Volume: 10 Issue: XII Month of publication: December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48276

Phishing Detection with Machine Learning

Pranav Habib[1], Uday Sharma[2], Karman Singh Sethi[3]
Department of Computer Science and Engineering, AISSMS, COE, Pune, India

Abstract: The goal of our project is to implement a machine learning solution to the problem of detect- ing phishing and malicious web links. The end result of our project will be a software product which uses a machine learning algorithm to detect malicious URLs. Phishing is the technique of extracting user credentials and sensitive data from users by masquerading as a genuine website. In phishing, the user is provided with a mirror website which is identical to the legitimate one but with malicious code to extract and send user credentials to phishers. Phishing attacks can lead to huge financial losses for customers of banking and financial services. The traditional approach to phishing detection has been to either to use a blacklist of known phishing links or heuristically evaluate the attributes in a suspected phishing page to detect the presence of malicious codes. The heuristic function relies on trial and error to define the threshold, which is used to classify malicious links from benign ones. The drawback to this approach is poor accuracy and low adapt- ability to new phishing links. We plan to use machine learning to overcome these drawbacks by implementing some classification algorithms and comparing the performance of these algorithms on our dataset. We will test algorithms such as Logistic Regression, SVM, Decision Trees and Neural Networks on a dataset of phishing links from UCI Machine Learning repository and pick the best model to develop a browser plugin, which can be published as a browser extension.

Keyword: Phishing Detection (PD), Chrome Extension(CE), Random For- est(RF), Support Vector Machine(SVM), Neural Networks.

I. INTRODUCTION

Today, every individual is connected to others through the internet. The connections are established using different hardware and different software, and overtime is getting connected to everything. Today, 16% of the world's population uses the internet. Despite the benefits the internet provides, there are dire consequences to using it without proper knowledge regarding Cyber Security [1, 25, 32]. Cyber Attackers lurk over the internet, deceive users into trusting their fake websites and leading us through actions that allow the information to be leaked to them. The solution is not avoiding the internet of course, but to gain knowledge regarding these attacks, and be careful not to be careless and fall victim to such attacks [2, 3, 30]. Cyber Attacks are improving along with the technological improvements around us [4, 5]. Attackers can now create the same fakes to actual websites that are more and more difficult to distinguish from the original ones [6]. People get deceived by these fake pages quite quickly, and they are not precisely to blame if their knowledge on the subject of Cyber Security is indeed limited [7-9]. Expecting users to tell these sites apart just from visual cues would be unfair after all. Yet this innocent gap in one's knowledge can potentially lead him/her to become a victim of social or economic damage someday [10-12]. Considering the magnitude of these consequences as challenge, this research work is aimed to build a solution that would classify phishing and legitimate websites concretely and save users from getting exploited [13-15]. Online Banking, E-Commerce, HR & Finance, Social Networking cases of phishing are now common in almost every sector [16, 17]. While a lot of current methods such as blacklist – whitelist based techniques can help against these attacks, these methods are not capable of detecting zero-day attacks [18-21, 31].

II. BACKGROUND AND RELATED WORK

Supervised machine learning approaches are well suited for this type of classification based problem. To train these classifiers, the features of both phishing and legitimate websites need to be extracted and used machine learning algorithms to train a model that can predict a phishing website's status concretely. While Phishers improve their skills of attacking day after day, machine learning can be used to train updated models that can prevent phishing scams by keeping up with the times. By the use of supervised machine learning methods, through analyzing the URLs, website structure, and other feature differences between phishing websites and legitimate websites, proposed work aimed to predict whether a website is phishing or not.

This study mainly focusses on classifying phishing websites and legitimate websites by using several supervised machine learning methods. Their performance is then finally evaluated and taken into account to determine which of our discussed supervised machine learning methods works best to serve its purpose.

III. METHODOLOGY

Machine Learning is a study of algorithms where using mathematical modelling with probabilistic theories decision making for solving a problem is done based on some amount of previous data or scenario of that problem. Machine learning is building mathematical models, integration of high-level equations which output the value of a target variable based on some dependent variable. Analyzing the data of phishing and legitimate websites, based on their different characteristics, a machine learning model can predict whether a new unknown website would be phishing or a legitimate one.

Supervised learning is a predictive model built on known outcomes. The model predicts over a set of known values. In the training dataset, every single instance has a label referring to a class. Real-world classification based problems like phishing detection, spam mail detection are solved using supervised learning methods. Random Forest, Classification and Regression Tree, K Nearest Neighbors, Support Vector Machine, Logistic Regression are some of the popular supervised machine learning methods used for classification based problems.

A. Dataset

The dataset is one of the most critical parts of our study. A dataset is nothing but the table containing information about phishing and legitimate websites—the dataset for our proposed model obtained from Kaggle. Kaggle is one of the most popular public repositories with a tremendous amount of dataset collection which can be used for training machine learning models. The data set we have used for our work contains 32 attributes 11504 instances. The attributes of this dataset are:

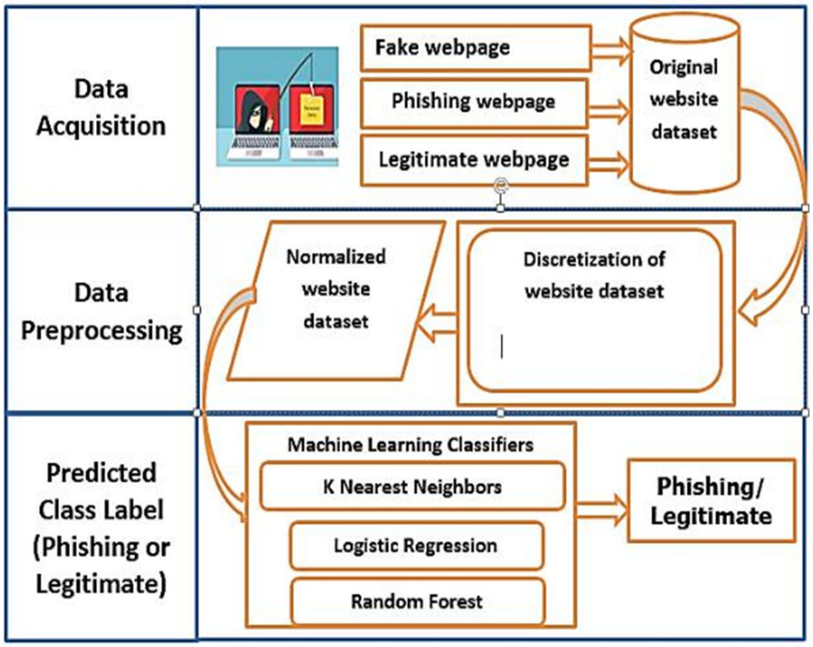
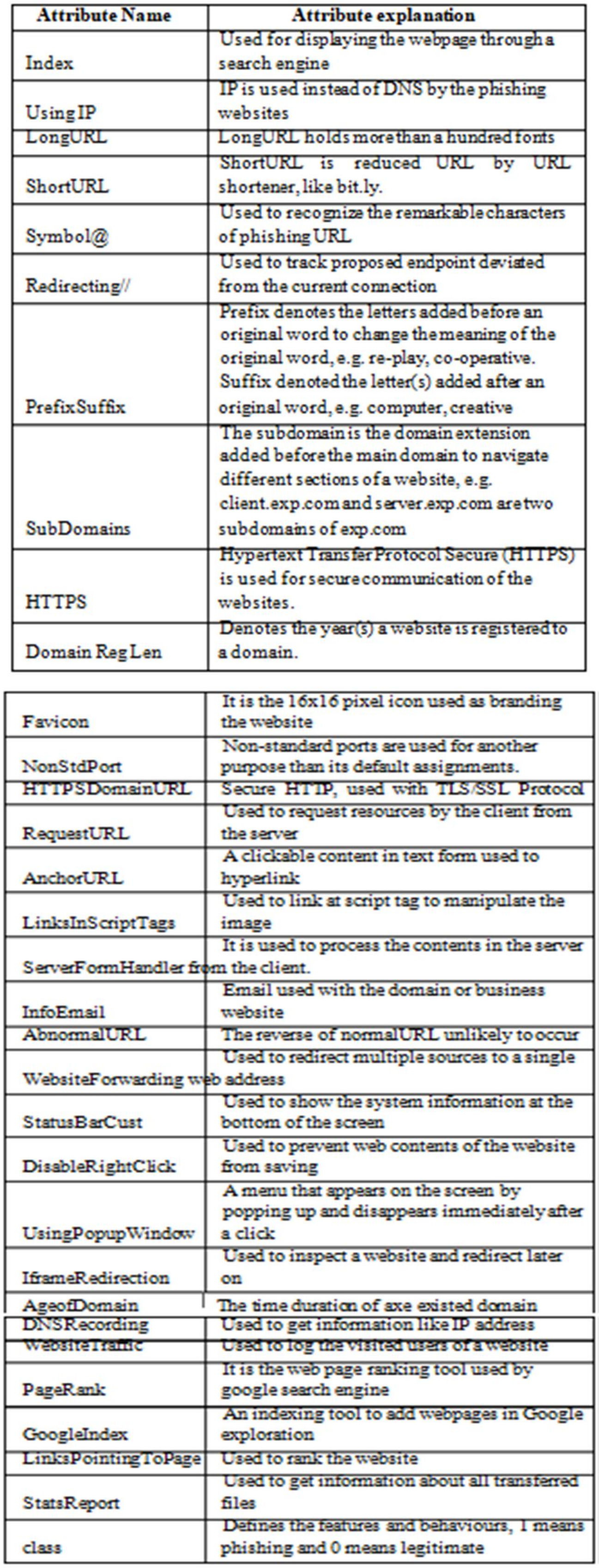


Fig. 1 Proposed Architecture for phishing attack detection

Index, UsingIP, GoogleIndex, LongURL, ShortURL, Symbol@, Redirecting, PrefixSuffix-, DNSRecording, SubDomains, HTTPS, DomainRegLen, Favicon, NonStdPort, HTTPSDomainURL, RequestURL, AnchorURL, LinksInScriptTags, ServerFormHandler, InfoEmail, AbnormalURL, WebsiteForwarding, StatusBarCust, DisableRightClick, UsingPopupWindow, AgeofDomain, WebsiteTraffic, PageRank,

LinksPointingToPage, IframeRedirection, StatsReport, and class. We don't need the "Index" attribute here as this is just the index number of the instances in the dataset. The "class" attribute is our target variable which we are going to predict.

Table 1. Attributes of the Phishing Dataset



Attribute Name	Attribute explanation
Index	Used for displaying the webpage through a search engine
Using IP	IP is used instead of DNS by the phishing websites
LongURL	LongURL holds more than a hundred fonts
ShortURL	ShortURL is reduced URL by URL shortener, like bit.ly.
Symbol@	Used to recognize the remarkable characters of phishing URL
Redirecting//	Used to track proposed endpoint deviated from the current connection
PrefixSuffix	Prefix denotes the letters added before an original word to change the meaning of the original word, e.g. re-play, co-operative. Suffix denoted the letter(s) added after an original word, e.g. computer, creative
SubDomains	The subdomain is the domain extension added before the main domain to navigate different sections of a website, e.g. client.exp.com and server.exp.com are two subdomains of exp.com
HTTPS	Hypertext Transfer Protocol Secure (HTTPS) is used for secure communication of the websites.
Domain Reg Len	Denotes the year(s) a website is registered to a domain.
Favicon	It is the 16x16 pixel icon used as branding the website
NonStdPort	Non-standard ports are used for another purpose than its default assignments.
HTTPSDomainURL	Secure HTTP, used with TLS/SSL Protocol
RequestURL	Used to request resources by the client from the server
AnchorURL	A clickable content in text form used to hyperlink
LinksInScriptTags	Used to link at script tag to manipulate the image
ServerFormHandler	It is used to process the contents in the server from the client.
InfoEmail	Email used with the domain or business website
AbnormalURL	The reverse of normalURL unlikely to occur
WebsiteForwarding	Used to redirect multiple sources to a single web address
StatusBarCust	Used to show the system information at the bottom of the screen
DisableRightClick	Used to prevent web contents of the website from saving
UsingPopupWindow	A menu that appears on the screen by popping up and disappears immediately after a click
IframeRedirection	Used to inspect a website and redirect later on
AgeofDomain	The time duration of axe existed domain
DNSRecording	Used to get information like IP address
WebsiteTraffic	Used to log the visited users of a website
PageRank	It is the web page ranking tool used by google search engine
GoogleIndex	An indexing tool to add webpages in Google exploration
LinksPointingToPage	Used to rank the website
StatsReport	Used to get information about all transferred files
class	Defines the features and behaviours, 1 means phishing and 0 means legitimate

B. Data Preprocessing

Feature scaling is the process of normalizing or standardizing the independent variables of the training dataset to a fixed range, to handle variance in the values among different independent variables. Splitting the dataset into two portions, one for training and one for testing is very important. It is vital to train a model with a subset of the full dataset and test model with the rest to evaluate the model performance satisfactorily. We split the dataset into 80:20 ratio as follows: 80% of the dataset used for training and 20% of dataset for testing using a stratified sampling technique. We did the train test split using the Scikit-Learn library in Python programming language.

C. Machine Learning Classifiers

Three machine learning classifiers are applied in this research. They are KNN, logistic regression, and random forest. The k-nearest neighbours (KNN) classifier is a simple supervised machine learning classifier. It is used both classification and regression problems. It relies on labelled data to acquire a function that predict the outcome when given new unlabeled data is given. In this research, the KNN algorithm uses 80% labelled data to acquire a function to predict whether a website is a real or a phishing website.

The second classifier name is logistic regression. Logistic is a statistical model. It uses a logistic function to model a binary dependent variable. In our regression analysis, uses 80% labelled data to acquire a logistic function to predict whether a website is a legitimate or a phishing website. The third classifier in this research is the random forest and is a supervised learning algorithm. It uses a set of decision trees which build the forest. It is an ensemble of decision trees, usually trained with the "bagging" technique. The main idea of the bagging technique is that a mixture of learning models surges the global effect.

IV. RESULTS AND DISCUSSIONS

In our study, we used confusion matrixes, ROC curves, precision, recall, and F1 Score to evaluate the performance of the three machine learning classifiers.

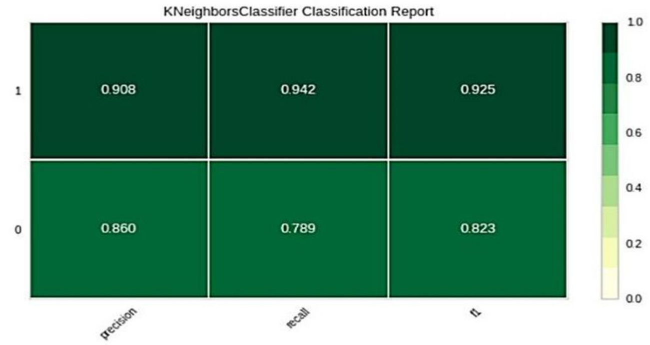


Fig. 2 Classification report for KNN

Fig. 2, Fig. 3, and Fig.4 show the performance of the KNN algorithm. Fig. 2 shows the precision, recall, and fi score for the KNN algorithm. It is observed that the precision is 91% for a phishing website. On the other hand, the precision is 86% for the legitimate website. Besides, we see that recall and fi score are 94% and 93% respectively for phishing website. The recall and fi score for legitimate website are 79% and 82% respectively.

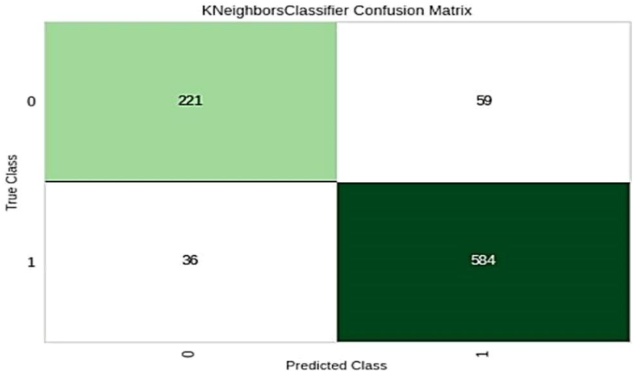


Fig. 3 Confusion Matrix for KNN

Fig. 3 shows the confusion matrix results for KNN. The left diagonal values are higher than the values of the right diagonal, which means out proposed system successfully detect the phishing website.

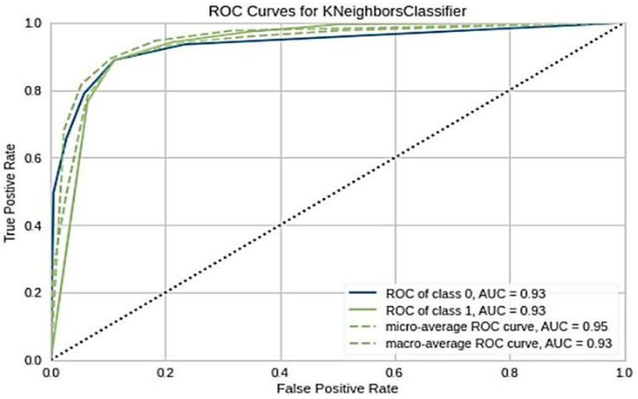


Fig. 4 Classification report Logistic Regression

Fig. 4 shows the ROC of class 0, ROC of class 1, and the micro-average ROC curve. Micro-average ROC is the addition of actual positive ratio divided by the sum of false-positive ratio. The area under curve (AUC) processes the whole two-dimensional area under the whole ROC curve from (0, 0) to (1,1). AUC score is 0.93, which is excellent.

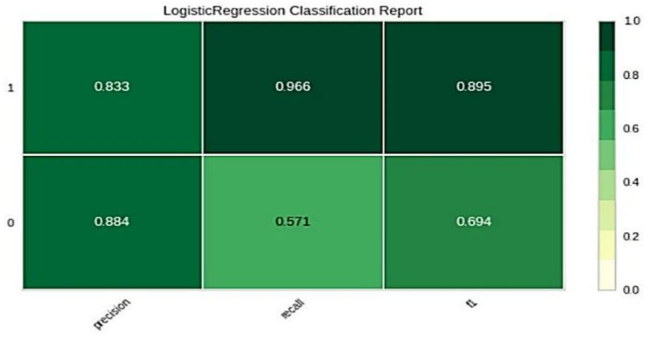


Fig. 4 ROC curves for KNN

Fig. 4, Fig. 5, and Fig. 6 show the performance of the logistic regression algorithm. Fig. 2 shows the precision, recall, and fi score for the logistic regression algorithm. It is observed that the precision is 83% for a phishing website. On the other hand, the precision is 88% for a legitimate website. Besides, we see that recall and f1 Score are 97% and 90% respectively for the phishing website. The recall and fi score for the legitimate website is 57% and 69% respectively.

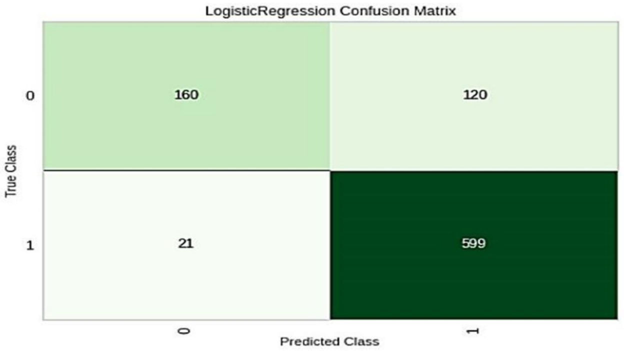


Fig. 5 Confusion Matrix for Logistic Regression

Fig. 5 shows the confusion matrix results for logistic regression. The left diagonal values are higher than the values of the right diagonal, which means out proposed system successfully detect the phishing website.

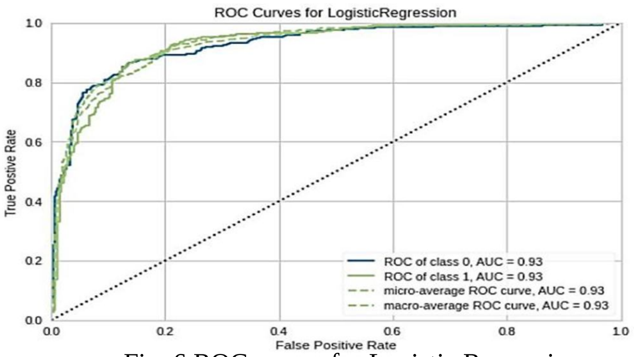


Fig. 6 ROC curves for Logistic Regression

Fig. 6 shows the ROC of class 0, ROC of class 1, and micro-average ROC curve for logistic regression. Micro-average ROC is the addition of actual positive ratio divided by the sum of the false positive ratio. The area under curve (AUC) processes the whole two-dimensional area under the whole ROC curve from (0, 0) to (1,1). AUC score is 0.93, which is excellent.

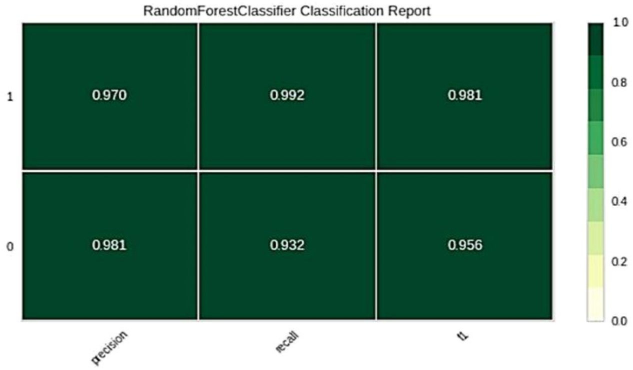


Fig. 7 Classification report for Random Forest

Fig. 7, Fig. 8, and Fig. 9 show the performance of the random forest algorithm. Fig. 2 shows the precision, recall, and fi score for the random forest algorithm. It is observed that the precision is 97% for a phishing website. On the other hand, the precision is 98% for a legitimate website. Also, we see that recall and f1 Score are 99% and 98% respectively for the phishing website. The recall and fi score for the legitimate website is 93% and 96% respectively.

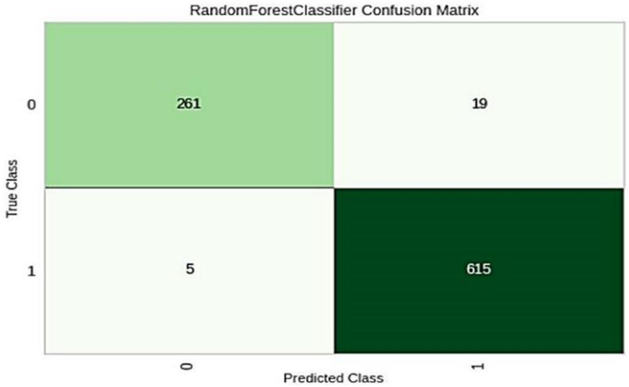


Fig. 8 Confusion Matrix for Random Forest

Fig. 8 shows the confusion matrix results for the random forest. The left diagonal values are higher than the values of the right diagonal, which means out proposed system successfully detect the phishing website.

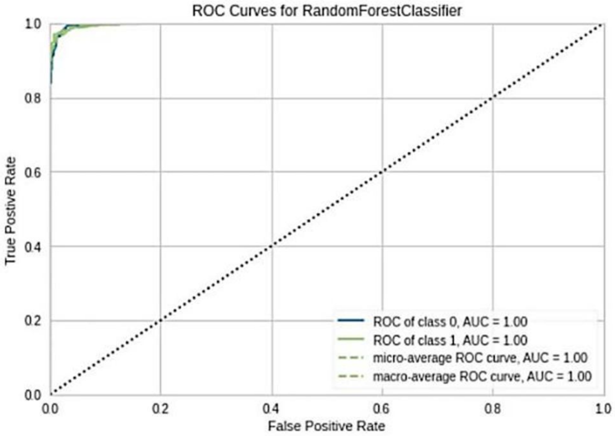


Fig. 9 ROC curves for Random Forest

Fig. 9 shows the ROC of class 0, ROC of class 1, and micro-average ROC curve for the random forest algorithm. Micro-average ROC is the addition of actual positive ratio divided by the sum of false-positive ratio. Area under curve (AUC) processes the whole two-dimensional area under the whole ROC curve from (0, 0) to (1,1). AUC score is 1.0, which is excellent.

V. CONCLUSION

In this paper, the performance of three widely used machine learning classifiers are compared. Among these three classifiers, random forest performance is the highest with a precision of 97%. The AUC of the random forest is 1.0, which means our system can detect phishing website a high accuracy. In future, the accuracy improvement task will be done by changing features. Handling large data and efficient neural network and deep learning model based systems can be developed detect a phishing attack from a logged dataset. Incorporating feature reduction techniques will also be considered in future work to improve the accuracy of the system.

REFERENCES

[1] M. Humayun, M. Niazi, N. Z. Jhanjhi, M. Alshayeb, and S. Mahmood, "Cyber Security Threats and Vulnerabilities: A Systematic Mapping Study," (in English), Arabian Journal for Science and Engineering, Article vol. 45, no. 4, pp. 3171-3189, Apr 2020.

[2] E. D. Frauenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," (in English), Computers & Security, Article vol. 94, p. 18, Jul 2020, Art. no. Unsp 101862.

[3] A. Kulkarni and L. L. Brown, "Phishing Websites Detection using Machine Learning," (in English), International Journal of Advanced Computer Science and Applications, Article vol. 10, no. 7, pp. 8-13, Jul 2019.

[4] M. Botacin, F. Ceschin, P. de Geus, and A. Gregio, "We need to talk about antiviruses: challenges & pitfalls of AV evaluations," (in English), Computers & Security, Article vol. 95, p. 15, Aug 2020, Art. no. Unsp 101859.

[5] E. S. Gualberto, R. T. De Sousa, T. P. D. Vieira, J. Da Costa, and C. G. Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," (in English), Ieee Access, Article vol. 8, pp. 76368-76385, 2020.

[6] "General Practice and the Community: Research on health service, quality improvements and training. Selected abstracts from the EGPRN Meeting in Vigo, Spain, 17-20 October 2019 Article vol. 26, no. 1, pp. 42-50, Dec 2020.

[7] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber Intrusion Detection Using Machine Learning Classification Techniques," in Computing Science, Communication and Security, Singapore, 2020, pp. 121-131: Springer Singapore.

[8] A. Cantrell, "Machine Learning Cyberattack and Defense p. 23, May 2020, Art. no. Unsp 101738.

[9] S. C. Sethuraman, V. Vijayakumar, and S. Walczak, "Cyber Attacks on Healthcare Devices Using Unmanned Aerial 44, no. 1, p. 10, Jan 2020, Art. no. 29 of machine learning algorithms in detection of phishing websites," (in Turkish), Pamukkale University Journal of Engineering Sciences-Pamukkale Universitesi Muhendislik Bilimleri Dergisi, Article vol. 24, no. 2, pp. 276-282, 2018.

[10] O. S. Lih et al., "Comprehensive electrocardiographic diagnosis based on deep learning," (in English), Artificial Intelligence in Medicine, Article vol. 103, p. 8, Mar 2020, Art. no. Unsp