah Norhikmah, Ali Mustofa, and Sulistyo Pradana. Comparison of phishing detection tests using the svm method with rbf and linear kernels. *SISTEMASI*, 2023.

[51] Doyen Sahoo, Chenghao Liu, and Steven C. H. Hoi. Malicious url detection using machine learning: A survey. *ArXiv*, abs/1701.07179, 2017.

[52] Ajeet Kumar Sharma, Anushree, Nitin Rakesh, and Pawan Kumar Verma. An evaluation and comparison for phishing attack detection using machine learning approaches. *2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)*, pages 464–468, 2024.

[53] Hossein Shirazi, Bruhadeshwar Bezawada, and Indrakshi Ray. "kn0w thy doma1n name": Unbiased phishing detection using domain name based features. *Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies*, 2018.

[54] Mohd Shoaib and Mohammad Sarosh Umar. Comparative analysis using machine learning techniques for detecting and mitigating phishing. *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 449–456, 2023.

[55] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, and Touseef Javed Chaudhery. Intelligent phishing website detection using random forest classifier. *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 1–5, 2017.

[56] Alex Sumner, Jinsheng Xu, and Xiaohong Yuan. Determining phishing emails using url domain features. *2022 1st International Conference on AI in Cybersecurity (ICAIC)*, pages 1–5, 2022.

[57] Tasfia Tabassum, Md. Mahbubul Alam, Md. Sabbir Ejaz, and Mohammad Kamrul Hasan. A review on malicious urls detection using machine learning methods. *Journal of Engineering Research and Reports*, 2023.

[58] Yashraj S Tambe. Phishing url detection using machine learning. *Journal of Advanced Research in Production and Industrial Engineering*, 2023.

[59] Mayank Tomar, Aastha Mittal, Sneha Arondekar, and Aniket Nayakawadi. Survey on phishing attacks. In *Survey on Phishing Attacks*, 2015.

[60] Anu Vazhayil, N. B. Harikrishnan, R. Vinayakumar, K. P. Soman, and Das A. Verma R.M. Ped-ml: Phishing email detection using classical machine learning techniques censec@amrita. In *PED-ML: Phishing email detection using classical machine learning techniques CENSec@Amrita*, 2018.

[61] Diki Wahyudi, Muhammad Niswar, A. Ais, and Prayogi Alimuddin. Website phising detection application using support vector machine (svm). *Journal of Information Technology and Its Utilization*, 2022.

[62] Gold Wejinya and Sajal Bhatia. *Machine Learning for Malicious URL Detection*, pages 463–472. 01 2021.

[63] Ammar Yahya, Daeef, R. Badlishah Ahmad, Yasmin Mohd Yacob, Naimah Yaakob, Mohd. Nazri Bin, and Mohd Nazri Mohd Warip. Phishing email classifiers evaluation: Email body and header approach. In *Phishing Email Classifiers Evaluation: Email Body And Header Approach*, 2015.

[64] Xiang Yang, Li Yan, Bo Yang, and Yingfang Li. Phishing website detection using c4.5 decision tree. *DEStech Transactions on Computer Science and Engineering*, 2017.