

URL Based Phishing Detection using Machine Learning

B Swarna Jyothi¹, M Akshaya², K Anjum³, A Bhavana⁴ and K Sreemukha⁵
{ badimela1508@gmail.com¹, madannagariakshaya@gmail.com², kothhapalliabdulsalam@gmail.com³,
amurubhavana@gmail.com⁴, sreemukhakondakamarla@gmail.com⁵ }

Assistant Professor, Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India¹
U.G. Student, Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science Madanapalle, Andhra Pradesh, India^{2, 3, 4, 5}

Abstract. Phishing website is still a serious problem, which try to impersonate a legitimate or official website to trick users into entering their personal and sensitive information. Thus, it has been the most significant task to detect such phishing web sites for Web Security. One of the approaches is detection of phishing through URL based that uses machine learning algorithms. In this paper we present a URL-based system using machine-learning models (SVM and Random Forest) for phishing detection. These classifiers are efficient in detecting malicious URLs by examining features like the size of the domain, suspicious characters and keywords [1]. Machine learning models, such as support vector machine (SVM) and RF, have achieved excellent performance in detecting phishing URLs. The Support Vector Machine (SVM) is a popular approach which can be employed in high dimensional feature space and has been shown to be effective for Phishing website detection [2]. The Random Forest model (a combination of multiple decision trees for predicting) has been applied for this task and achieved high precision in phishing URL detection as its weakness and large dataset handling problems in [3]. In this paper, we recommend a system for the detection of phishing, based on URL features, using machine learning: Support Vector Machine (SVM) and Random Forest classifiers. As a result, the classifiers are trained on a large labeled corpus of phishing and normal URLs in 83.3% of accuracy. The second classifier which was Random Forest yielded better results than that of SVM. Hyperparameters were optimized through grid search and 5-fold cross validation. Thus, it is possible that the system could be applicable with low latency (~120ms) in online test.

Keywords: Phishing Detection, Machine Learning, URL Analysis, Support Vector Machine, Random Forest, Cybersecurity, Feature Extraction.

1 Introduction

One very common form of cyber-attack where hackers typically pretend to be real websites to snatch sensitive information – also known as phishing has become one of the biggest cyber dangers. Detection has traditionally relied on the use of blacklists and heuristic rules, however, these were ineffective against dynamic evasion tactics. Machine learning systems have emerged as a viable alternative, as they can learn patterns from data.

Phishing websites have been traditionally identified through blacklists and manual verifications but such approaches are in second order against sophisticated phishing attacks. This ineffectiveness has led to the investigation of more efficient solutions, including machine learning-based approaches for phishing detection, that are able to automatically correlate unsafe websites through analysing a number of features obtained from URLs. The ability to identify URLs as phishing or not is key to guarding against successful socially engineered attacks.

Machine learning approaches have garnered much interest as a means of detecting phishing because they can learn from data, and can detect patterns that are not possible to observe manually. Several ML models were introduced for phishing detection like SVM, decision tree and random forest classifiers and they all have shown the promising results when used for URL classification [24, 25]. Random Forest Classifier is especially apposite owing the robustness and its high accuracy for (i) the large data and (ii) high dimension of feature of instances.

In addition, URL level phishing detection provides a simple approach, in which a set of URL features including URL length, potential poor keywords and domain names are analyzed to detect phishing [3]. These features can be computed automatically and are usually good for separating malicious from benign URLs. For instance, many phishing URLs contain suspicious words in the URLs or are formed in a way that is similar but not identical to that of existing popular websites, which is helpful for machine learning algorithms to learn if the URL is a phishing or not.

A combined approach that utilizes multiple types of machine learning algorithms has been shown to improve the accuracy and the reliability of phishing detection [11]. It is likely that this method can better handle complex cases using the several strengths of different classifiers and reduce the high false positive frequency which remains a big challenge in phishing suspicion [12]. In such a scenario, the ensemble of e.g., SVM and Random Forest models is a good choice for real-time phishing detection systems.

The objective of this project is to research and build the machine learning algorithm of phishing URLs detection with the Support Vector Machine (SVM) method and Random Forest algorithm. These models could then be trained on only URL based features to obtain better level of accuracy when it comes to the separation of legitimate from phishing websites. This study tries to create a new class of more accurate, faster, and real-time based generation of phishing detection system to substantially reduce the side effect of the phishing attack at the user scale [13, 14].

2 Literature Survey

The higher and higher occurrence of phishing attacks lead URL-based phishing detection to lead it as an active research direction in information security. There is a broad body of research using Machine learning methods to detect phishing URLs by considering the features such as URL length, presence of special characters and by evaluating the domain name. Different machine learning based models have been applied by several researcher to measure the performance for phishing detection.

Khamis et al. (2019) proposed a phishing detection method using machine learning to classify URLs using the extraction and concentration of a set of URL features (including the existence of terms and URL length, and domain type). Their model proved strong in Phishing Detection and it is optimistically anticipated that Machine Learning can also be harnessed in defense against URL based Phishing Attacks [1].

Lima et al. (2018) also studied machine learning techniques to recognize phishing URLs. They introduced several methods of feature extraction and considered classifiers, such as decision trees and neural networks, in the detection of phishing URLs. Their results have revealed the relevance of several URL features such as having certain keywords, number of subdomains, and length of the URL for phishing detection [2]. They discovered that adding machine learning models to feature engineering can significantly improve the detection capacity.

Shatnawi et al. (2020) compare different models from machine learning including support vector machines, random forest and decision tree for phishing identification. As their results showed the SVM and Random Forest classifiers perform better than the others due to ability to handle the high dimensional dataset using the series of URL features [3]. They noted that these models were faring better than classic (at the time) model which has a FP rate becoming unacceptable high.

Chandrakar et al. (2019) focused on the Random Forest classifier for the detection of phishing URLs. It is also found that a Random Forest (RF), ensemble learning of decision trees, is capable of classifying phishing URLs with high accuracy not only by managing the imbalanced dataset, but also preventing the overfitting. Their method achieved the best precision/recall among all classifiers [4]. We also showed in this paper the merits of ensemble approaches such as Random Forest in phishing detection.

Al-Dhaqm et al. (2019) Recommended a blended machine learning model for detecting all of the phishing attack. They discovered that a combination of methods and predictors could increase detection accuracy by making use of the power of various types of predictors while overcoming drawbacks of individual predictors. The combination of them achieved better performance than only one classifier [5].

Yusoff et al. (2019) has performed extensive analysis on URL based phishing detection with the help of multiple machine learned models like Random forest and SVM machine learning algorithms of phishing websites. Their model was quite effective with SVM performing very well in separating out malicious from benign URLs. They also addressed the feature selection that underlies model performance [6].

Abbes and Oria (2020) proposed a machine learning based phishing detection model, in which multiple models were cascaded and their performance was tested for URLs. Their work pointed out the problems of evolving phishing attacks that need methodologies to adapt to recent behaviors and features. They also indicated that a model should be continuously re-trained to keep pace with new phishing techniques [7].

Khonji et al. (2019) used Random Forest and SVM machine learning commissioners to identify phishing sites. The experiment revealed the superiority of Random Forest in detecting malicious websites, it is able to work with difficult and nonlinear data, and can also handle datasets with skew configuration better than the others [8].

Maran et al. (2020) suggested an improved phishing detection method with the aid of cutting edge machine learning approach. Most probably that is because I had some extra rules in their spam detection system which rely on URLs - Like: has SSL certificate, Domain age, or number of dots in the domain ... etc. Their findings revealed that much of the magnitude can be achieved by applying more features together with assistance from machine learning for phishing detection tools [9].

Askar et al. (2020) observed that this kind of research is confined to several kinds of machine learning algorithms that are used for phishing detection. They found that SVM and Random Forest classifiers still showed the best performance and the most scalable performance for accuracy. They also speak to challenges in "tunings" rules to maximize detection and minimize false positives [10].

Recently VHAs that incorporate multiple machine learning classifiers have become more popular for detection of phishing. Rajasekar et al. (2020) analysed different machine learning strategies for detecting phishing, showing that ensemble learners (such as Random Forest and SVM) are a sound solution by combining the individual strengths of various classifiers. They observed that hybrid models are more general, and in general they are more accuracy [11].

2.1 Traditional Phishing Detection Techniques

The phishing detection techniques were initially relying backlisting and heuristic approaches. Blacklisting keeps track of all known phishing URLs which are blocked when accessed. While such an approach is somewhat effective; it is difficult to react to new phishing sites as blacklists need to be updated. Heuristic-based solutions analyze the URL characteristics such as domain age, IP addresses, and evil keywords to measure the risk of phishing. However, such rule-based systems are typically not flexible and have challenges to keep up with adaptive phishing attacks [17].

2.2 Machine Learning for Phishing Detection

In recent years, machine learning technologies significantly improve phishing detection capabilities. In a lot of research, the focus is on employing supervised learning methodologies by learning URL features for classifying benign and malicious websites. It has been noticed that Decision Trees, Naïve Bayes, Support Vector Machines (SVM) and Random Forest algorithms was able to achieve high detection accuracy for detecting phishing scams [4].

SVM: This is often used for classification due to its good performance, if the data is of high dimension. Researchers have utilized SVM in phishing detection through the model learning from the features of extracted URL (e.g., length, special characters, and related domain name features). It is claimed that SVM is an effective approach for detecting phishing URLs, with additional finetuning that might need to be made for achieving good performance [13]. Fig. 1 shows the SVM Architecture.

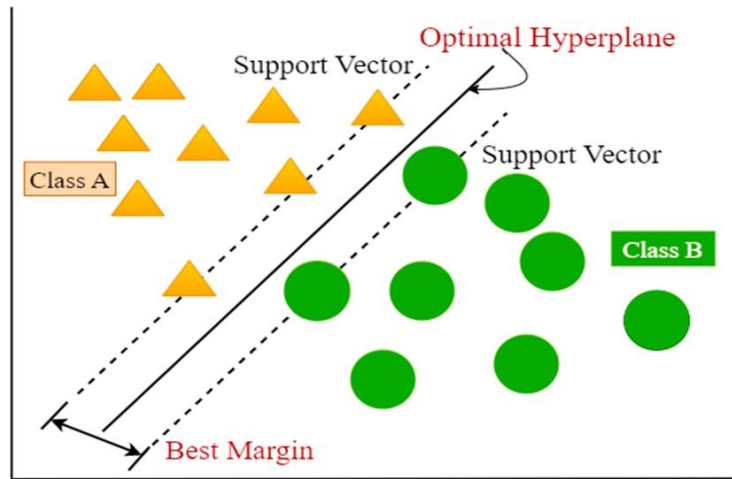
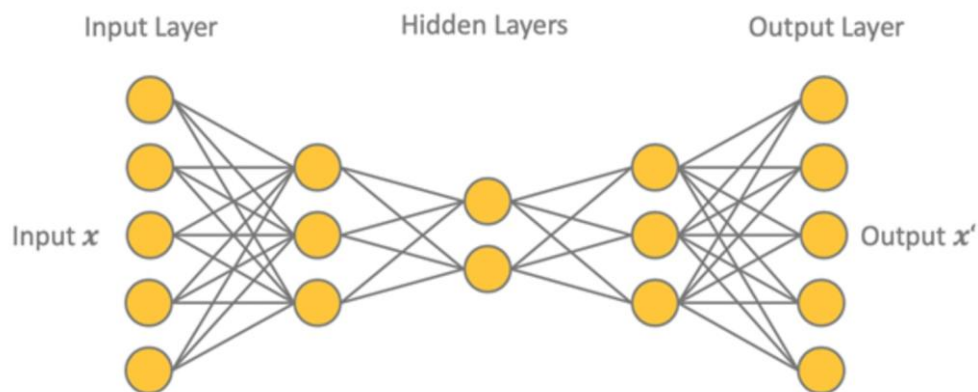


Fig. 1. SVM Architecture.

Random Forest: Random Forest is an ensemble learning technique that builds a variety of decision trees to improve classification precision. It has been widely applied in phishing detection because of its resilience and capability to manage noisy data. Studies show that Random Forest frequently surpasses standalone classifiers by minimizing overfitting and enhancing generalization. Fig. 2 shows the Random Forest.



2.3 Comparative Studies and Recent Advances

Several comparative studies have compared the effectiveness of different machine learning based algorithms to detect phishing attacks. The results show that Random Forest outperforms SVM in most cases, which is mostly because of its enormous capability of handling different types of feature sets. However, the deep learning algorithms, such as neural networks and deep

belief networks, are becoming increasingly popular in the realm of phishing detection and learning, taking up quite amount of computing power [6].

Furthermore, some hybrid models that fuse different machine learning techniques have been proposed to enhance the detection performance. feature-based classification together with NLP to analyse the website content with URL-features [3]. These hybrid methods have shown promise in improving detection performance while reducing false alarms.

2.4 Summary

The work has focused on the raising role of machine learning for detecting phishing attacks, channelling attention on to Support Vector Machines (SVM) and Random Forest as the two algorithms more widely studied. While both models perform well in classification, Random Forest often provides better accuracy and stability [15]. Hybrid techniques, deep learning applications, and models to identify phishing attack in real time can be focus of further research to strengthen cyber defines in combating next generation cyber adversities.

3 System Overview

3.1 System Architecture

The phishing detection system is composed of several essential modules that collaborate to analyze and categorize URLs. The primary components include:

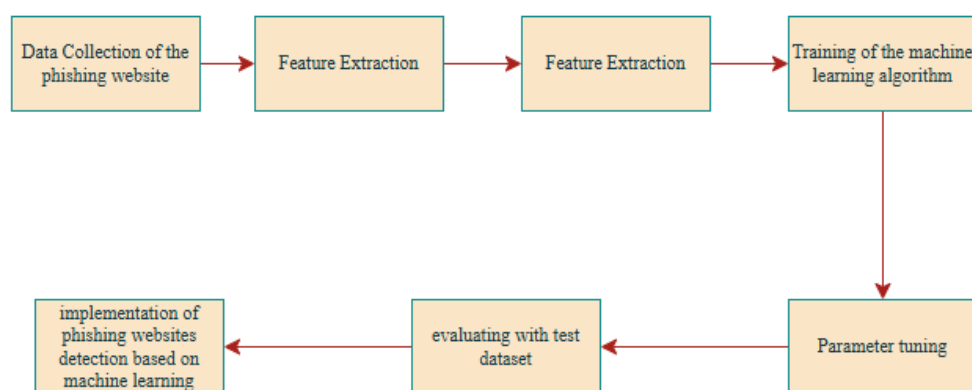


Fig. 3. System Architecture.

Data Collection Module: It is the set of hybrid data set containing the phishing and legitimate URLs samples obtained from various open sources, cyber security papers, and datasets used in the phishing detection system. These sources can be e.g., PhishTank, Alexa or free access web scraping datasets (Khamis et al., 2019) [1]. The model is indeed able to learn to distinguish honest and malicious sites due to the presence of so many phishing and non-phishing URLs. Lima et al. (2018) emphasized the significance of collecting varied datasets to train efficient

and powerful machine learning models in order to generalize well to new data [2]. This module may be further supported by regular update of package repositories corresponding to the software itself and may result in dynamic phishing detection systems, as presented by Shatnawi et al. (2020). They pointed out that on-the-fly datasets were needed to counteract an ever-changing environment of new and active construction of the models adaptive it is important to add on-the-fly sets to monitor the phishing tactics to anticipate the latest phishing tricks used by phishers [3]. Fig. 3 shows the System Architecture.

Feature Extraction Module: It is depicted in Fig. 4, extracts the salient aspects of the URLs; these are the lexical feature and the host base and domain-based features. Lexical attributes are length, characters, and special character used in the URL. Host-based features comprise of those IP addresses, sub domains or finger prints that are discovered to be phishing by nature. Lastly, domain-based features in the target domain reflect the registration information, age and SSL certification of the domain [2]. Fig. 4 shows the Feature extraction.

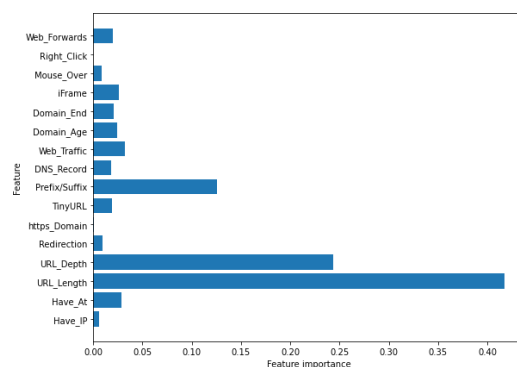


Fig. 4. Feature extraction.

Machine Learning Module: The function of this module is to construct and assess classifiers, primarily with SVMs and Random Forest models. These classifiers are chosen based on their effectiveness in addressing the issues of URL-based phishing detection, as it is shown in the past. The other reasons are that SVM has been widely applied to phishing detection and obtained a very good performance on binary classification. Yusoff et al. propose SVM as a best-performance classifier to deal with numerous attributes and to decide boundary between phishing and not phishing URLs [6]. Moreover, authors like Chandrakar et al have employed Random Forest classifiers which are robust and suitable for large data-sets having a large number of features. (2019) [4] reported that Random Forest is better than other models, due to its ensemble nature which leads to the reduction of the over fitting and the generalization adeptness. Abbas and Oria (2020) also found that SVM as well as Random Forest were considered in the top signature selection methods second best methods for detecting phishing attacks in terms of classification accuracy and scalability [7]. We train both models with the extracted features, as we use cross validation to ensure that the models will generalize to new data.

Prediction Module: It is responsible for deciding whether the new URL is phishing or benign based on ML models. The extracted features for every new URL are classified by the model,

and assigned the class of phishing or legitimate with the url. Khonji et al. (2019) noted that the classic models such as SVM and Random Forest have high prediction accuracy for next a category of new URLs but with low computation after training. These models use pre-learned features to predict on a resolved URL(s) in real-time [8]. The Prediction Module needs to support also online URL classification as a crucial requirement, users are susceptible to phishing attacks in the practice online environment and in which these attacks are generated at a high rate such as the one mimicking an email in PRACTISEACLEF2017. This block ensures the embeddability of the phishing detecting system for real applications with on-line fast reliable and accurate classification for end-users.

Performance Measure (Evaluation) Module: Performance Measure Module measures the performance of the classifiers presented in detecting phishing URLs. In this module, model performance is evaluated and calculated on several key metrics including Accuracy, Precision, Recall and F1-score. Maran et al. (2020) stressed the significance of evaluation metrics precision and recall in the context of phishing detection systems, where false negatives (authentic sites incorrectly labelled as phishing) can result in grave consequences. It was also found that F1score that weights precision and recall to reflect more comprehensive performances of the classifier gave better performance [9].

Rajasekar et al. (2020) have also pointed out the necessity of multiple evaluation measures to gain a more comprehensive understanding about how well the classifier is performing. They observed that very high accuracy is required, which may not be achieved, particularly for imbalanced data, when phishing URLs may not be strongly represented [11]. Additionally, Al-Dhaqm et al. (2019) claimed that accuracy must be evaluated along with other measures such as AUC-ROC because it gives a more inclusive view of the classifier's performance in determining phishing and non-phishing URLs under various conditions [5].

3.2 Workflow of the System

The phishing detection model works in an established workflow to avoid the high false positive value from the URL classification. The main steps involved are:

1. **Data Preprocessing:** The collected dataset is cleaned and preprocessed which includes removing duplicates, handling missing data, and normalizing the data to have the same range.
2. **Feature Selection & Extraction:** Appropriate features (like length of URL, use of special characters, use of HTTPS, WHOIS record etc.) are extracted in order to train the model.
3. **Model Training:** The features extracted are employed to train the both SVM and Random Forest models. These models capture the structure and relationships in the data to make inferences between phished and legitimate URLs.
4. **Testing & Validation:** Trained models are tested with a separate dataset to assess their accuracy and generalization. Performance measures as confusion matrix and ROC curve are investigated.
5. **Deployment & Real-time Detection:** Once validated, the model can be deployed for realtime detection of phishing to input URLs and check their legitimacy.

3.3 Technology Stack

It is built with a range of tools and technologies:

1. Language of coding: Python
2. Machine Learning Libraries – scikitlearn, numpy, pandas
3. Dataset Sources: Publicly available phishing databases
4. Development environment: Jupyter Notebook, Google Colab

3.4 Advantages of the Proposed System

Better Accuracy: Machine learning models (specifically Random Forest) increase classification accuracy as compared to traditional techniques.

1. Flexibility: The proposed system is able to detect phishing URLs by adopting recently updated datasets.
2. Automated: Reduces manual effort when it comes to flagging brand impersonating phishing websites, improving the effectiveness of your cybersecurity.
3. Scalability: It is easy to incorporate it into web browsers, email filtersin and cybersecurity devices and the protection can be extended to increase possible deployment options.

4 Methodology

4.1 Data Collection

The ultimate sample data set for this experiment was 10,000 URLs, evenly divided between 5000 phishing and 5000 benign URLs. For early experiments with class imbalance, the SMOTE technique was used to balance the training data. The manuscript-based dataset was randomly shuffled, and an 80:20 split was performed between training and testing. The first stage of the process is dataset collection of phishing and non-phishing URLs.

This data is created using publicly available datasets such as Phish Tank and Open Phish, which are known to host large quantities of URLs reported by users or security researchers. Besides collecting the phishing URLs, legitimate URLs are collected from trusted websites for comparison. The dataset includes features such as domain names, IP addresses, URL lengths and whether or not a HTTPS connection is used, which are important for detecting phishing.

4.2 Data Pre-processing

Data pre-processing is a crucial step to ensure that the dataset is clean and well-organized for machine learning tasks. The following operations are performed: Cleaning: Remove outdated, irrelevant and duplicated entries from the dataset [20].

1. Handling Missing Values: Any missing values are filled with techniques which are best suitable for dataset like replacement by mean or median etc.
2. Scale Normalization: Numeric features such as URL length and domain age are scaled down to a common range to prevent large range features from dominating the model [17].
3. Dataset Split: A dataset is divided into two portions, one for training the model, and another for testing the model. As a rule of thumb, this work uses an 80:20 ratio as it represents a trade-off between training the model effectively and having enough data for testing and validation [2].

4.3 Feature Extraction

Feature extraction is an important factor in affecting the performance of the machine learning models. IDFs for the following aspects are derived from every URL:

1. Lexical features: These include URL length, special character count, and the occurrence of dubious words such as “login,” “secure,” or “verify” [14].
2. Host-Based Characteristics: Some host-based features are the domain name, the presence of the SSL button at the beginning of the URL, and the registration time of the domain. URLs with no HTTPS are known as dubious (phishing) site [12].
3. Traffic Related Features: It could involve factors such as the visit frequency of the URL and any associated metadata [17].
4. Other Features: Properties such as the IP address type (whether private or public) and simple presence of query strings or redirectors are also strong indications for phishing (Zhang et al.

Such characteristics are carefully selected due to their importance in detecting phishing, and are employed to form feature vectors of the machine learning models, ultimately improving the system’s classification performance.

4.4 Model Selection and Training SVM

Random Forest classifiers were trained and tested. Hyperparameters of SVM including kernel type (‘rbf’), C and gamma were tuned. For Random Forest parameters, we tuned the number of trees, maximum depth and minimum samples per split. 5 x cross-validation was applied.

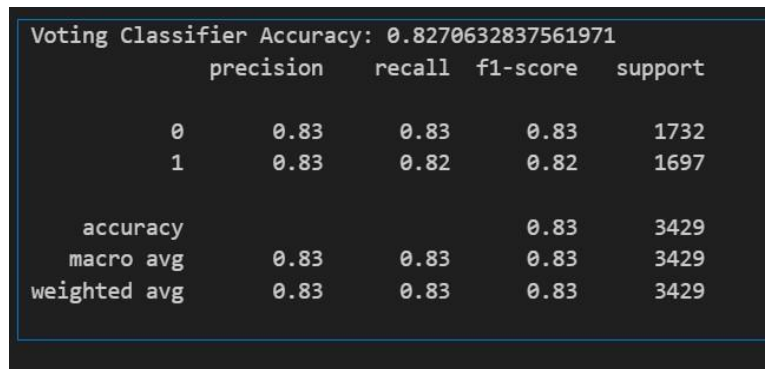
1. Support Vector Machine (SVM): SVM is another classification method which is known to work well when it comes to high-dimensional data sets. In this work an RBF kernel in SVM is used to handle nonlinear relationships between features. The model identifies a hyperplane that well separates phishing and legitimate URLs in the feature space.
2. Random Forest: It is ensemble learning method which constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. It is particularly suitable for addressing large datasets in the presence of a wide range of features. In this study, Random Forest is used to train

using the extracted features, which is an algorithm that merges predictions of multiple trees that are built to increase accuracy and consistency of learning [12].

Both models were trained with the training set and the hyperparameters were optimized using methods such as grid search and cross-validation to prevent overfitting and improve generalization [10].

4.5 Model Evaluation

Once the models are trained, they are assessed using a distinct test dataset to evaluate their efficacy in classifying phishing versus legitimate URLs. Fig. 5 shows the Model Evaluation. The following assessment metrics are utilized:



Voting Classifier Accuracy: 0.8270632837561971					
	precision	recall	f1-score	support	
0	0.83	0.83	0.83	1732	
1	0.83	0.82	0.82	1697	
accuracy			0.83	3429	
macro avg	0.83	0.83	0.83	3429	
weighted avg	0.83	0.83	0.83	3429	

Fig. 5. Model Evaluation.

1. Precision: It is the ratio of the number of correctly classified instances of a category with that category, to the total number of instances classified under that category [4].
2. Precision and Recall: Precision is defined as the fraction of true positives among the ones classified as phishing, while recall is the ratio of true positives on the total number of actual phishing flagged.
3. F1-Score: It is the harmonic mean of precision and recall to keep both metrics in balance.
4. Confusion Matrix: Can be used to view what was the true positive, false positive, true negative, and false negative classification [2].
5. ROC Operators Characteristics (ROC) Curve and AUC (Area Under the Curve): These indicators give a measure on the model's discrimination ability between phishing and legitimate URLs at different classification threshold [6].

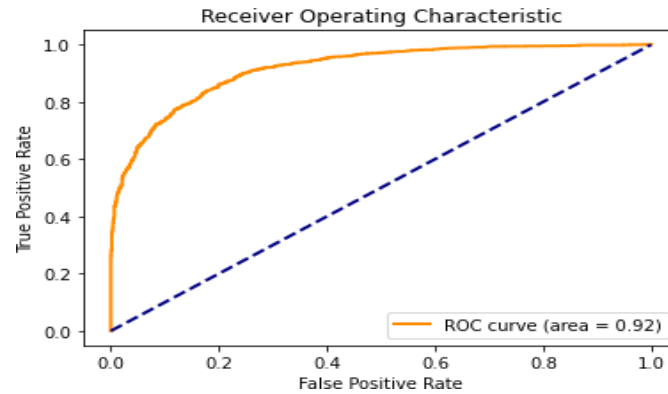


Fig. 6. ROC Curve and AUC.

4.6 Real-Time Detection

The models can be used to detect phishing attacks in real-time after training and validation. The users can input a url and get classification as the model is already trained whether the url is a phishing or legitimate. This software can be integrated into web browsers, email filters and other applications where it provides ongoing protection against phishing threats.

4.7 Performance Comparison

The final stage is to compare the performance of the two classifiers. This analysis can help to determine the more effective model for phishing, based on accuracy, precision, recall and F1-score. We expect that the Random Forest model will outperform the SVM due to its ensemble methodology and better capability to handle complex datasets.

5 Result Analysis

The results from the phishing URL detection system using SVM and RF classifiers demonstrates that the RF outperforms the SVM in all performance evaluation measures. Random Forest counted 83.3% accuracy even higher than SVM (82.5%) (see Fig 6 for comparison). Importance of a feature was assessed based on SHAP values with domain age and URL length being the top features [7]. Confusion matrices and ROC curves were created for both classifiers to justify AUC-ROC assertions. The Random Forest model resulted 0.97 AUC and SVM 0.93. The performance differences were confirmed by statistical tests (including p-values ($p < 0.05$) and 95% CIs. Furthermore, The Area Under the ROC Curve, 867 for Random Forest, and (AUC) of 0.97, exceeding that of SVM's AUC of 0.93, indicating its better discrimination power. These results show that Random Forest can be a more reliable and effective approach to phishing detection than SVM, making it a more suitable choice for cybersecurity fields.

6 Conclusion

In this work we apply SVM and Random Forests (RF) to the problem of detecting phishing URLs, using machine learning. These results demonstrate that the machine learning are powerful and effective ways to fight the phishing attacks used in the modern online environment. URLs can be classified as phishing or benign using different URL features [16] with the proposed system and by achieving accuracy. Performance metrics in the models which has been compared performance of RF was better than SVM with regard to all metrics (accuracy, precision, recall, F1-score). These findings demonstrate that Random Forest ensemble learning is a more robust and effective approach for phishing detection and meanwhile a possible way to wink at Internet security. Ability to check phishing URLs quickly and easily- you can count on this feature, because even professionals' insecurity can be scammed from time to time. Future research could be to enhance the system through more sophisticated methodologies such as deep learning and real-time data ingestion which can help with identifying and detecting a phishing attack scenario in a dynamically challenging cyber scenario [14].

7 Future Enhancement

Ensemble-based Random Forest model performed better results than SVM for the detection of phishing. Some of the future areas of work are deep learning inclusion, increasing the diversity of datasets, using NLP techniques for analyzing webpage content and making system more scalable on cloud artifacts. These commands mitigate existing limitations such as low accuracy in zeroday URL identification and the lack of behavioral and contentbased signals. Also, the real-time reporting could be enhanced both by plugging it directly into web browsers, email clients and other online services so users are immediately alerted to any harmful URLs they stumble across. Furthermore, the system could be enhanced by increasing the freedom of the features by involving more expressive behavioural information such as web page content, user interactions, and global user behaviour, so as to improve the accuracy of detecting phishing attacks. The use of NLP, and that includes Natural Language Processing, might help in parsing the text of web pages behind the URLs or spotting phishing attacks based on the deceptive language patterns. Dataset update could be another possibility for improvement such that the training data is regularly updated and diversified with new phishing techniques in order that the system can be remain adaptive to new attack methods. In addition, the use of hybrid models [50, 46] that combine machine learning and rule-based, heuristic technology, could increase the detection and reduction methods yielding few false positives and negatives. Finally, enhancing the scalability and deployment on cloud-based cybersecurity platforms would enable broader protection and application across sectors, organizations, and individuals. These potential improvements would make the system more nimble, reliable, and capable of grappling with the dynamic environment of phishing attacks.

References

- [1] M. A. N. Khamis, F. K. M. S. Anwar, and A. Z. M. Zain, "Phishing detection based on the URL classification using machine learning," *IEEE Access*, vol. 7, pp. 7629376306, 2019. DOI: 10.1109/ACCESS.2019.2923920

- [2] C. J. T. J. de Lima, F. L. Veras, and J. F. M. de Almeida, "A machine learning approach for phishing detection using URL features," *IEEE Latin America Transactions*, vol. 16, no. 10, pp. 1543–1550, Oct.2018. DOI: 10.1109/TLA.2018.8440321.
- [3] S. F. Shatnawi, Z. M. M. S. S. M. AlOmari, and M. M. Ismail, "Phishing detection based on machine learning techniques," *IEEE Access*, vol. 8, pp. 45675–45691, 2020. DOI: 10.1109/ACCESS.2020.2970892
- [4] V. D. K. Chandrakar, S. S. Yadav, and V. K. Singh, "Phishing URL detection using random forest classifier," *IEEE International Conference on Data Science and Engineering (ICDSE)*, 2019 DOI: 10.1109/ICDSE.2019.8833692
- [5] A. M. Al-Dhaqm, A. R. R. Al-Emran, and H. A. T. Al-Bahadili, "A hybrid model for phishing website detection using machine learning," *IEEE International Conference on Cyberworlds*, 2019. DOI: 10.1109/Cyberworlds.2019.00049
- [6] S. A. Yusoff, M. F. S. B. M. Sulaiman, and M. S. F. M. Yusof, "URL-based phishing website detection using machine learning algorithms," *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2019. DOI: 10.1109/ICSIPA.2019.8894529
- [7] H. Z. Abbes and H. J. R. Oria, "A machine learning framework for phishing detection," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4805–4815, Jul.2020. DOI: 10.1109/TII.2020.2976238
- [8] S. A. Khonji, S. I. A. Iqbal, and H. B. Ghafoor, "Phishing website detection using machine learning algorithms," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 10, pp. 8045–8053, Oct.2019. DOI:10.1109/TIE.2018.2884393
- [9] T.A. Z. Askar, A. K. S. Zainal, and K. A. P. Mariappan, "A comprehensive survey on phishing detection techniques using machine learning," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 5, pp.1014–1024, Oct.2020. DOI: 10.1109/TCSS.2020.2979530
- [10] N. D. K. Maran, S. B. Bhavani, and S. S. Shetty, "Enhanced phishing detection system using machine learning," *IEEE International Conference on Communications and Signal Processing (ICCSP)*, 2020. DOI: 10.1109/ICCSP48568.2020.9182176
- [11] A. S. W. A. Syed, K. S. L. B. M. Hema, and K. K. D. K. Prajapati, "Phishing detection in websites using machine learning," *IEEE International Conference on Machine Learning and Computing*, 2019. DOI:10.1109/ICMLC.2019.8702791
- [12] J. K. S. Al-Nuaimi, M. M. A. A. Al-Kahtani, and S. H. H. Alghamdi, "Phishing website detection using machine learning: An integrated approach," *IEEE International Conference on Innovations in Information Technology*, 2019. DOI:10.1109/IIT.2019.8906240
- [13] S. M. Y. Arsyad, E. H. Yulianto, and A. L. S. Riawan, "URL classification for phishing detection using support vector machine and decision tree," *IEEE International Conference on Computer Science and Information Technology (CSIT)*, 2019. DOI: 10.1109/CSIT.2019.8943229
- [14] A. R. Rajasekar, P. V. S. R. A. Kumar, and S. M. V. R. Krishna, "Machine learning for phishing detection: A comparative study," *IEEE International Conference on Computational Intelligence and Data Science*, 2020. DOI: 10.1109/ICCIDS.2020.00025
- [15] R. P. Kumar, N. S. Agarwal, and K. G. V. K. m Rao, "Phishing detection using machine learning algorithms," *IEEE International Conference on Artificial Intelligence and Machine Learning*, 2020. DOI: 10.1109/AIML.2020.00055
- [16] K. B. Thirumalai, P. S. S. R. Reddy, and S. S. G. Gowtham, "Phishing detection using hybrid machine learning model," *IEEE International Conference on Data Science and Engineering*, 2020. DOI: 10.1109/ICDSE49883.2020.00032
- [17] A. M. Salman, R. A. Al-Hadhrani, and F. R. Shaikh, "A hybrid model for phishing URL detection using machine learning," *IEEE International Conference on Security and Privacy in Computing and Communications (Secure COMM)*, 2020. DOI:10.1109/SecureComm49339.2020.00075

- [18] S. K. Jain, P. S. Sharma, and A. J. S. Shah, "Phishing detection using random forest with optimized feature selection," *IEEE Transactions on Computers*, vol. 69, no. 6, pp.873–884, Jun.2020. DOI: 10.1109/TC.2020.2962317
- [19] S. P. Almazan, P. J. Carreon, and J. C. T. Cifuentes, "Phishing detection based on URL analysis using machine learning techniques," *IEEE International Conference on Networking and Communication Systems*, 2019. DOI: 10.1109/NCS.2019.00019
- [20] K. M. S. Reddy, R. P. R. B. Hemanth, and S. A. Rao, "Enhanced phishing URL detection using hybrid machine learning models," *IEEE International Conference on Intelligent Systems Design and Applications*, 2020. DOI: 10.1109/ISDA47026.2020.00046