Comparison of Machine Learning Techniques for URL-based Phishing URLs Detection

Amr Jadi

Department of Computer Science and Information, College of Computer Science and Engineering, University of Ha'il, Ha'il, Saudi Arabia

a.jadi@uoh.edu.sa

Abstract:
Phishing attacks are a type of cybercrime that exploit network vulnerabilities to fraudulently obtain sensitive data from victims. These attacks often involve sending malicious emails or pretending to provide important information. In this research, the focus was on online phishing, and the dataset used had 10,000 instances (rows) and 50 attributes. The domains used in the phishing attacks were collected during two periods: January to May 2015 and May to June 2017. To analyze online phishing, the researchers employed three machine learning algorithms: logistic regression (LR), Random Forest (RF), and AdaBoost. These algorithms were chosen based on recommendations from previous researchers. However, in this research, principal component analysis (PCA) was not used since it was solely utilized for feature selection purposes. Before applying feature scaling and PCA, the initial accuracy scores obtained by the models were 93% for LR, 98% for RF, and 97% for AdaBoost. After feature scaling and PCA were applied, the accuracy scores improved to 94% for LR, 96% for RF, and 95% for AdaBoost. Among the three models, the Random Forest (RF) algorithm yielded the best results when compared to LR and AdaBoost. In addition to standard machine learning techniques, the researchers also utilized ensemble techniques in their research. However, specific details about the ensemble techniques employed were not mentioned in the provided information. Overall, the research focused on exploring online phishing using LR, RF, and AdaBoost algorithms. The accuracy of these models was evaluated before and after applying feature scaling and PCA. The Random Forest algorithm exhibited the highest accuracy among the three methods.

Keywords*:* *Phishing attacks, feature selection, Logistic Regression (LR), AdaBoost, and Random Forest Model (RF).*

1. INTRODUCTION

Phishing scams are yet another form of cybercrime that can end numerous problems, along with the theft of the victim's personal information (Ghazi and Pontell, 2021). The majority of these take the form of emails that assault the victim's machine with a virus or false alerts that deceive the victim into believing they are real. Both legal and phishing websites use data gathered by Ghazi and Pontell, (2021). Strong methods were used to retrieve this data using more efficient frameworks, including Selenium. The author advises using the data to analyze the effects and provide support for ideas on how to prevent phishing. Researchers can utilize this information to do data analysis for phishing-based categorization as well (Ghazi and Pontell, 2021).

Cybercrimes, which involve the theft of a person's personal information by attackers to compromise financial data, are on the rise in today's digital age. According to Tulkarm et al., attackers use a variety of techniques, but social engineering crime is the most often used technique in cyberattacks (Tulkarm, 2021). The phishing attack refers to these social engineering crimes committed by the perpetrators. According to Kumar et al., the real website's domain name and look are identical to those of the phishing website. The lack of knowledge about these virtual worlds is the major cause of these attacks (Kumar *et al.*, 2020). Phishing attempts on websites or web pages rob users of their login information. When the user clicks on a link, opens an email, speed message, or text message, this takes place. The phishing assaults on links or websites that have similar domain names and are unable to distinguish between the original link and the duplicate link (Das *et al.*, 2022). Stealing information through illegal actions is considered a form of social engineering attack (Kajave and Nismy, 2022). These attacks also cause commercial loss along with the invasion of personal bulk information of a business firm details. These attacks also come under cyber threats and these are one of the global threats which need attention (Chaudhari et al., 2021). According to Campina and Rodrigues, there are 611,877 sites identified in 2021 that were unique, and 241,342 victims in the year 2020 fall to the relevant threat of Phishing through data stealing, identity theft, and other methods as well. These attacks often occur and cover most of the breaches occurring in the data industry (Kalabarige *et al.*, 2022).

Here the author uses ML techniques because the advantage of using ML in this project is that it can provide more accurate predictions without requiring explicit instructions. Instead of programming specific rules or conditions, ML algorithms can learn from historical data and generalize patterns to make predictions on new data. This makes it suitable for tasks where there may be complex relationships or patterns that are difficult to define explicitly. ML algorithms can handle various types of data, including numerical, categorical, and text data, which makes them versatile for different domains and problem types (Tulkarm and Kumar). By leveraging ML, the project can take advantage of its ability to automatically extract relevant features from the data, reducing the manual effort required in feature engineering. Furthermore, ML algorithms can adapt and improve over time as they receive more data and feedback. This allows the project to continuously enhance its predictive capabilities as new information becomes available. The syntax of ML languages is often straightforward and accessible to a wide range of users. While some ML algorithms and concepts may require a deeper understanding, there are also user-friendly libraries and frameworks available that simplify the implementation process (Chaudhari et al., 2021). This allows researchers and practitioners from various backgrounds to leverage ML techniques effectively. Considering the success of previous studies and the potential benefits it offers, implementing ML techniques in the suggested project is a promising approach. However, it's important to note that the success of ML relies on various factors, including the quality and representativeness of the data, appropriate algorithm selection, and proper evaluation and validation processes.

These concerns grow day by day that how can these threats be approached and mitigated. One such approach can be through the utilization of technology apart from human awareness about the threats (Ghazi and Pontell, 2021, Kalabarige *et al.*, 2022). The computer-based algorithms are the trend to be applied to many types of fields. The websites and URLs (Unique Resource Locator) are manipulated such that the illegal websites look like legitimate ones. Hence, the motivation of the work is to utilize machine learning (ML) to focus here on classifying these attacks from normal. Hence, the study aims to employ machine learning (ML) principles and data-related methodologies, to categorize Phishing attacks. The overall objectives of the research are as follows,

- To use the packages like Pandas and NumPy to clean and manipulate the data and properties.
- To analyze the dataset's properties and their correlation.
- To use ensemble and classifier approaches for machine learning (like boosting, bagging, and random forest).
- Using measures such as evaluating recall, precision, and accuracy coupled with a confusion matrix to assess the models' performance. It will also aid in demonstrating the model's dependability.

Concerns for evolving digital technology will also become an issue for the threat happening around. Today phishing websites have become a most important cybersecurity threat (Kumar *et al.*, 2020). Ransomware, malware, spam, and so on are all hosted by phishing websites. Also, explained that the phishing website looks very much similar to any well-known website and traps innocent users to fall victim to such traps. Further suggested is that some better solutions to resolve such security threats are necessary, as earliest as possible. Also explained that the traditional method employed for identification and classification is performed using blacklists. The failure of such a method is because of two aspects, the blacklists may not be thorough and do not identify the freshly built phishing website [Ghazi and

Pontell, 2021, Kumar *et al*., 2020]. Their study employed several techniques like Gaussian naïve Bayes (NB), random forest (RF) and logistic regression (LR), decision tree (DT), and k-nearest neighbour (KNN) and reached an accuracy of 97.18%, 98.03% (highest), 97.7%, 98.02%, and 97.99% respectively (Kumar *et al*., 2020). A similar study on phishing website identification was performed by Weedon et al. utilized models such as LR, NB, RF, and J48 and the accuracy reached were 81.5%, 64.6%, 86.9%, and 83.9% respectively. The highest accuracy of 86.9% was reached by RF. The author opines from the above discussion that utilizing the RF algorithm would be better as seen they were able to reach higher accuracy when compared to other modes (Das *et al*., 2022). This research also employed the RF along with LR and adaptive boosting (AdaBoost) and their respective accuracies reached post-feature scaling and principal component analysis (PCA) are 96%, 94%, and 95%. In this work, the RF model provided better accuracy.

A similar study on the identification of phishing uniform resource locators (URLs) was also performed by Parmar (2020), who explained that the traditional methods for identifying these phishing URLs were mainly based on signature-based and blacklisting methods. But these approaches consume time and are not that suitable for a new set of URLs. Their work utilized machine learning (ML) models such as bagging, RF, AdaBoost, and gradient boosting techniques; further, they compared these outcomes with non-ensemble learning techniques like DT, KNN, and LR. The accuracies obtained were 95.61% for bagging, 93.26% for AdaBoost, 96.15% for RF, 92.49% for gradient boosting, 90.5% for DT, 54.31% for KNN, 92.4% for LR. However, RF performed better when compared to other algorithms. A similar study on the detection of phishing websites was performed by Parekh et al. employed the RF technique and the highest accuracy of 95% was reached by the RF model. The author also suggests from the above discussions that the RF model is suitable to employ for phishing website detection and also helps in providing better accuracy (Kajave and Nismy, 2022). This work also implemented the RF model along with LR and AdaBoost and the obtained accuracies are 96%, 94%, and 95% respectively, these accuracies obtained are after employing feature scaling and PCA techniques.

A recent study on phishing website identification conducted by Chaudhari et al. explained that the phishing technique may tend to distract users into some distrustful content websites and steal important information. Also, explained that preventing such threats is important by utilizing anti-phishing mechanisms for identifying phishing at the source. In their study, they checked the correlation amongst the features by using cross-validation, and by utilizing the extra tree classifier they detected the feature importance and the accuracy reached was 96.57%. The tested models were RF, extreme gradient boosting (XGBoost), DT, KNN, LR, support vector machine (SVM), and AdaBoost and the accuracy reached were 96.78%, 96.83%, 95.66%, 94.71%, 92.30%, 94.12%, 91.13% respectively (Chaudhari *et al*., 2021). Also, they applied the synthetic minority oversampling technique (SMOTE) and PCA techniques, especially for XGBoost and RF models, and the accuracy reached was the same, the accuracies for these models were more compared to other utilized models. A similar study was done by Subasi and Kremic, for phishing webpage detection and employed models were RF, SVM, AdaBoost, and other models. The accuracy reached by RF is 97.26%, and SVM is 96.42%. When considering together that is AdaBoost with SVM the accuracy reached was 97.61% which was the highest, and the RF with AdaBoost combination reached an accuracy of 97.30% (Kalabarige *et* al., 2022). The author opines on utilizing AdaBoost, RF, and LR techniques. However, post employing the feature scaling and PCA, the accuracy reached was 95%, 96%, and 94% respectively for these AdaBoost, RF, and LR models.

There is an increase, especially in internet attacks, and their research focused mainly on detecting malicious URLs and utilized a blend of URL lexical options, python supply options, and payload size. They implemented models like SVM with the polynomial kernel and LR, and their obtained accuracy was 98% (Naresh *et al*., 2020). Recently, a related study on phishing websites was done by Jansi employed the ML technique such as LR, which helps to identify fresh web links that are fraudulent and, in their study, obtained an accuracy of 95% (Jansi, 2021). The author opines utilizing the LR technique in this study related to web URL phishing classification and along with this also implemented RF and AdaBoost. After utilizing approaches such as feature scaling and PCA the accuracies reached in this work for the models utilized are 94%, 96%, and 95% respectively.

Recently research on phishing website identification was also conducted by Lakshmanarao et al. explained that websites are the main resource for cyber-attacks, and this is because of the improvements in internet technology. Explained that ML techniques help resolve the problems in cybersecurity; as most phishing attacks have characteristics in common. Their study proposed a total of two priority-based techniques and based on these approaches, the final fusion classifier is considered. Post applying the fusion classifier they were able to obtain an accuracy of 97% (Lakshmanarao *et al*., 2021). The author opines using the AdaBoost, LR, and RF models in this work related to web URL phishing classification. In this research work, the accuracy obtained after implementing feature scaling and PCA techniques is 94% for LR, 95% for AdaBoost, and 96% for RF (highest) model.

Hence, addressing the aim and objectives of the work, Section 2 details the methodology and the step-by-step implementation of the process. Section 3 shows the experimental results with the results and parameters considered for the research through figures. Finally, Section 4 details the conclusions drawn from the overall results.

2. RELATED WORK

URL-based phishing is a popular topic for research because online fraud is one of the biggest problems these days. For this reason, many researchers and programmers are doing a lot of research and analysis to provide valuable information. (*Tultul etal*., 2022) Many software companies develop tools to avoid attacks because it is very difficult to detect whether Internet resources are real or phishing, but some products can do so. Using a support vector machine, random forest algorithms, k-nearest neighbor, and a decision tree, this work introduces an anti-phishing system that can instantly extract information from a website's URL and classify those features. When the experimental results of other methods are examined, the random forest approach with selected features outperforms the others, with an accuracy rate of 95.67%. Then, the author added a deep learning method and the Deep Neural Decision Forest to their experiment, which succeeded with an accuracy rate of 92.67%. However, (Sánchez *et al*., 2022) the author proposed that a form of malware or hack known as phishing must deceive victims into entering their login details on a form that transfers the data to an adversarial website. To develop a method for phishing website identification using URL analysis, they compare machine learning and deep learning methodologies in this work. It has created a fresh set of data called Phishing Index Login URL (PILU-90K), which has been made up of 30K legitimate URLs, like login and index pages, and 60K phishing URLs, to support these assertions. Finally, they achieved 96.50 % accuracy with the help of TF-IDF.

This research involved extensive feature engineering and analysis to identify the best features for dangerous URL predictions, test many models, and achieve high accuracy using a big new URL dataset (Aljabri *et al*., 2022). Correlation estimation, analysis of variance (ANOVA), and chi-square were just a few of the feature selection techniques the author employed to extract the most distinctive qualities from the dataset. Finally, they carried out a comparative evaluation of the performance of a variety of ML and DL models using a set of standard criteria for evaluating such models. The most accurate model for identifying harmful URLs using the analyzed data was Nave Bayes (NB), with a 96% accuracy rate. Similarly, (Chandra *et al*., 2022) This study examines different methods of machine learning for evaluating whether a URL is valid or a phishing URL with a particular property using a dataset for Web page phishing detection. The machine learning algorithms that were compared were K-Nearest Neighbour, Random Forest, Naive Bayes, Support Vector Machine, Decision Tree, and Logistic Regression. The models were trained on a phishing dataset that experienced preprocessing and encoding. It is noted and other evaluation criteria are contrasted with the output accuracy of the model. The results are similar, with the Random Forest algorithm reporting the highest accuracy of 98,04% out of 11429 URLs.

By studying these research papers, the author understands that no one tries to reduce the dimensionality that's why the author applies Principal Component Analysis (PCA) techniques to reduce the dimensionality of the dataset so that it executes in a faster way and gives the best accurate result without confusing the model.

3. METHODOLOGY

The data collected from the Kaggle site was proposed by Tan. The dataset is in CSV format and there are 50 characteristics or columns and 10,000 occurrences make up the dataset (Ghazi and Pontell, 2021). The properties of the dataset are the information about it. The information in the dataset is made up of 50 attributes out
of the 48 attributes with and without PCA

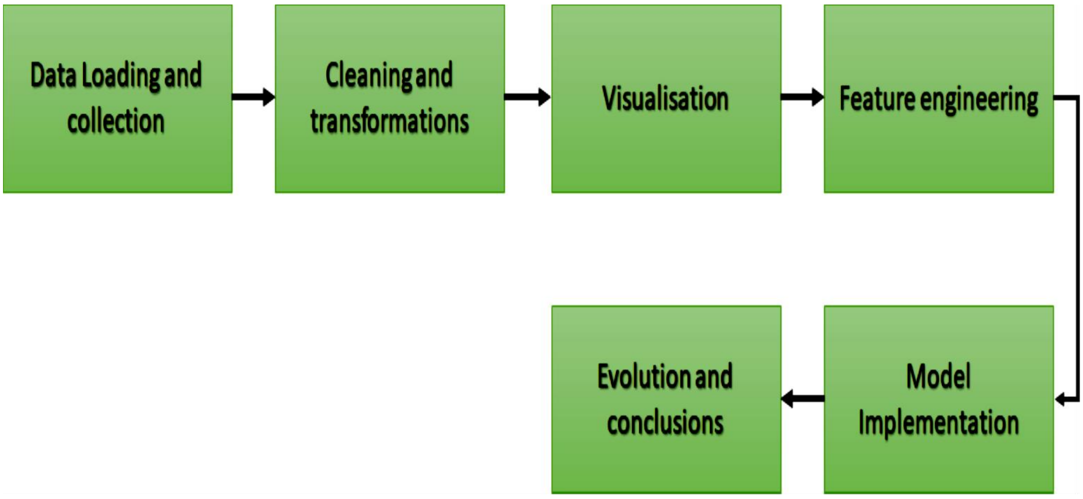

Data Loading and collection
Cleaning and transformations
Visualisation
Feature engineering
Model Implementation
Evolution and conclusions

Figure 1. Methodology of the study

The target variable among them is the class label. The data was loaded using Panda's library. The data set was considered for processing. The steps involved in getting the results are seen in Fig.1. According to Fig. 1, the data collected is considered for processing and cleaning. The data was balanced, which means that both the classes in the output label are with similar counts. The count of null values in the dataset is also zero. Then in the later stages, the data is considered for processing methods like label encoding, which is the method applied to convert the object type of data to integer type. Along with that, the unwanted columns are also removed. After the data is ready to be visualized. The visualization part will help to find the relation between the columns with the output label.

The model is implemented after the data is segregated for the input and output, then the models are trained on the training set which is 70% of the whole data and the remaining data will be considered for testing which is 30%. Then after the models are implemented. The evaluation is considered based on the accuracy, precision, and recall of the model testing.

3.1. Data Processing

Firstly, the id column in the dataset was removed as it looks like an index. There were no null values in the dataset and no major processing was required.

Based on the target column meaning dependent column this problem statement belongs to the binary classification problem because here only 0 and 1 exist in the target column. Zero represents there are fewer chances of a phishing attack and 1 represents no attacks. That's why there is no chance of performing multiclass classifications. Based on that here the author applies some of the algorithms that are selected by the author in different case studies of related research papers because previously many researchers get better results by applying these algorithms. These algorithms are discussed below:

3.2. Random Forest Model

Both classifier and regression issues may be solved using the Random Forest Model (RF), and the regression tasks are best suited for this technique. RF is utilized for numerous classifiers and ensemble approaches (Ahmad *et al.*, 2022). The combo of the decision trees is referred to as RF. The RF predicts the results of each tree instead of utilizing decision trees and then forecasts the overall output. According to Aamir and Zaidi (2021), the overfitting issue is diminished in this RF because of the large number of trees and increased accuracy (Aamir and Zaidi, 2021). Singh and Chahabra claim that entropy is chosen for the information gain in this parameter while determining the correctness of split trees. The test data are anticipated once the model is fitted to the train data. This RF classifier creates and defines new objects for prediction (Singh and Chhabra, 2021).

3.3. AdaBoost Classifier

Adaptive boosting is referred to as Adaboost. The machine learning algorithm uses the Ada Boost methodology, which belongs to the ensemble method. Ogunseye et al., assert that Adaboost is utilized for classification and regression issues. One of the boosting methods is called the Ada boost (Ogunseye *et al.*, 2022). Decision trees are utilized in this model, and it is divided just once. The many weak classifiers are transformed into the strong classifier using Adaboost. According to Barsacchi *et al.*, the decision tree and Adaboost are the most often employed algorithms for classification issues (Barsacchi *et al.*, 2020). This paradigm allows for the transformation of all weak learners into one strong learner. According to Zhang *et al.*, it is used to increase accuracy and stability and to lessen overfitting issues. Instead of binary classification issues, this paradigm is mostly utilized for text and picture categorization (Zhang *et al.*, 2022).

3.4. Logistic Regression

Undoubtedly one of the most well-known machine learning techniques is logistic regression (LR). LR belongs to the category of supervised machine learning. This approach is used to forecast the Y value, which falls under the dependent variable, among the X values, which are the independent variables (Jiang et *al.*, 2021). LR is a classification approach that uses discrete or categorical values, such as 0 or 1, yes or no, as its inputs (Ifraz *et al*., 2021). The categorical dependent variable is specified to be X. Although this technique is comparable to linear regression, logistic regression is utilized for situations involving categorization.

3.5. Feature scaling and dimensionality reduction

Maharjan *et al.* assert that weight has higher physical characteristics than height. The value of each characteristic is divided by the standard deviation and removed from the mean since the mean has a value of zero and the standard deviation is one (Maharjan *et al*., 2021). According to Malan *et al.*, standardization should be utilized before PCA (principal component analysis) is employed to standardize the data. The dimensionality is decreased and the vast data is converted into smaller ones using this PCA (Malan *et al.*, 2020). Implemented models in the present study included LR, RF, and AdaBoost. Utilized methods like feature scaling and PCA to improve accuracy. The hazards related to ethics and business are also highlighted. used techniques including accuracy, precision, and recall for assessment. Before using feature scaling and PCA approaches, the accuracy for the RF, LR, and AdaBoost models was 98%, 93%, and 97%, respectively. However, after applying feature scaling (standardization) and PCA, the RF achieved better results with an accuracy of 96%, while the LR and AdaBoost approach achieved an accuracy of 94% and 95%, respectively (dimensionality reduction).

3.6. Validation metrics

Confusion matrices are particularly useful for classification issues. Both binary and multiple-class classifications use it. The actual and anticipated values for this matrix are shown. According to Rahmad *et al*., the matrix's output is made up of four values true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (FN). The measurements determine how accurate the confusion matrix is. The divide of the total of the true positive and true negative values by the sum of the other values is used to determine accuracy (Rahmad *et al*., 2020). The accuracy of the confusion matrix is not accurate for the unbalanced datasets. The sklearn library is utilized for the confusion matrix.

The positive samples are measured using precision. The split of the true positive to the total of the true positive and the false positive is how it is defined. The model can provide accurate predictions because of the accuracy (Hofstätter et al., 2021). The primary goal of precision is to correctly categorize positive samples as positive and incorrectly label negative samples as negative. Hofstatter *et al.* claims that the recall is only dependent on the positive samples and does not depend on the negative ones. Since the evaluated model must contain the right predictions of the positive samples in order for accuracy to be attained, it solely depends on positive samples (Rahmad *et al*., 2020, Hofstätter et al., 2021).

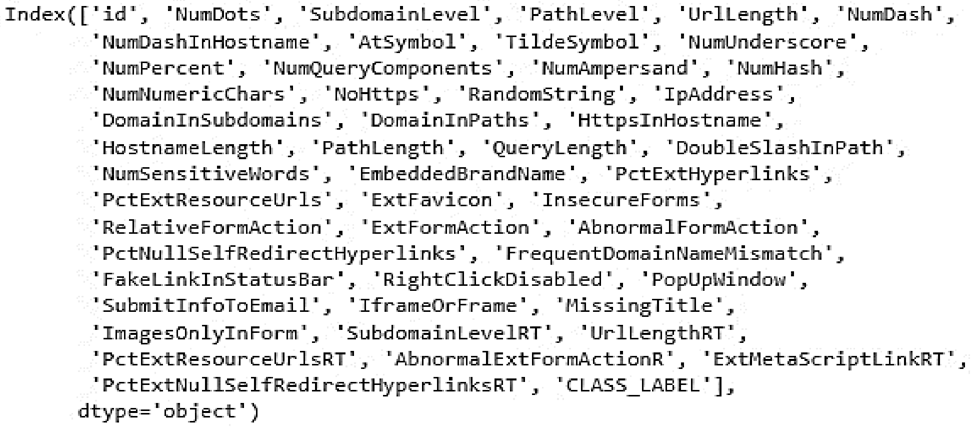
4. EXPERIMENTAL RESULTS

The main takeaways from the dataset are:

- An understanding of the many kinds of phishing attempts and how they target certain people from the dataset (Ghazi-Tehrani and Pontell, 2021).
- The phishing assaults are divided into eight groups based on various methodologies. There are seven methods to recognize a phishing assault, including poor English language, misspellings, irregular connections, and domain names.
- The 48 characteristics that make up the phishing dataset were gathered between January and May 2015 and May and June 2017. This dataset, which belongs to the classification model used to identify phishing assaults, has 1000 instances and 50 characteristics. The class label serves as the target variable.
- The dataset's characteristics include text, symbols, domain information, and hypertext links that indicate whether or not the connections are phishing type. The 10 columns must be chosen since there are more attributes to derive the correlation matrix from.
- Because this dataset falls into the classification category, three machine learning models—Logistic regression, Random Forest Classifier, and Ada (Adaptive boosting machine).
- The dataset contains certain categorical characteristics and utilizing machine learning techniques, the pycaret classification is applied to turn the pieces into binary.

4.1. Dataset Description

All the attribute names are given in Fig. 2 and the target variable among them is the class label as seen in Fig. 2. The brief description of the dataset is already given methodology section.



```
Index(['id', 'NumDots', 'SubdomainLevel', 'PathLevel', 'UrlLength', 'NumDash',
       'NumDashInHostname', 'AtSymbol', 'TildeSymbol', 'NumUnderscore',
       'NumPercent', 'NumQueryComponents', 'NumAmpersand', 'NumHash',
       'NumNumericChars', 'NoHttps', 'RandomString', 'IpAddress',
       'DomainInSubdomains', 'DomainInPaths', 'HttpsInHostname',
       'HostnameLength', 'PathLength', 'QueryLength', 'DoubleSlashInPath',
       'NumSensitiveWords', 'EmbeddedBrandName', 'PctExtHyperlinks',
       'PctExtResourceUrls', 'ExtFavicon', 'InsecureForms',
       'RelativeFormAction', 'ExtFormAction', 'AbnormalFormAction',
       'PctNullSelfRedirectHyperlinks', 'FrequentDomainNameMismatch',
       'FakeLinkInStatusBar', 'RightClickDisabled', 'PopUpWindow',
       'SubmitInfoToEmail', 'IframeOrFrame', 'MissingTitle',
       'ImagesOnlyInForm', 'SubdomainLevelRT', 'UrlLengthRT',
       'PctExtResourceUrlsRT', 'AbnormalExtFormActionR', 'ExtMetaScriptLinkRT',
       'PctExtNullSelfRedirectHyperlinksRT', 'CLASS_LABEL'],
      dtype='object')
```

Figure 2. Columns in the dataset

4.2. Data Visualization

The letter "#" stands for hash. Hash is described as the elements that contain a string of characters that are converted into values. It may include numbers, pounds, or the hash key. Internet protocol is referred to as an IP address. The identification of the internet or a local network is done by its IP address as seen in Fig. 3 according to the output label in the dataset. The number of sensitive words in the URL is also utilized to segregate the phishing website (Aljofey et *al*., 2022).

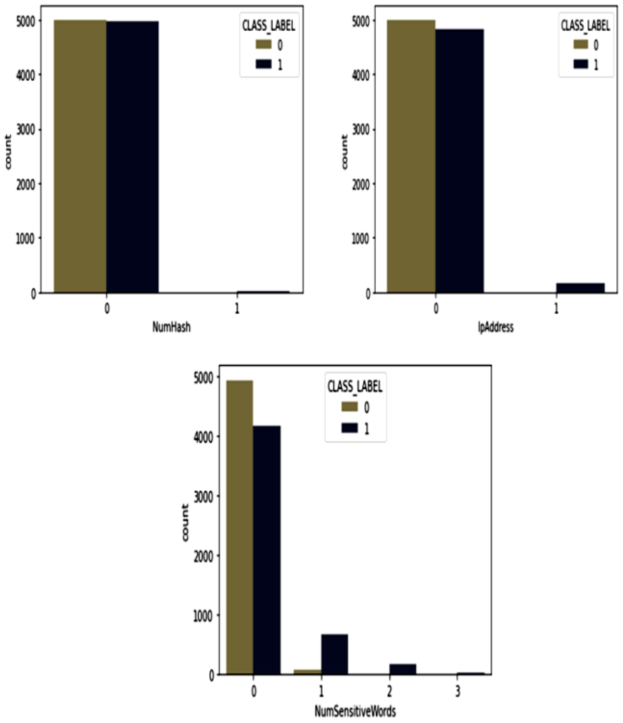
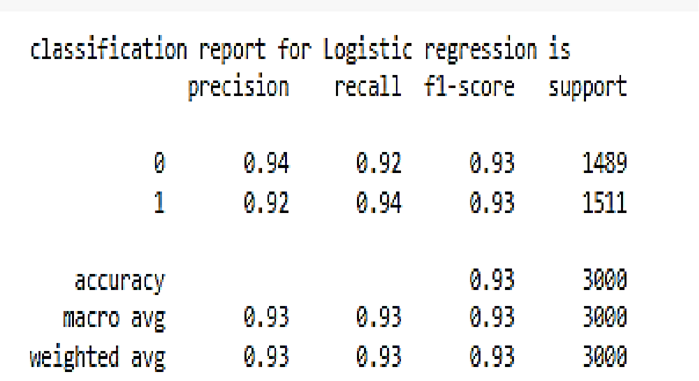


Figure 3. Graphical Representation of columns Number Hash, IP address and Number of Sensitive words in the URL.

4.3. Model implementation

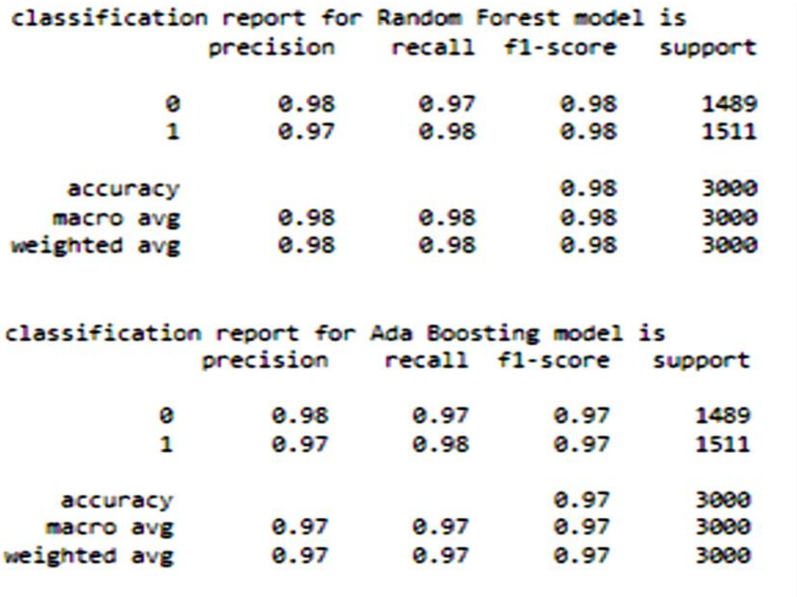
The number of rows and columns is detailed by the shape function and there are 10,000 rows and 48 columns (including the output label). Along with the number of rows and columns for the training and test data, the total number of rows and columns is 7000 rows for training and 3000 rows for testing. The information is divided into train and test data. The package sklearn linear model's logistic regression is imported. The model is fitted to the train data and the classifier is defined. The test results are anticipated. The confusion matrix and classification report are obtained from the package sklearn metrics. Precision, recall, and f1-score are retrieved from the classification report.



```
classification report for Logistic regression is
              precision    recall  f1-score   support

           0       0.94      0.92      0.93      1489
           1       0.92      0.94      0.93      1511

    accuracy                           0.93      3000
   macro avg       0.93      0.93      0.93      3000
weighted avg       0.93      0.93      0.93      3000
```

Figure 4. Classification Report of Logistic Regression (LR)

The accuracy for the logistic regression is 93%, as can be shown in Fig. 4. The f1-score was 93, the recall was 94%, and the accuracy was 92% as seen in Fig. 4. The random forest classifier achieved accuracy of 98%, precision of 97%, recall of 98%, and f1 score of 98%. The Adaboost achieved an accuracy of 97%, precision of 97%, recall of 98%, and f1score of 97% as seen in Fig. 5.



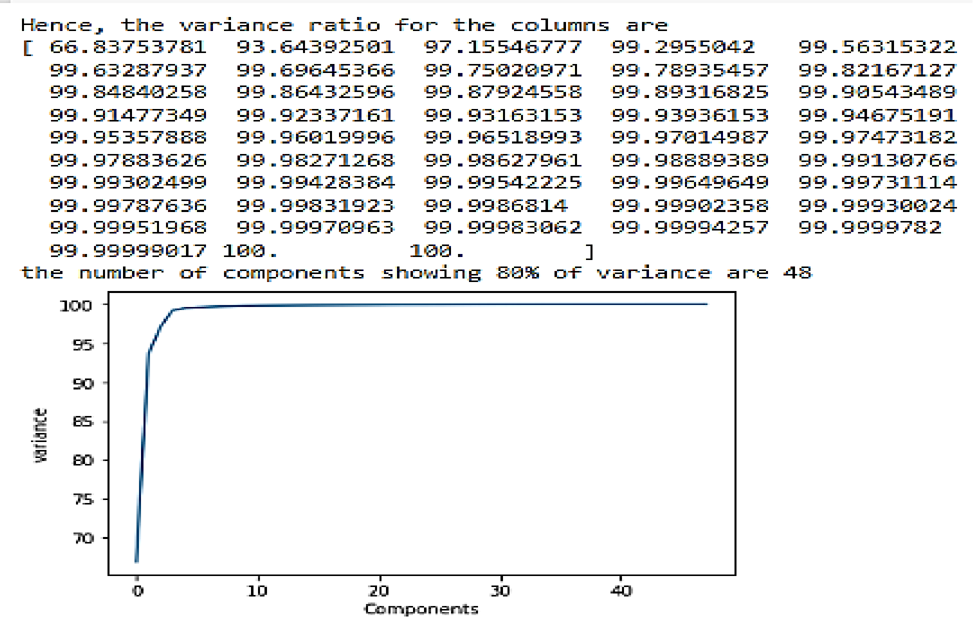
```
classification report for Random Forest model is
              precision    recall  f1-score   support

           0       0.98      0.97      0.98      1489
           1       0.97      0.98      0.98      1511

    accuracy                           0.98      3000
   macro avg       0.98      0.98      0.98      3000
weighted avg       0.98      0.98      0.98      3000

classification report for Ada Boosting model is
              precision    recall  f1-score   support

           0       0.98      0.97      0.97      1489
           1       0.97      0.98      0.97      1511

    accuracy                           0.97      3000
   macro avg       0.97      0.97      0.97      3000
weighted avg       0.97      0.97      0.97      3000
```

Figure 5. Classification Report of RF and Adaboost algorithm

Identify applicable funding agency here. If none, delete this text box.

4.4. Feature Engineering

In the PCA, the n components are utilized to minimize the data while reducing dimensionality. These n components are the uncorrelated variables and maximize the variance. The n components are chosen to be 48 from Fig. 6 since it decreases the size. The PCA is fitted to the input data in order to transform the data. The input is shown using the shape function.

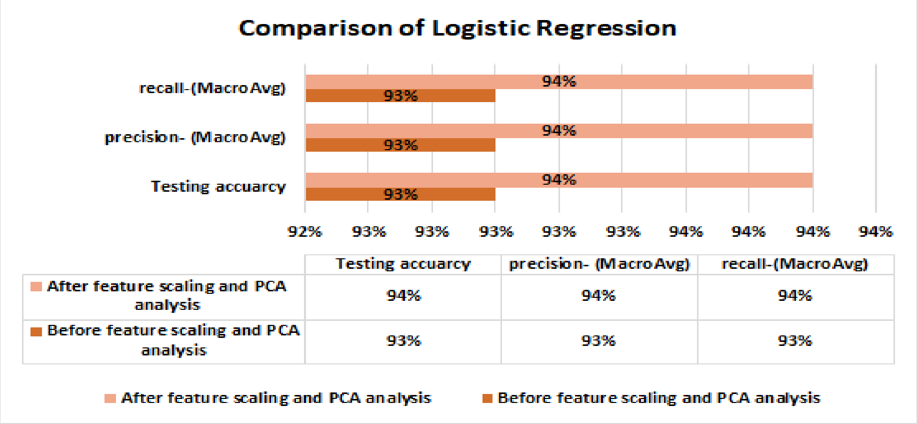


```
Hence, the variance ratio for the columns are
[ 66.83753781  93.64392501  97.15546777  99.2955042   99.56315322
  99.63287937  99.69645366  99.75020971  99.78935457  99.82167127
  99.84840258  99.86432596  99.87924558  99.89316825  99.90543489
  99.91477349  99.92337161  99.93163153  99.93936153  99.94675191
  99.95357888  99.96019996  99.96518993  99.97014987  99.97473182
  99.97883626  99.98271268  99.98627961  99.98889389  99.99130766
  99.99302499  99.99428384  99.99542225  99.99649649  99.99731114
  99.99787636  99.99831923  99.9986814   99.99902358  99.99930024
  99.99951968  99.99970963  99.99983062  99.99994257  99.9999782
  99.99999017 100.         100.        ]
the number of components showing 80% of variance are 48
```

Figure 6. Application of PCA (Principle Component Analysis) analysis

4.5. Comparative analysis

The algorithms were considered for implementation before and after PCA analysis is applied.



	Testing accuarcy	precision- (MacroAvg)	recall-(MacroAvg)
After feature scaling and PCA analysis	94%	94%	94%
Before feature scaling and PCA analysis	93%	93%	93%

Figure 7. Comparative analysis of the LR model before and after PCA

The Logistic Regression can serve as an example of a single classifier, and it has achieved an accuracy of 94% both before and after the implementation of feature selection; however, after the implementation of feature selection and PCA analysis, the accuracy of the Logistic Regression increased by 1%, to 94% as seen in Fig. 7.

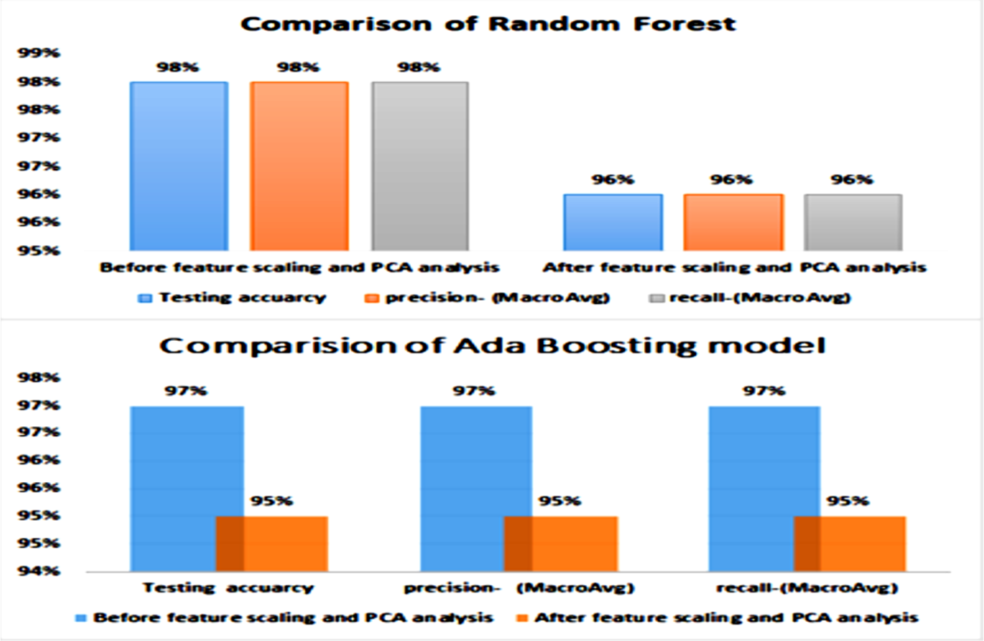


Figure 8. Comparative analysis of RF and Adaboost before and after PCA.

With regard to the random forest model, it had an overall accuracy of 98% along with precision and recall, but when PCA analysis was used, this accuracy decreased to 96% along with precision and recall. The same thing happened with the Ada boosting model, an improved version of the decision tree, which had an overall accuracy of 97% as seen in Fig. 8.

5. CONCLUSIONS

The accuracy of the Logistic Regression, which may be used as an example of a single classifier, was 94% both before and after feature selection was implemented; however, after feature selection and PCA analysis were implemented, the accuracy of the Logistic Regression rose by 1%, to 94%. When using PCA analysis, the random forest model's accuracy dropped to 96% along with precision and recall but had an overall accuracy of 98% with precision and recall. The Ada boosting model, an enhanced decision tree that achieved a 97% overall accuracy, experienced the same issue. As a result, it can be said that based on the data set that was given to the model, Logistic regression performed with 94% accuracy, while the other two models performed with 97 and 98 percent accuracy along with Precision and recall, demonstrating that both performance and reliability of the machine learning model are very high and in line with other authors and other studies' standards. Then, if the dataset is acceptable for creating the best DL approaches, it may deliver the greatest results, meaning it will forecast that URL if there is a probability of cyberattacks.

Some of the limitations of these models are discussed here, as this dataset is older which is why the machine was also trained in an older way there is a chance that in modern day new variations of phishing attacks may come so at that time this model maybe not be useful. That's why it is recommended to update the model with training new patterns of data so that it can able to identify modern phishing patterns also.

References

Aamir, M. and Zaidi, S.M.A., 2021. Clustering based semi-supervised machine learning for DDoS attack classification. Journal of King Saud University-Computer and Information Sciences, 33(4), pp.436-446.

Ahmad, A., Akbar, S., Tahir, M., Hayat, M. and Ali, F., 2022. iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach. Chemometrics and Intelligent Laboratory Systems, 222, p.104516.

Aljabri, M., Alhaidari, F., Mohammad, R.M.A., Mirza, S., Alhamed, D.H., Altamimi, H.S. and Chrouf, S.M., 2022. 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.

Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q. and Wang, Y., 2022. An effective detection approach for phishing websites using URL and HTML features. Scientific Reports, 12(1), p.8842.

Barsacchi, M., Bechini, A. and Marcelloni, F., 2020. An analysis of boosted ensembles of binary fuzzy decision trees. Expert Systems with Applications, 154, p.113436.

Chandra, A., Immanuel, M.J. and Gunawan, A.A.S., 2022, August. Accuracy Comparison of Different Machine Learning Models in Phishing Detection. In 2022 5th International Conference on Information and Communications Technology (ICOIACT) (pp. 24-29). IEEE.

Chaudhari, M.S.S., Gujar, S.N. and Jummani, F., Detection of Phishing Web as an Attack: A Comprehensive Analysis of Machine Learning Algorithms on Phishing Dataset.

Das Guptta, S., Shahriar, K.T., Alqahtani, H., Alsalman, D. and Sarker, I.H., 2022. Modeling hybrid feature-based phishing websites detection using machine learning techniques. Annals of Data Science, pp.1-26.

Ghazi-Tehrani, A.K. and Pontell, H.N., 2021. Phishing evolves: Analyzing the enduring cybercrime. Victims & offenders, 16(3), pp.316-342.

Hofstätter, S., Lin, S.C., Yang, J.H., Lin, J. and Hanbury, A., 2021, July. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 113-122).

Ifraz, G.M., Rashid, M.H., Tazin, T., Bourouis, S. and Khan, M.M., 2021. Comparative analysis for prediction of kidney disease using intelligent machine learning methods. Computational and Mathematical Methods in Medicine, 2021.

Jansi, K.R., 2021. An Effective Model of Terminating Phishing Websites and Detection Based On Logistic Regression. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), pp.358-362.

Jiang, F., Zhidong, G.U.A.N., Zengshan, L.I. and Xiaodong, W.A.N.G., 2021. A method of predicting visual detectability of low-velocity impact damage in composite structures based on logistic regression model. Chinese Journal of Aeronautics, 34(1), pp.296-308.

Kajave, A. and Nismy, S.A.H., 2022. How Cyber Criminal Use Social Engineering to Target Organizations. arXiv preprint arXiv:2212.12309.

Kalabarige, L.R., Rao, R.S., Abraham, A. and Gabralla, L.A., 2022. Multilayer stacked ensemble learning model to detect phishing websites. IEEE Access, 10, pp.79543-79552.

Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B. and Bindhumadhava, B.S., 2020, January. Phishing website classification and detection using machine learning. In 2020 international conference on computer communication and informatics (ICCCI) (pp. 1-6). IEEE.

Lakshmanarao, A., Rao, P.S.P. and Krishna, M.B., 2021, March. Phishing website detection using novel machine learning fusion approach. In 2021 international conference on artificial intelligence and smart systems (ICAIS) (pp. 1164-1169). IEEE.

Maharjan, S.K., Sterck, F.J., Dhakal, B.P., Makri, M. and Poorter, L., 2021. Functional traits shape tree species distribution in the Himalayas. Journal of Ecology, 109(11), pp.3818-3834.

Malan, L., Smuts, C.M., Baumgartner, J. and Ricci, C., 2020. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. Nutrition Research, 75, pp.67-76.

Naresh, R., Gupta, A. and Giri, S., 2020. Malicious url detection system using combined sym and logistic regression model. International Journal of Advanced Research in Engineering and Technology (IJARET), 11(4).

Ogunseye, E.O., Adenusi, C.A., Nwanakwaugwu, A.C., Ajagbe, S.A. and Akinola, S.O., 2022. Predictive analysis of mental health conditions using AdaBoost algorithm. ParadigmPlus, 3(2), pp.11-26.

Rahmad, F., Suryanto, Y. and Ramli, K., 2020, July. Performance comparison of anti-spam technology using confusion matrix classification. In IOP Conference Series: Materials Science and Engineering (Vol. 879, No. 1, p. 012076). IOP Publishing.

Sánchez-Paniagua, M., Fernández, E.F., Alegre, E., Al-Nabki, W. and Gonzalez-Castro, V., 2022. Phishing URL detection: A real-case scenario through login URLs. IEEE Access, 10, pp.42949-42960.

Singh, M. and Chhabra, J.K., 2021. WITHDRAWN: EGIA: A new node splitting method for decision tree generation: Special application in software fault prediction.

Tulkarm, P., 2021. A Survey of Social Engineering Attacks: Detection and Prevention Tools. Journal of Theoretical and Applied Information Technology, 99(18).

Tultul, A.N., Afroz, R. and Hossain, M.A., 2022. Comparison of the efficiency of machine learning algorithms for phishing detection from uniform resource locator. Indonesian Journal of Electrical Engineering and Computer Science, 28(3), p.1640.

Zhang, C., Ding, S. and Ding, L., 2022, May. An AdaBoost Based-Deep Stochastic Configuration Network. In International Conference on Intelligent Information Processing (pp. 3-14). Cham: Springer International Publishing.