PUMMP: PHISHING URL DETECTION USING MACHINE LEARNING WITH MONOMORPHIC AND POLYMORPHIC TREATMENT OF FEATURES

S. Chanti[1], T. Chithralekha[2], and K. S. Kuppusamy[3]

[1]Department of Banking Technology, Pondicherry University, Puducherry, India
[2,3]Department of Computer Science, Pondicherry University, Puducherry, India

ABSTRACT

Phishing scams are increasing drastically, which affects Internet users in compromising personal credentials. This paper proposes a novel feature utilization method for phishing URL detection called the Polymorphic property of features. In the initial stage, the URL-related features (46 features) were extracted. Later, a subset of features (19 out of 46) with the polymorphic property of features was identified, and they were extracted from different parts of the URL (the domain and path). After extracting the features, various machine learning classification algorithms were applied to build the machine learning model using monomorphic treatment of features, polymorphic treatment of features, and both monomorphic and polymorphic treatment of features. By the polymorphic property of features, we mean that the same feature provides different interpretations when considered in different parts of the URL. The machine learning models were built on two different datasets. A comparison of the machine learning models derived from the two datasets reveals the fact that the model built with both monomorphic and polymorphic treatment of features yielded higher accuracy in Phishing URL detection than the existing works.

While testing the model on phishing URL datasets, the most challenging thing we noticed was detecting the phishing URLs with a valid SSL certificate. The existing works on detecting phishing URLs, using only digital certificate-related features, are not up to the mark. We combined certificate-related and URL-related features to improve the performance to address the problem.

KEYWORDS

Phishing, Anti-Phishing, Non-Content based approach, Monomorphic Features, Polymorphic Features, URL phishing, Credential Stealing

1. INTRODUCTION

Phishing is a malicious activity through which the phishers lure the victims' credentials like username, password, credit card number, CVV number, bank account information, and so on to gain unauthorized access to their account(s). One of the predominant goals of the phisher is to gain financial benefits from the stolen credentials. Phishing attacks have increased drastically during the pandemic. The attackers register domain names resembling prominent Organizations, Financial Institutions, Brands, etc. For example, an attacker can register a domain that resembles World Health Organization (WHO) and send phishing e-mails to victims by pretending themselves as WHO and asking them to click on the links provided in the mail for COVID-19 Solidarity Response Funds [1]. According to the 4th quarter report of the Anti-Phishing Working Group (APWG)in 2021[2], phishing attacks have doubled.

Phishing detection has become more challenging as many phishers use HTTPS-enabled URLs to bypass the filters created by anti-phishing tools. According to the APWG trend report published in 2021, over 82% of phishing URLs use X.509 SSL certificates to fool Internet users [3].

Anti-phishing techniques/solutions have been developed in the form of browser extensions/ toolbars to overcome phishing attacks. The existing anti-phishing solutions can be categorized into two types, i.e., content-based and non-content-based approaches. The content-based approaches analyze the content of either the web page, e-mail, URL or all of them to identify a phishing attack[4], [5]. Heuristic-based, content similarity, machine learning-based, and pattern matching-based approaches fall under this category. Non-content-based approaches like Blacklist, Whitelist, Domain popularity, DNS-based, and Layout similarity-based do not analyze content[6],[7],[8]. Instead, they compare a pre-existing pattern matching to detect phishing attacks. While considering phishing URL detection, machine learning-based solutions perform better in classifying phishing URLs. The performance of the machine learning-based approaches for phishing URL detection mainly depends on the number of features selected. Most of the existing works provide good accuracy, but the error rates are not minimal. There is much scope to improve the performance and reduce the error rate.

The existing works on HTTPS-enabled phishing detection focused on digital certificate-related features (SSL certificate, Certifying Authority, Root Certifying Authority, etc.,). Only a few works focus on some additional features. Detection of phishing URLs with valid SSL certificates is very challenging if the certificate-related features are used alone. For example, if the attacker registers a domain, gets a valid digital certificate, and later uses that domain for malicious activities (creating a clone webpage of any popular domains). The certificate-related features can only detect phishing URLs that use fake, expired, and no certificates. To detect the HTTPS-enabled phishing URLs, the features from URL, website, e-mail metadata, visual similarity features[9], etc., must be used along with the certificate-related features.

In this paper, a novel machine learning-based phishing URL detection model with the monomorphic and polymorphic treatment of features has been proposed and it is found to exceed the performance of existing machine learning methods used for phishing URL detection. The proposed model is capable enough to distinguish the HTTPS-enabled phishing URLs with the help of Public Key Infrastructure (PKI).

The further sections of this paper are as follows. Section 2 explains the methodology of the proposed work. Section 3 discusses the existing literature on phishing URL detection using machine learning. Section 4,5,6,7, and 8 explains the proposed work on Phishing URL detection starting from dataset collection to machine learning model construction. Section 9 provides an Experimental Design and Discussion of the Results. Section 10 is about the application of the proposed model in detecting HTTPS-enabled phishing URLs. Section 11 gives the conclusion of the work.

2. METHODOLOGY

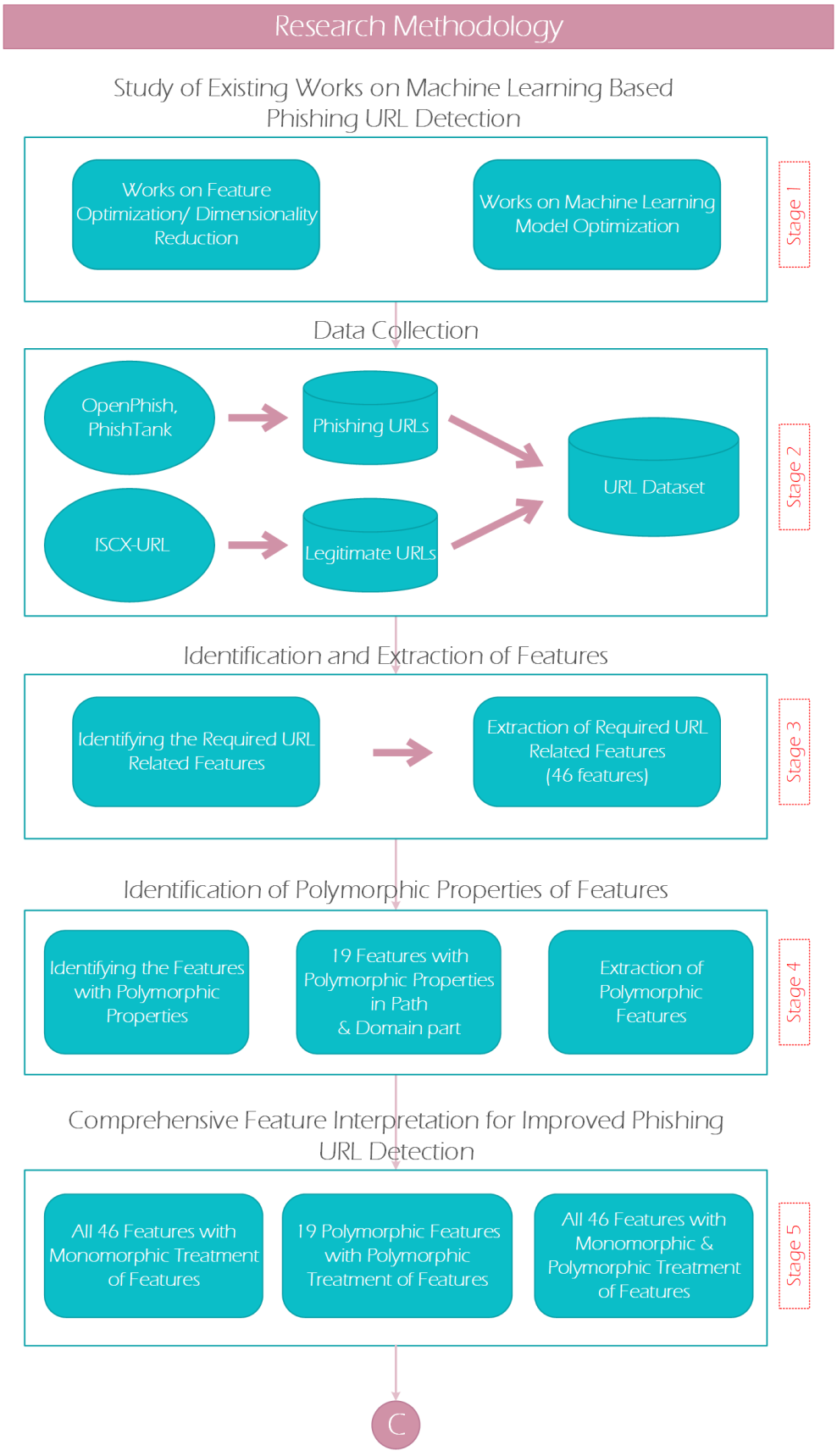
The phishing URL detection on SNS using the novel feature property and feature interpretation has been carried out using the following steps:

1) *Study of Existing Works on Machine Learning-Based Phishing URL Detection*: Researchers have developed anti-phishing solutions to overcome the phishing problem. There are numerous anti-phishing solutions, among which machine learning-based approaches perform better. The existing works on machine learning to detect phishing URLs are considered in this work. It is also essential to determine the other possibilities

to improve the accuracy of phishing URL identification by any other novel approach. In order to perform the same, a URL dataset comprising legitimate and phishing URLs needs to be collected.

2) *Dataset Collection*: The phishing and legitimate URLs are collected from different sources like PhishTank[10], OpenPhish[11], and ISCX-URL-2016 [12]. The reason for collecting the URLs from different sources is that no standard dataset is available for phishing URL detection.
3) *Identification and Extraction of Features*: Existing works on phishing URL detection were thoroughly analyzed by examining various research works, as well as the codes available on GitHub as a reference to identify the essential features required for phishing URL detection. As a result of the analysis, 46 URL-related features which have been found to play a critical role in detecting phishing URLs with higher accuracy have been identified from the existing works. These features have been extracted from the URL datasets collected.
4) *Identification of Polymorphic Properties of Features*: The features with polymorphic properties have been identified and extracted from the URL datasets. The polymorphic property of features means that the same feature provides different interpretations when considered in different parts of the URL. The features that do not exhibit the polymorphic property of features are considered as monomorphic features. A more detailed explanation is given in Section 5.
5) *Comprehensive Feature Interpretation for Improved Phishing URL Detection*: A novel feature interpretation method involving the monomorphic and polymorphic properties of features has been formulated to increase the performance of the phishing URL detection model. Using the identified feature interpretation method, the machine learning model is constructed with all available classification algorithms to determine the best-performing algorithm that yields the highest phishing URL detection accuracy.
6) *Machine Learning Model Construction and Performance Evaluation*: All classification algorithms are applied to the datasets formed by using the various feature treatment methods identified in the previous step. Phishing URL detection was carried out using each feature treatment method, and the classification's accuracy has been observed to determine which of the treatment methods yields better results.
7) *Experimental Design and Discussion of Results*: The experiment procedure was conducted in four stages. Experiment 1 analyzed the performance of the machine learning models built using the monomorphic treatment of all 46 features. Experiment 2 analyzed the performance of the machine learning models which were constructed with the only polymorphic treatment of features (19 out of 46). Experiment 3 assessed the performance of the machine learning models built using both the monomorphic and polymorphic treatment of features. The best results have been obtained by considering the phishing and legitimate URLs in equal proportions. To make sure that the results obtained in experiment 3 are not affected by considering different proportions of phishing and legitimate URLs, experiment 4 is conducted. In experiment 4, the machine learning model is built using the same dataset used in experiment 3, with different proportions of phishing and legitimate URLs.

The complete overview of the phishing URL detection on SNS using the monomorphic and polymorphic properties of features is depicted in Figure1.


Research Methodology
Study of Existing Works on Machine Learning Based Phishing URL Detection
Works on Feature Optimization/ Dimensionality Reduction
Works on Machine Learning Model Optimization
Stage 1
Data Collection
OpenPhish, PhishTank
Phishing URLs
ISCX-URL
Legitimate URLs
URL Dataset
Stage 2
Identification and Extraction of Features
Identifying the Required URL Related Features
Extraction of Required URL Related Features (46 features)
Stage 3
Identification of Polymorphic Properties of Features
Identifying the Features with Polymorphic Properties
19 Features with Polymorphic Properties in Path & Domain part
Extraction of Polymorphic Features
Stage 4
Comprehensive Feature Interpretation for Improved Phishing URL Detection
All 46 Features with Monomorphic Treatment of Features
19 Polymorphic Features with Polymorphic Treatment of Features
All 46 Features with Monomorphic & Polymorphic Treatment of Features
Stage 5
C

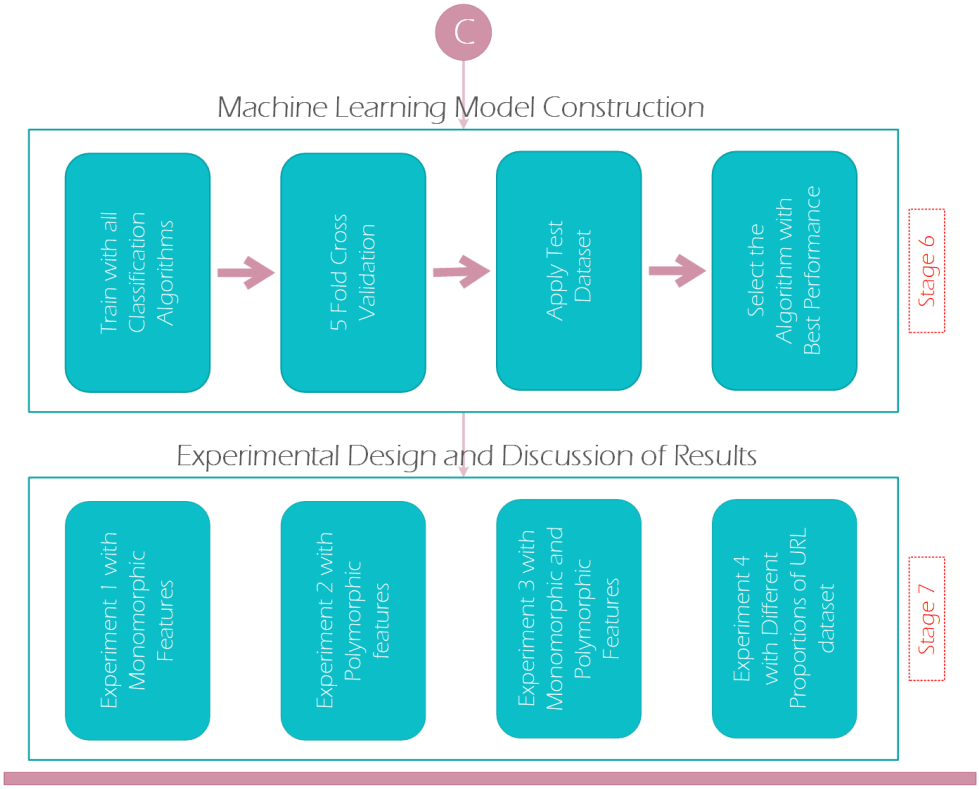


Figure 1. Proposed Methodology for Phishing URL Detection

3. LITERATURE SURVEY

This literature survey was carried out to review the existing machine learning techniques for phishing URL detection. Most of the existing machine learning-based phishing detection techniques are found to use either of the following approaches for phishing URL detection:

- Feature Optimization/ Dimensionality Reduction
- Machine Learning Model Optimization

3.1. Research Works on Feature Optimization/ Dimensionality Reduction

Feature Optimization (FO) uses feature selection methods and Dimensionality Reduction (DR) is typically used to obtain an optimal set of features by eliminating the least important features. However, there is a tiny difference between these two methods. The feature selection methods will simply select the highest-performing features from the complete feature set. In dimensionality reduction, features are combined to generate a new set of optimum features (transformation of features into a lower dimension). Principal Component Analysis (PCA) is one best example for dimensionality reduction. The existing works on phishing detection using FO/DR are described below:

In [13], they developed a machine learning-based framework (PhishMon) for phishing website detection with 15 features which is not dependent on any third-party services for extracting those features. The developed PhishMon achieved 95.40% accuracy with a false positive rate of 1.3% by using Random Forest.

In [14], the authors used 9 features for phishing detection. In their work, all classification algorithms are applied to the proposed model with different feature combinations and found that Random Forest gives better accuracy with 93.20% for nine features, 96.30% for six features, and 84.8% for five features.

Odeh et al. [15] proposed a novel phishing detection model by using feature selection methods to filter the highly correlated features and also employ an adaptive boosting approach with multiple classifiers to increase the accuracy of the model. They used an adaptive boosting classifier and achieved an accuracy of 98.9%.

Almseidin et al. [16] developed a machine learning-based phishing detection model with a novel dataset that contains 5000 phishing and 5000 legitimate web pages. Forty-eight were extracted from the dataset, and different feature optimization techniques were applied to improve the performance. The experimental results were better with 20 features out of 48 with an accuracy of 98.11% using a Random Forest classifier.

3.2. Research works on Machine Learning Model Optimization

Machine learning model optimization is performed to improve the performance by optimizing the model. Machine learning model optimization can be done in the following ways:

- *Tuning the Hyper-parameters*: For instance, in neural networks, the number of hidden layers can be increased or decreased while training the model to improve its performance.
- *Ensemble Learning*: Ensemble learning helps to combine the outcomes of two or more models to achieve better performance in terms of accuracy.

Some of the existing works on phishing detection using machine learning model optimization are given below:

Adeyemo et al. [17] proposed an Ensemble-based Logistic Model Tree (LMT), which is a combination of logistic regression and tree induction methods called AdaBoostLMT and BaGgingLMT. The proposed model is trained on two datasets collected from UCI machine learning archives [18]. The first dataset contains 30 features, and the second dataset contains 10 features. The experiments showed that AdaBoostLMT and BaGgingLMT performed better with an accuracy of 97.42% and 97.18%.

To increase the performance in detecting phishing attacks, [19] presented Convolutional Neural Network (CNN) which is a deep learning-based method. The CNN model is assessed with the dataset collected from UCI machine learning archives [18]. They achieved better accuracy of 97.30% by applying the different configurations in constructing the CNN model from experimental results.

Similarly, Shahrivari, Darabi, and Izadi[20] presented different machine-learning classifiers with 30 features for Phishing detection. On evaluating the classifiers, [20] found the ensemble model (SVM and XGBoost) can provide better accuracy 98.32% with less error rate.

Sonowal and Kuppusamy [21] developed MMSPhiD to detect Typosquatting and phoneme-based phishing attacks. The authors focused on visually impaired people who are highly prone to such kinds of attacks. The proposed model is a machine learning based approach that includes Doublemetaphone and editdistance to correctly identify the phishing domains that look similar to the original one. The machine learning-based approach achieved 94.39% accuracy with a 5.13%

error rate from the experimental results. The Phoneme-based based approach performed better with an accuracy of 99.03% and an error rate of 1.4%.

Wu,Kuo, and Yang[22] proposed URL based phishing detection model using machine learning. The URL(s) from the web page and the web page's source code are two different sources considered for extracting 14 features. Levenshtein distance is used to distinguish the strings. The proposed model achieved an accuracy of 92.60% and a false positive rate of 7.40%, with Support Vector Machine.

3.3. General Machine Learning Based Approaches

There have been few studies on using machine learning for phishing URL detection that does not use Feature Optimization/ Dimensionality Reduction and machine learning model optimization techniques. Instead, the authors have applied various classification algorithms on URL datasets to identify the best performing machine learning algorithm for the given dataset.

Yadollahi et al.[23] developed a phishing detection model by hybridizing the features related to the categories of length, count, suspicious URLs, and hyperlink information into four groups (38 features). The learning classifier system called XCS is used on these 38 features in four groups, individually and all together. The proposed learning classifier system achieved an accuracy of 98.30% with a false positive rate of 1.59%.

Xuan, Nguyen, and Nikolaevich[24] investigated phishing URLs' behaviour and attributes to detect malicious URLs. The performance has been improved by analyzing the malicious behaviour of the URL and exploiting the big data technology. The model is trained with 54 features using Random Forest (RF) and Support Vector Machine (SVM) with varying parameters. The proposed model achieved an accuracy of 99.77% with RF and SVM.
In [25], the authors compared different machine learning techniques on website features categorized as URL lexical structure-based, the domain name associated, and page-based features to detect phishing URLs. Using a random forest algorithm, they trained the model with 26 features and achieved an accuracy of 98.03 %.

Outcome of the Study

The improvement in the performance of phishing detection models mentioned above focuses on either finding the optimal features by applying feature selection methods or optimizing the machine learning models. Table1 presents the outcome of the study in the form of a comparison of the existing works on phishing detection using machine learning. The motivation of this work was to determine if there is any scope to improve the accuracy of phishing URL detection even further by any other novel approach. To perform the same, the URL dataset comprising legitimate and phishing URLs needs to be collected.

4. DATASET COLLECTION

The life span of phishing URLs is very short because the website goes offline or invalid once the attack is executed successfully. If they keep the site online for a long period, then there is a chance of tracing back the attacker. That's how the phisher escapes by stealing the users' personal credentials. So it is very difficult to detect new phishing attacks, with those URLs. The phishing URLs that are active and recent phishing scams reported are required for effective classification of phishing from legitimate URLs. Therefore, to detect the phishing attacks, an elaborate URL collection is required. Both legitimate and phishing URL datasets are required for

Table 1. Existing works on Phishing URL Detection using Machine Learning

S.no.	Author	Features	Dataset	ML Algorithm	Performance Metrics						FO/ DR	ML MO
					FPR	TPR	Accuracy	Precision	Recall	F1 score		
1	[14]	9	UCI Repository	RF, DT	-	-	94.10%	-	-	-	yes	no
2	[17]	30, 10	UCI Repository (2 Datasets)	-	0.028%	97.40%	97.42%	97.40%	97.40%	97.40%	no	yes
3	[13]	15	PhishTank, Alexa	RF	1.30%	-	95.40%	-	-	-	yes	no
4	[19]	30	UCI Repository	CNN	-	-	97.30%	97.00%	98.20%	97.60%	no	yes
5	[22]	14	PhishTank, DMOZ	SVM	7.40%	92.60%	-	-	-	-	no	no
6	[20]	30	PhishTank	RF, XGBOOS	-	-	98.32%	98.72%	98.10%	97.68%	no	yes
7	[23]	38	-	XCS	1.59%	98.19%	98.39%	98.39%	-	98.29%	no	no
8	[16]	48	PhishTank, OPenPhish, Alexa	RF	-	-	98.11%	-	-	-	yes	no
9	[15]	30	PhishTank, MillerSmilesArchive, Google Search	Adaboost	-	-	98.90%	99.00%	98.60%	98.80%	yes	no
10	[21]	12	PhishTank, Alexa	-	1.40%	98.60%	99.03%	99.79%	98.60%	99.19%	no	no
11	[24]	54	PhishTank, Alexa	RF, SVM	-	-	99.77%	98.75%	97.85%	-	no	no

phishing URL detection. The legitimate URLs are collected from ISCX-URL-2016 [12]. The phishing URLs are collected from two authentic sources, i.e., PhishTank [10] and OpenPhish[11]. PhishTank is a non-profit organization that regularly updates phishing attacks reported globally [10]. The phishing URLs that have been reported as phishing by Internet users and experts are verified and included in archives in different file formats. In this work, 15000 phishing URLs were downloaded from the PhishTank website in the form of Comma Separated Value (CSV) files. OpenPhish is another authentic source that provides URLs that have been verified as phishing [11]. The URLs in the archive are updated every 24 hours. To validate this work with the current and updated phishing URLs, the phishing URLs were collected on a daily basis from OpenPhish over a period of two months, which are 80,000 phishing URLs.

After collecting the URLs from the sources mentioned above, two datasets were created:

- Dataset 1 was created by collecting the legitimate URLs from ISCX-URL-2016 [12] and phishing URLs from OpenPhish[11]. A total of 81916 URLs were considered in dataset 1 with equal proportions (50% phishing and 50% legitimate URLs).
- Dataset 2 is created by collecting the legitimate URLs from ISCX-URL-2016 [12] and phishing URLs from PhishTank [10]. Dataset 2 contained 27556 URLs with equal proportions (50% phishing and 50% legitimate URLs).

5. IDENTIFICATION AND EXTRACTION OF FEATURES

Identifying the features related to phishing URL detection was carried out as specified in the following sections.

5.1. Identifying all the Possible URL Related Features

Since the proposed work's objective is to detect phishing URLs, possible URL-related features were collected. The existing works related to phishing URL detection were studied thoroughly along with GitHub codes to gather URL-related features [16], [19], [24]. After a thorough investigation, 46 features were identified as the best-performing features for phishing URL detection. Only URL-related features were chosen because the other phishing-related features like e-mail metadata, the contents within the website were not appropriate when the URL appears as a tweet/ post on social networking sites. All 46 URL-related features identified should be extracted from the URL dataset.

5.2. Extraction of URL Related Features

For every URL collected in dataset 1 and 2 mentioned in Section 4, the required 46 features were extracted and stored as a separate CSV file. The extraction process is explained in the following steps and depicted in Figure 2.

The feature extraction process can be explained as:

1) The URLs in dataset 1 and dataset 2 were taken.
2) For each URL in the dataset, the required 46 URL-related features were extracted.
3) Since numerical features are ideal for training with machine learning algorithms, the output of every feature extracted from the URL was represented in binary format.
4) All the extracted features were stored in a separate CSV file.

6. IDENTIFICATION OF POLYMORPHIC PROPERTIES OF FEATURES

In search of finding other methods or techniques for improving the performance of machine learning models, it is found that some of the features may have polymorphic properties (same feature but different properties in different parts of the URL). For example, if the domain name wxyz.com appears in the domain part of the URL, it is normal. But if the same domain name appears in the path part of the URL; then, it signifies a phishing URL. Thus, the same feature, when considered in different parts of the URL, is found to exhibit different properties. Similarly, if special characters like '*', '#', '%' are present in the path part of the URL, then it is normal, but if the same is present in the domain part of the URL, then it signifies phishing. Likewise, 19 features out of 46 were identified as possessing polymorphic properties from the list of features collected previously, as shown in Table 2. Table 3 presents the features with polymorphic properties and illustrates how the same feature behaves differently when extracted from different parts of the URL. Identifying such properties can enhance the ability to classify phishing URLs correctly.

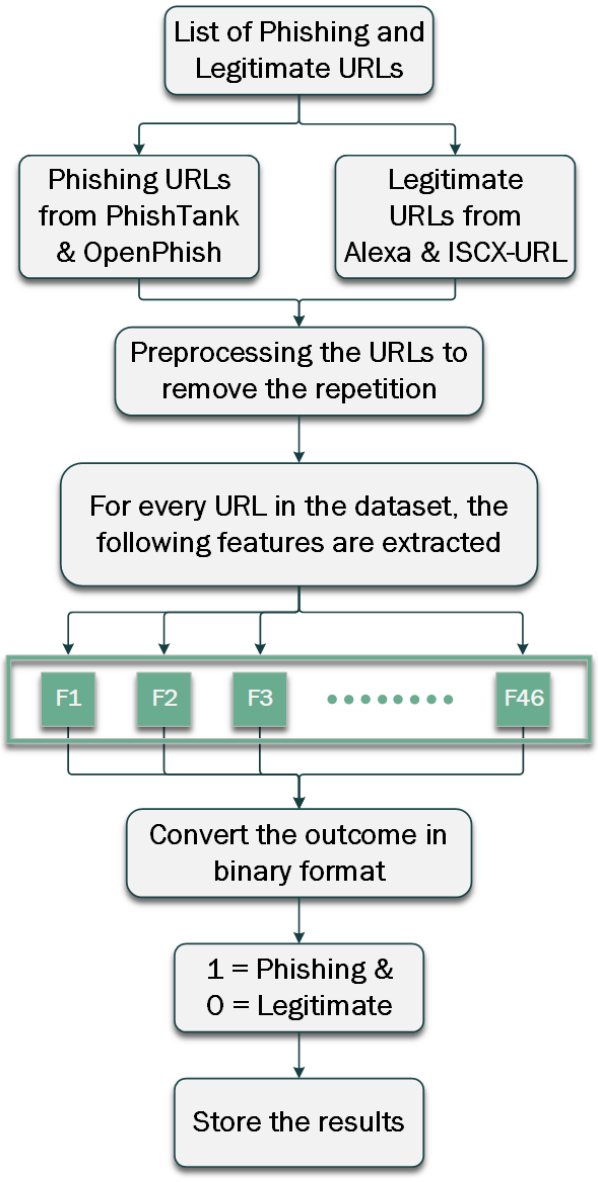


Figure 2. Feature Extraction Process

Table 2. List of URL Features

S. No.	Feature Name	Feature Type		S. No.	Feature Name	Feature Type
1	count dots	Polymorphic		24	entropy	Monomorphic
2	count Hyphen	Polymorphic		25	count subdomain	Monomorphic
3	count Underscore	Polymorphic		26	count Queries	Monomorphic
4	count Fslash	Polymorphic		27	len of domain	Monomorphic
5	count Qmark	Polymorphic		28	countdots subdomain	Monomorphic
6	count equal	Polymorphic		29	is ip	Monomorphic
7	count At Sign	Polymorphic		30	shortening service	Monomorphic
8	count And	Polymorphic		31	SSL Final State	Monomorphic
9	count Exlamation Mark	Polymorphic		32	favicon	Monomorphic
10	count Space	Polymorphic		33	Request URL	Monomorphic
11	count Tilda	Polymorphic		34	URL of Anchor	Monomorphic
12	count Comma	Polymorphic		35	Links In Tags	Monomorphic
13	count Plus	Polymorphic		36	Submit To E-mail	Monomorphic
14	count Star	Polymorphic		37	Abnormal URL	Monomorphic
15	count Hash	Polymorphic		38	Redirect	Monomorphic
16	count Dollar Sign	Polymorphic		39	On Mouseover	Monomorphic
17	count Percent	Polymorphic		40	Right Click	Monomorphic
18	length of URL	Monomorphic		41	PopUp Window	Monomorphic
19	get digit count	Monomorphic		42	Iframe	Monomorphic
20	get double slash	Monomorphic		43	DNS Record	Monomorphic
21	https protocol count	Polymorphic		44	Page Rank	Monomorphic
22	http protocol count	Polymorphic		45	Links Pointing Page	Monomorphic
23	number of Numbers	Monomorphic		46	get www	Monomorphic

Table 3. List of features with polymorphic property

Feature Name	Entire URL	Domain	Path
count dots	Normal to have dots in the URL	Presence of dots is normal but if there is more number of dots then suspicious	suspicious if the count of dots is greater than 1
count Hyphen	Normal to have hyphen in the URL but not in domain part	suspicious if it is present in domain part	Normal if it is present in path
count Underscore	Normal to have underscore in path part of an URL	suspicious if it is present in domain part	Normal if it is present in path

count Fslash	Normal to have forward slash in the URL	suspicious if it is present in domain part	Normal if it is present in path
count Qmark	Normal to have question mark in the URL	suspicious if it is present in domain part	Normal if it is present in path
count Equal	Normal to have equal symbol in the URL path	suspicious if it is present in domain part	Normal if it is present in path
count At Sign	Suspicious if it is present in the URL	suspicious if it is present in domain part	suspicious if it is present in path
count And	Normal to have and symbol in the path of an URL	suspicious if it is present in domain part	Normal if it is present in path
count Exclamation Mark	Normal to have exclamation mark in the path of an URL	suspicious if it is present in domain part	Normal if it is present in path
count Space	Normal to have space in the form of \%20 encoding format	suspicious if it is present in domain part	Normal if it is present in path
count Tilda	Normal to have Tilda in the URL	suspicious if it is present in domain part	Normal if it is present in path
count Comma	Normal to use comma in encoded form in path part of the URL	suspicious if it is present in domain part	Normal if it is present in path
count Plus	Normal to have plus symbol in the path of an URL	suspicious if it is present in domain part	Normal if it is present in path
count Star	Normal to have star in the URL	suspicious if it is present in domain part	Normal it it is present in path
count Hash	Normal to have Hash symbol in the URL	suspicious if it is present in domain part	Normal if it is present in path
count Dollar Sign	Normal to have Dollar sign in the URL path of the URL	suspicious if it is present in domain part	Normal if it is present in path
count Percent	Normal to have \% in the URL	suspicious if it is present in domain part	Normal if it is present in path
https protocol count	Normal if it is present at the beginning of the URL	suspicious if it is present in domain part	suspicious if it is present in path
http protocol count	Normal if it is present at the beginning of the URL	suspicious if it is present in domain part	suspicious if it is present in path

7. COMPREHENSIVE FEATURE INTERPRETATION FOR IMPROVED PHISHING URL DETECTION

The comprehensive feature interpretation on the polymorphic property of features for phishing URL detection can be done by extracting the identified polymorphic property of features from the domain and path part of the URL and combining them with the complete URL features considered in this work.

At first, all 46 URL-related features (monomorphic features) extracted from the URL datasets were considered. All machine learning algorithms for the classification of phishing and legitimate URLs are applied with these 46 features. Figure3 shows the performance with the monomorphic

treatment of features. All the URL features with polymorphic properties alone were considered here. Out of 46 features, 19 have polymorphic properties. As such, these 19 features were extracted from the domain and path parts of the URL, and performance with these features alone is examined here. The results of the polymorphic treatment of features alone are shown in Figure 5. The performance of features with polymorphic properties alone is insufficient to classify phishing URLs effectively. Therefore, the monomorphic and polymorphic treatment of features were combined to improve the performance of the phishing URL detection model. As a result, the implementation of the phishing URL detection model with the monomorphic and polymorphic treatment of features increased the accuracy, as shown in Figure6.

8. MACHINE LEARNING MODEL CONSTRUCTION AND PERFORMANCE EVALUATION

8.1. Machine Learning Model Construction

To construct a machine learning model for phishing URL detection, the features extracted from different datasets were considered to train the model. The machine learning model was built using various classification algorithms viz. Decision Trees, K Nearest Neighbours (KNN), Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Kernel SVM, and Random Forest. The machine learning model outcome of each algorithm is subject to performance evaluation to determine which algorithm yields the best outcome. While training, 5-fold cross-validation was also performed for better training. The random forest algorithm is found to perform the best when compared to other classification algorithms.

8.2. Performance Evaluation

In this phase, a new set of data is given to the prediction model to predict whether the given URL is phishing or not. To know how the machine learning model performs with the selected dataset, the following performance metrics were computed [26].

8.2.1. Accuracy

Accuracy is a simple metric used to predict the number of values correctly classified over the total number of values.

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

Where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

8.2.2. Precision

Precision specifies to us how good the proportion of positive predictions was. It can be calculated by counting the true positive samples (TP) and dividing them by the total positive, correct, or wrong predictions (TP, FP).

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

8.2.3. Recall

The recall is much like the precision to measure the correctly identified proportion of the actual positive values.

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

8.2.4. F1 Score

F1 score indicates the harmonic mean of Precision and Recall.

$$F1\ \text{Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

8.2.5. ROC_AUC

AUC is the area under a curve, which measures the whole area below the ROC curve.

8.2.6. Mean Absolute Error

The mean absolute error is the average of the differences between the true and predicted values. It is mathematically formulated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$$

Where:

N = Number of data points
x_i = Actual values
$\hat{x}_i$ = Predicted values from regression model

9. EXPERIMENTAL DESIGN AND DISCUSSION OF RESULTS

The experimental process in this work has been performed on two datasets in four different experiments which are explained in the following sections.

9.1. Experiment 1: Analysis of the Performance with Monomorphic Treatment of Features (46 features)

Experiment 1 was performed with two different datasets with 46 features extracted from each URL. Dataset 1 contained 81916 URLs and dataset 2 had 27556 URLs. In both datasets, the phishing and legitimate URLs were taken in equal proportion. In monomorphic treatment, all 46 features were extracted for each dataset first, and then the performance evaluation with respect to machine learning algorithms was carried out. Figure3 depicts the performance of the phishing URL detection model, and Figure 4 shows the Mean Absolute Error (MAE) rate on two different datasets. From the results it is found that the performance with monomorphic treatment on dataset 1 and 2 is 99.66% and 99.67% respectively.

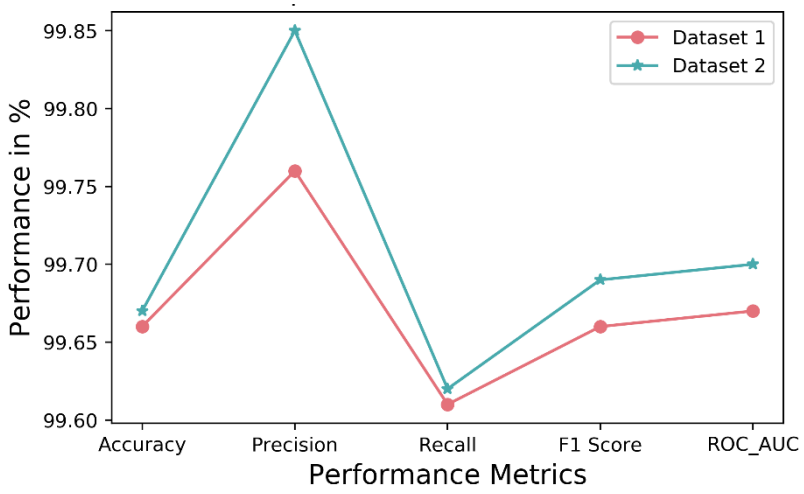


Figure 3. Performance Evaluation with Monomorphic Treatment on Datasets 1 and 2 with 46 Features

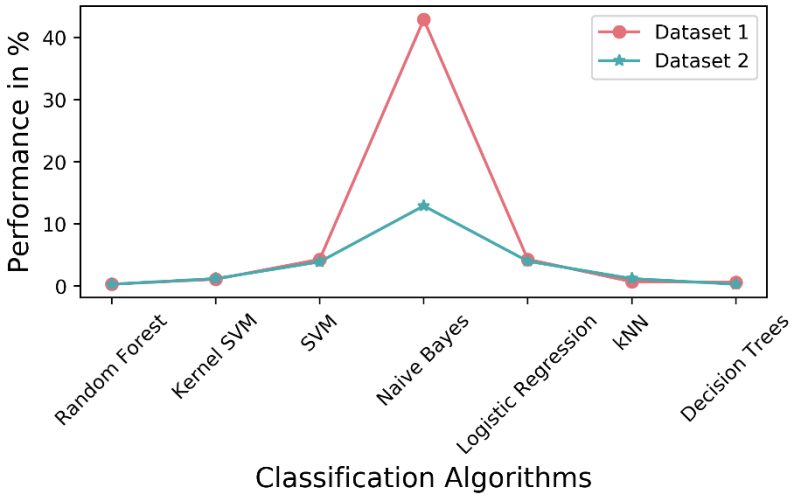


Figure 4. Mean Absolute Error for Monomorphic Treatment on Datasets 1 and 2

9.2. Experiment 2: Analysis of the Performance with Polymorphic Treatment of Features (19 Features)

In this experiment, the features with polymorphic properties alone are considered. Out of 46 features, 19 features were identified with polymorphic properties. These 19 features were extracted from the domain part and the path part of the URL. The machine learning model was built on combining the polymorphic features extracted from the path and the domain part of the URL. This is because only features extracted from either the domain part or path part alone cannot decide whether the given URL is phishing or not. For example, the presence of a domain name in the domain part of the URL is normal but the presence of a domain name in the path part of the URL is not normal. If the model is built on the domain part or path part of the URL alone, it can lead to misclassification. Table 3 shows the list of polymorphic features and also specifies how the same features give different results when applied to different parts of the URL. After applying the machine learning models to the collected datasets, it was found that Random Forest performed better with an accuracy of 91.47% on dataset 1 and 94.80% on dataset 2, as shown in Figure5.

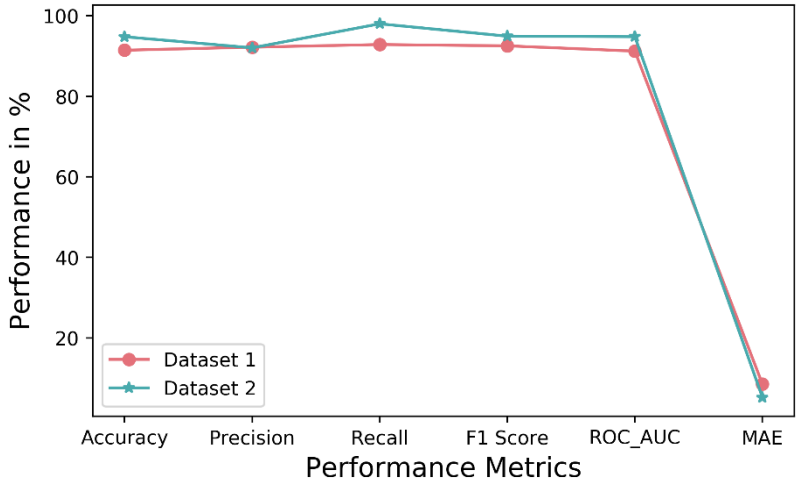


Figure 5. Performance Evaluation with Polymorphic Features on Datasets 1 and 2

9.3. Experiment 3: Analysis of the Performance by Combining both monomorphic and Polymorphic Treatment of Feature

In experiment 1, it was found that the proposed model performs better in classifying the phishing URL with a monomorphic treatment of 46 features. This experiment was about improving the performance of the proposed model by combining monomorphic (46 features) and polymorphic (19 features) treatment of features. Even though the polymorphic features are the subset of monomorphic features, the reason for not excluding those 19 features from monomorphic features is that the monomorphic features are applied on the entire URL. Whereas, the polymorphic features are applied only on the domain and path part of the URL. After combining monomorphic and polymorphic features, the overall accuracy of the proposed model increased to 99.85% and the MAE was 0.2%. Figures6 and 7illustrate the performance of the machine learning models and their error rates on two different datasets.

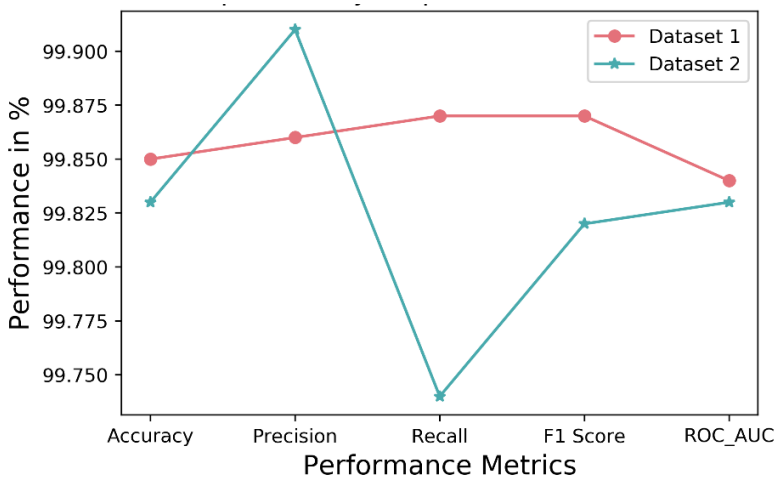


Figure 6. Performance Evaluation for Monomorphic and Polymorphic Treatments of Features on Datasets 1 and 2.

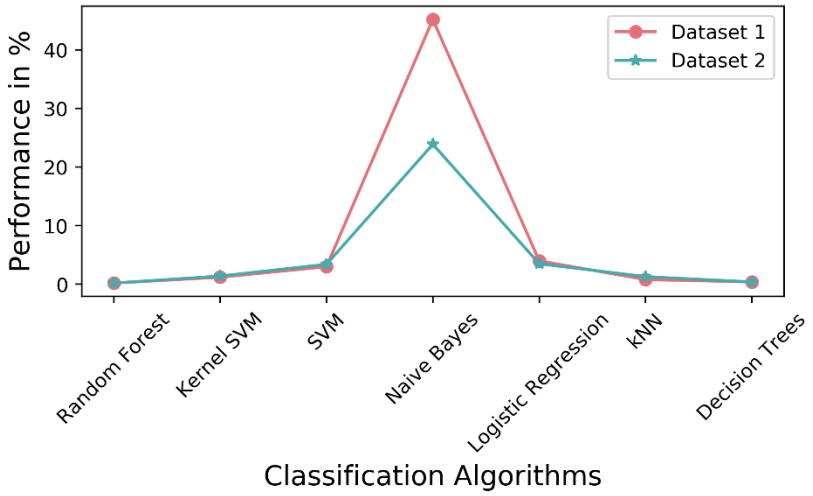


Figure 7. MAE for Monomorphic and Polymorphic Treatments of Features on dataset 1 and 2 with different machine learning algorithms

9.4. Experiment 4: OpenPhish Dataset with 40958 Records (46 Features)

As in the above experiments, the phishing and legitimate URLs were taken in the equal proposition. To prove that the proposed model provides the same performance even if the proportion of the phishing and legitimate URLs in the dataset is changed, a different proportion of the phishing and legitimate URLs were considered and the performance of the machine learning model was analyzed. In this experiment, the dataset contains 40958 phishing and legitimate URLs. The experiment was performed with five different combinations of URLs. Starting from a 70% - 30% proportion, the phishing URLs are modified to 60%, 50%, 40% and 30%, whereas the legitimate URLs are increased to 40%, 50%, 60%, and 70%.

After calculating the performance in terms of accuracy, it is the same or close to the actual accuracy for all these five combinations. The results are tabulated in Table 4.

Table 4. Performance Metrics with Different Proportions using 46 Monomorphic Features

Phishing - Legitimate URLs	Testing Accuracy	Precision	Recall	F1 Score	ROC - AUC	MAE
70 - 30	99.66%	99.80%	99.72%	99.76%	99.63%	0.3%
60 - 40	99.67%	99.78%	99.67%	99.73%	99.67%	0.3%
50 - 50	99.67%	99.72%	99.62%	99.67%	99.67%	0.3%
40 - 60	99.67%	99.65%	99.53%	99.59%	99.65%	0.3%
30 - 70	99.70%	99.64%	99.38%	99.51%	99.61%	0.3%

9.5. Discussion of Results

The above experimental results convey that the proposed model performs better in classifying phishing and legitimate URLs. Initially, in experiment 1, the monomorphic treatment with 46 features was used for phishing URL detection, resulting in an accuracy of 99.67% with a Mean Absolute Error (MAE) of 0.3%. A subset of features (19 out 46) was identified as the features with the polymorphic property. The performance with polymorphic features alone is 94.80% with an MAE of 5.80% as shown in experiment 2. To improve the overall performance, experiment 3

was conducted by combining monomorphic treatment with polymorphic treatment, and, the accuracy of the proposed model increased by 0.19% i.e., 99.85% with an MAE of 0.2%. A Comparison of the proposed phishing URL detection model with the existing works is tabulated in Table 5. The important contribution of the proposed work is as follows:

- Identification of features with polymorphic properties.
- A novel method for improving the performance of a machine learning model using monomorphic and polymorphic treatment of features, which is different from the existing features optimization and machine learning model optimization methods.
- Identification of HTTPS-enabled phishing URLs that use a valid SSL certificate but redirect Internet users to a spoofed or fake site.

Table 5. Comparison of Existing Works with Proposed Work

Comparison of Existing Works	FPR	TPR	Accuracy	Precision	Recall	F1 measure
[20]	-	-	98.32%	98.72%	98.10%	97.68%
[23]	1.59%	98.19%	98.39%	98.39%	-	98.29%
[15]	-	-	38.90%	99.00%	98.60%	98.80%
[21]	1.40%	98.60%	99.03%	99.79%	98.60%	99.19%
[24]	-	-	99.77%	98.75%	97.85%	-
Monomorphic Treatment	**0.22%**	**99.38%**	**99.67%**	**99.85%**	**99.62%**	**99.69%**
Polymorphic Treatment	**8.37%**	**98.01%**	**94.80%**	**92.05%**	**98.01%**	**94.94%**
Monomorphic + Polymorphic Treatments	**0.18%**	**99.87%**	**99.85%**	**99.86%**	**99.87%**	**99.87%**

10. APPLICATION OF PROPOSED MODEL IN DETECTING HTTPS ENABLED PHISHING URLS

The proposed phishing URL detection model improves the performance by applying both polymorphic treatments of features and digital certificate-related features. Features like HTTPS protocol count, SSL final and DNS record from the list of features considered in this work will help in identifying the phishing URLs, even if they are HTTPS enabled. HTTPS protocol count is used to check whether the URL contains more than one HTTPS protocol used. If so, it is considered phishing. For example, the following phishing URL *https://10jt78ulye.s3.us-south.objectstorage.softlayer.net/epiblemata/index.html?key=d653fefd64a5fb5f0a63d46c4620a667&redirect=https://www.amazon.com* has two domains used and among that only softlayer.net is the actual domain, and the amazon.com appearing at the end of the URL is simply a trick to fool the victim. The SSL final state will verify the domain's digital certificate and the certifying Authority (CA) who signed the certificate. DNS record verification will help to check whether the attacker had manipulated the IP address of the domain or not.

With these three features alone, it is difficult to tell whether the HTTPS-enabled URL is phishing or not. A fake or spoofed webpage can be hosted on a trusted domain, which can be misclassified as a legitimate URL. Along with these three features, the remaining URL-related features listed in Table 2 are also considered to classify the phishing URLs correctly.

11. CONCLUSIONS

In this paper, a novel feature interpretation method on the polymorphic treatment of features was introduced to increase phishing URL detection performance. Initially, the phishing URL detection model was developed using 46 features, resulting in an accuracy of 99.67%. Later, a subset of features, say 19 from 46 features, were selected as polymorphic features and extracted from the domain and path parts of the URL. These polymorphic features were combined with the remaining features to improve the performance of the proposed phishing URL detection model. The accuracy of the proposed machine learning model with monomorphic and polymorphic treatment is 99.85%. Thus, the interpretation of the monomorphic treatment of features along with the polymorphic treatment of features is found to be the best in improving the performance of the phishing URL detection model.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] World Health Organization, "Beware of criminals pretending to be WHO," 2020. https://www.who.int/about/communications/cyber-security (accessed May 05, 2021).

[2] APWG, "APWG Phishing Trends Report 4th Quarter 2021," 2021.

[3] APWG, "APWG Phishing Trends Report 2nd Quarter 2021," 2021.

[4] M. Dunlop, S. Groat, and D. Shelly, "GoldPhish: Using images for content-based phishing analysis," *5th International Conference on Internet Monitoring and Protection, ICIMP 2010*, pp. 123–128, 2010, doi: 10.1109/ICIMP.2010.24.

[5] H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, "Textual and visual content-based anti-phishing: A Bayesian approach," *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1532–1546, 2011, doi: 10.1109/TNN.2011.2161999.

[6] N. Vaishnaw and S. R. Tandan, "A Bird's Eye View of Anti-Phishing Techniques for Classification of Phishing E-Mails," *International Journal for Research in Applied Science & Engineering Technology*, vol. 3, no. 6, pp. 263–275, 2015.

[7] Y. Zhang, L. Cranor, S. Egelman, and J. Hong, "Phinding phish: Evaluating anti-phishing tools," *HumanComputer Interaction Institute*, no. Paper 76, 2006, [Online]. Available: http://repository.cmu.edu/hcii/76/.

[8] A. P. E. Rosiello, E. Kirda, C. Kruegel, and F. Ferrandi, "A layout-similarity-based approach for detecting phishing pages," *Proceedings of the 3rd International Conference on Security and Privacy in Communication Networks, SecureComm*, pp. 454–463, 2007, doi: 10.1109/SECCOM.2007.4550367.

[9] S. Al-Ahmadi and Y. Alharbi, "A Deep Learning Technique For Web Phishing Detection Combined Url Features And Visual Similarity," *International Journal of Computer Networks and Communications*, vol. 12, no. 5, pp. 41–54, 2020, doi: 10.5121/ijcnc.2020.12503.

[10] PhishTank, "PhishTank Dataset," 2020. http://www.phishtank.com/index.php (accessed Aug. 23, 2020).

[11] OpenPhish, "Phishing Feeds from OPenPhish." 2021, [Online]. Available: https://openphish.com/.

[12] University of New Brunswick, "URL datasets (ISCX-URL2016)," 2016. https://www.unb.ca/cic/datasets/url-2016.html (accessed May 05, 2021).

[13] A. Niakanlahiji, B.-T. Chu, and E. Al-Shaer, "PhishMon: A machine learning framework for detecting phishing webpages," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2018, pp. 220–225.

[14] T. Chandrasegar and P. Viswanathan, "Dimensionality reduction of a phishing attack using decision tree classifier," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2019, vol. 1, pp. 1–4.

[15] A. Odeh, I. Keshta, and E. Abdelfattah, "PHIBOOST-a novel phishing detection model using Adaptive boosting approach," *Jordanian Journal of Computers and Information Technology*

(JJCIT), vol. 7, no. 01, 2021.

[16] M. Almseidin, A. A. Zuraiq, M. Al-Kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," 2019.

[17] V. E. Adeyemo, A. O. Balogun, H. A. Mojeed, N. O. Akande, and K. S. Adewole, "Ensemble-Based Logistic Model Trees for Website Phishing Detection," in *International Conference on Advances in Cyber Security*, 2020, pp. 627–641.

[18] UCI, "UCI Machine Learning Repository," *UCI Irvine Machine Learning Repository*, 2021. https://archive.ics.uci.edu/ml/index.php (accessed May 27, 2021).

[19] S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," in *2020 3rd International Conference on Computer Applications \& Information Security (ICCAIS)*, 2020, pp. 1–6.

[20] V. Shahrivari, M. M. Darabi, and M. Izadi, "Phishing Detection Using Machine Learning Techniques," *arXiv preprint arXiv:200911116*, 2020.

[21] G. Sonowal and K. S. Kuppusamy, "Mmsphid: a phoneme based phishing verification model for persons with visual impairments," *Information \& Computer Security*, 2018.

[22] C.-Y. Wu, C.-C. Kuo, and C.-S. Yang, "A phishing detection system based on machine learning," in *2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, 2019, pp. 28–32.

[23] M. M. Yadollahi, F. Shoeleh, E. Serkani, A. Madani, and H. Gharaee, "An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features," in *2019 5th International Conference on Web Research (ICWR)*, 2019, pp. 281–286.

[24] C. D. Xuan, H. D. Nguyen, and T. V Nikolaevich, "Malicious URL Detection based on Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 148–153, 2020.

[25] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Bindhumadhava, "Phishing website classification and detection using machine learning," Jan. 2020, doi: 10.1109/ICCCI48352.2020.9104161.

[26] Eugenio Zuccarelli, "Performance Metrics in ML," 2020. https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92 (accessed May 05, 2021).