e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:05/May-2020 www.irjmets.com

PHISHING WEBSITE ANALYZER TO SECURE E-BANKING AND E-COMMERCE WEBSITES

Suraj J Pai[*1], Rakshitha Gokuldas[*2], Rahul Kakkadan[*3], Sourabh Hegde[*4], Ms. Saritha Suvarna[*5]

[*1,2,3,4]UG Students, Department of Computer Science and Engineering, Canara Engineering College, Benjanapadavu, India.

[*5]Asst. Prof, Department of CSE, Canara Engineering College, Benjanapadavu, India.

ABSTRACT

Phishing is a fraudulent attempt to obtain sensitive information such as usernames, passwords, and credit card details by disguising oneself as a trustworthy entity in an electronic communication. Phishing is one kind of cyber-attack and at the same time, it is the most dangerous and common attack. Typically carried out by email spoofing or instant messaging, it often directs users to enter personal information at a fake website that matches the look and feel of the legitimate site. Phishing websites have certain characteristics and patterns, and to identify those features can help us to detect phishing. The E-Banking and E-Commerce sector was targeted by phishing more than in any other industry sector. Using Phishing Website Analyzer can help the users to predict the legitimacy of the website they are using. It helps them learn if the website being visited is a phi shy website or not. The Analyzer works by collecting the parameters of the website and evaluating them using different algorithms. The results of the different algorithms are compared against each other and the algorithm that best predicts, is chosen for classification. The use of machine learning algorithms, makes the application better than most other approaches.

KEYWORDS: Phishing Website Detection, Machine Learning, Naïve Bayes, Random Forest, LMT.

I. INTRODUCTION

Phishing is a form of fraud in which an attacker pretense as a trustworthy entity or individual in an email or other communication channels. The attacker might use phishing emails to distribute malicious links or attachments that can achieve various functions, which may include extraction of login credentials or account information from victims. It is a sort of fraud in which the attacker gains full access to others private information. A fake website indistinguishable to the original one can be easily created by an expert designer and so identifying the website as fake could be troublesome. These websites ask users to enter their credentials by claiming itself as a reliable site, for e.g. by using HTTPS. This convinces a user to trust that fake site. They assure privacy and security but, instead, harvest the user's credentials.

These fake web pages have highly visual similarities with real pages to cheat their victims. Also, some of the fake web pages look the same as real web pages. A number of careless internet users easily get cheated by this type of fake pages. These fake pages will collect the personal information of their victims. This information could be their credit card number, password, bank account and also some important information that could be more confidential.

1. Heuristic Based Approach- This technique makes use of heuristics to classify URLs. Heuristics are the features that are considered to check a website. The heuristics like IP address in domain part, '@' symbol in URL, right-click disabled, pop-up windows for passwords etc. to derive rules on these heuristics and decide a threshold for it.

2. Content-Based Approach- The comparison of two web pages is made based on the similar contents on the web page. This technique makes use of the Term Frequency/Inverse Document Frequency (TF-IDF). TF-IDF compares the terms in the original website to the phi shy one. Another approach is to capture the screenshots of a website and then process it to compare. This information retrieved from screenshots after processing can be

given to a search engine to acquire its page rank and check the legitimacy of the website by comparing the content on it. The website logo can be used in this method to analyze the webpage using the Google image database.

3. Blacklist Based Approach- A blacklist includes a list of websites that are declared as spam. Organizations like Google maintain such blacklists. Spam URLs are added to this blacklist. The disadvantage of this approach is that a newly created phishing URL may not be present in the blacklist. Thus, such URLs will be left undetected. URLs present in the blacklist have denied access. This means a user cannot surf this webpage. This method makes use of Google's PageRank value given to each site that is available on the web.

4. Machine Learning Approach- In this technique, features are extracted and they are classified using machine learning techniques. The classification accuracy depends on the algorithm chosen. We can see that more than one ML method are experimented on the same dataset to find the best suitable one. Such comparisons of algorithms can help to give better accuracy in experimentation.

5. Hybrid Approach- In this technique, different techniques are combined to detect if a website is fake or real. For e.g., heuristics and blacklisting of URL can be combined to form a better system. Another hybrid model is a combination of the Machine Learning algorithm. In such a model, the dataset is trained using the first algorithm, and then the result is again passed to the second algorithm for training.

II. METHODOLOGY

The main objective of this paper is to accurately predicting website legitimacy using Machine Learning algorithms and then comparing accuracy between them, thus predicting the best algorithm. Proposed system is an enhanced Phishing Website Analyzer. Through our analyzer, a user can train a machine learning model and check whether the website is legitimate or not. Our proposed system has the following advantages:

- User friendly interface
- Less error
- Remote Analysis
- Centralized data system.

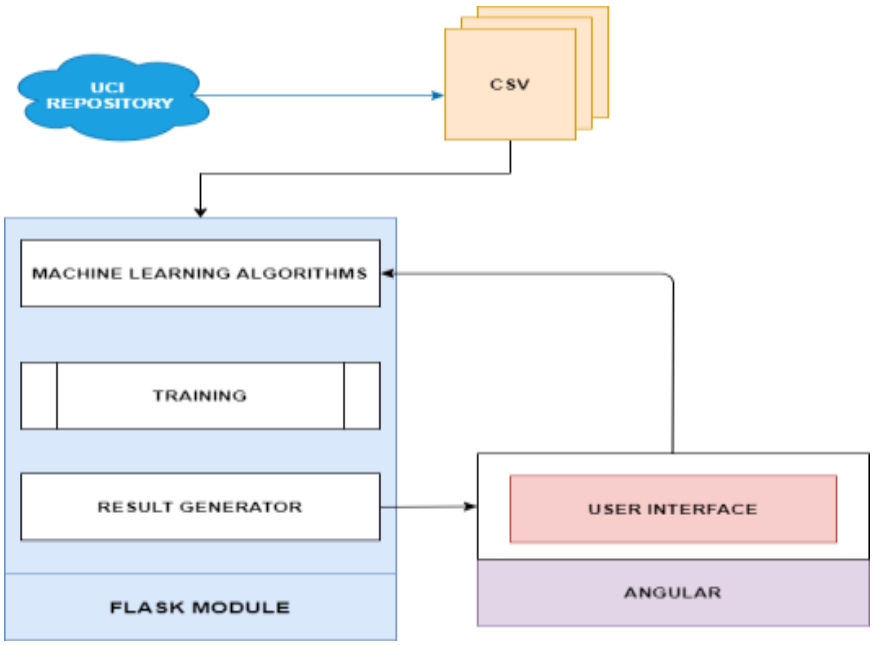

Fig-1: Architectural Design

The Figure-1 shows the architectural design of the process. Here the data collected from UCI repository is pre-processed and passed to Flask module for further processing.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:05/May-2020 www.irjmets.com

III. DATA MINING CLASSIFICATION TECHNIQUES

Data mining is the process of extracting useful and relevant information based on the required purpose. Figure-2 shows the Complete System Flow Diagram of the Phishing Website Analyzer. This diagram shows the complete working of the proposed system. The user has to select the algorithm and pass the parameters for the selected algorithm. This is passed to the train module. The train module trains the model based on the parameters passed. This model is then saved. The saved model is passed to the test module and accuracy of the model is found. Then the model along with prediction data is passed to the predict module. Predict module predicts the output.

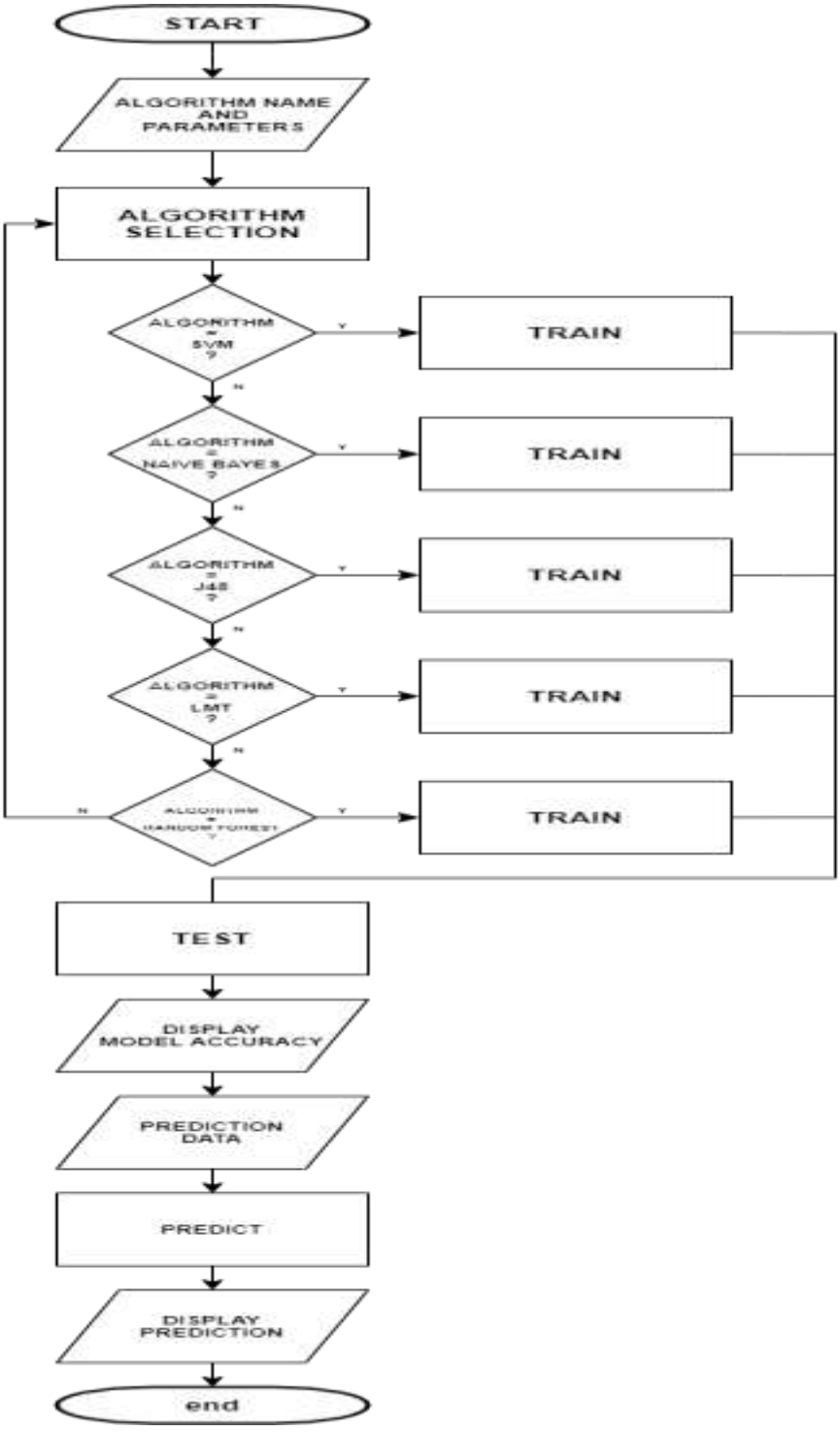

Fig-2: Complete System Flow Diagram

A. J48

- Decision tree J48 is the implementation of algorithm ID3.
- The WEKA project team developed it.
- J48 is an open Java Source code of C4.5 algorithm in the WEKA for data mining.

B. Random Forest

- Random forest algorithm can be used both for classification and the regression kind of problems.
- In the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

C. Naive Bayes

- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.
- Every pair of features being classified is independent of each other.
- Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.
- Bayes' theorem is stated mathematically as the following equation:
  P(A) =P(B|A) P(A)P(B)

Finding the probability of event A, given that the event B is true.

- Event B is also termed as evidence.
- P(A) is the probability of A.
- The evidence is an attribute value of an unknown instance.
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

D. LMT

- A logistic model tree (LMT) is a classification model with an associated supervised training algorithm that combines logistic regression (LR) and decision tree learning.
- The logistic model tree is a decision tree that has linear regression models at its leaves to provide a piecewise linear regression model.

E. SVM

- Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.
- An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.
- New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

IV. IMPLEMENTATION

We implemented the front end using Visual Studio. Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It is used to develop computer programs, as well as web sites, web apps, web services and mobile apps. Visual Studio uses Microsoft software development platforms such as Windows API, Windows Forms, Windows Presentation Foundation, Windows Store and Microsoft Silverlight. It can produce both native code and managed code.

Visual Studio includes a code editor supporting IntelliSense (the code completion component) as well as code refactoring. The integrated debugger works both as a source level debugger and a machine-level debugger. Other built-in tools include a code profiler, forms designer for building GUI applications, web designer, class designer, and database schema designer. It accepts plug-ins that enhance the functionality at almost every level—including adding support for source control systems (like Subversion) and adding new tool sets like

editors and visual designers for domain-specific languages or toolsets for other aspects of the software development lifecycle (like the Team Foundation Server client: Team Explorer).

Visual Studio supports 36 different programming languages and allows the code editor and debugger to support (to varying degrees) nearly any programming language, provided a language-specific service exists.

A. Python

Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as java and C++. The language provides constructs intended to enable clear programs on both small and large scales.

Python supports multiprogramming paradigms, including object-oriented imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library. Python interpreters are available for installation on many operating systems, allowing Python code execution on a wide variety of systems.

B. FLASK

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools. Extensions are updated far more frequently than the core Flask program.

C. Angular

Angular is a TypeScript-based open-source web application framework led by the Angular team at Google and by a community of individuals and corporations. Angular does not have a concept of "scope" or controllers, instead it uses a hierarchy of components as its primary architectural characteristic. It recommends the use of Microsoft's TypeScript language, which introduces the following features:

- Static typing, including Generics
- Annotations

D. Heroku

Heroku is a cloud platform as a service (PaaS) supporting several programming languages. One of the first cloud platforms, Heroku has been in development since June 2007, when it supported only the Ruby programming language, but now supports Java, Node.js, Scala, Clojure, Python, PHP, and Go. For this reason, Heroku is said to be a polyglot platform as it has features for a developer to build, run and scale applications in a similar manner across most languages. Applications that are run on Heroku typically have a unique domain (typically "applicationname.herokuapp.com") used to route HTTP requests to the correct application container or dyno. Each of the dynos are spread across a "dyno grid" which consists of several servers. Heroku's Git server handles application repository pushes from permitted users. All Heroku services are hosted on Amazon's EC2 cloud-computing platform.

V. RESULTS AND DISCUSSION

Predictions have been made by us using our application for classification and accuracy by applying different algorithmic approaches. we are comparing the results in the two ways firstly we find the best algorithm by using the comparison of the different attributes like Correctly Classified Instances, Incorrectly Classified Instances, Mean absolute error and kappa statistic and so on. The selected algorithm makes the website analyzing process automated. Before making payment on any e-commerce website, this prediction model can be used for determining the legitimacy of that website.

A. Random Forest

We use the random forest algorithm in our application to analyse the legitimacy of the websites. In the result we are extracting some statistical information about the algorithm that shows different parameters to describe the accuracy of the algorithm as shown in "Table 1" classification accuracy achieved shows that 96.9609% out of total 2764 instances from which 2680 are correctly classified and 84 are not correctly classified, mean absolute error is 0.0607, kappa statistics is 0.9383 are outputs. Also, in "Fig. 3" we are visualizing the different parameters in a bar chart that can show the accuracy of that algorithm in more precisely. As in description, there is there a series of the bar has shown in a bar chart. Series1 shows the bar of the Fraud websites, series2 show the bar of the Legitimate websites and series3 shows the weighted average of these parameters that are defined in the bar chart.

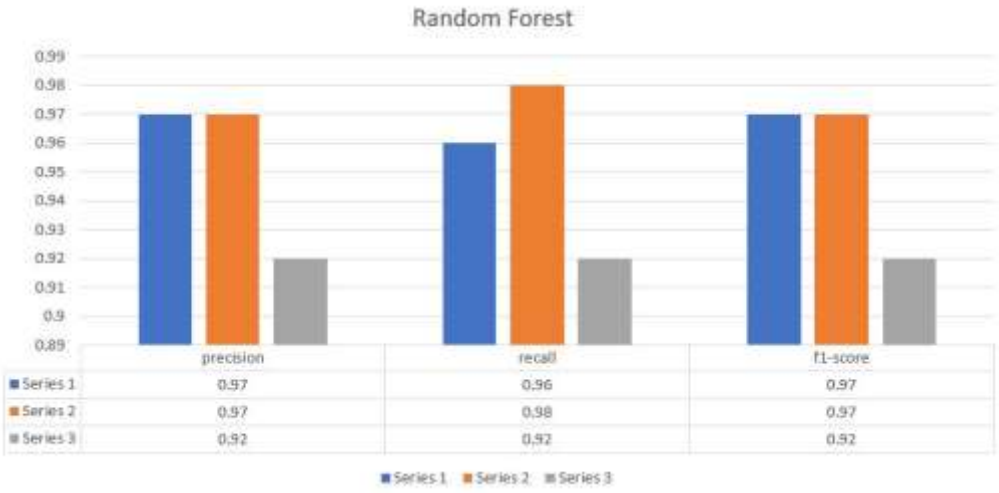

Fig-3: Bar chart Representation of the Random Forest algorithm

Table-1: Statistical Information Of Random Forest Algorithm

Correctly Classified Instances	2680	96.9609%
Incorrectly Classified Instances	84	3.0390%
Kappa Statistics	0.9383	
Mean Absolute Error	0.0607	
Root Mean Squared Error	0.3486	
Total Number of Instances	2764	

B. LMT

We use the LMT algorithm in our application to analyse the legitimacy of the websites. In the result we are extracting some statistical information about the algorithm that shows different parameters to describe the accuracy of the algorithm as shown in "Table 2" classification accuracy achieved shows that 92.3661% out of total 2764 instances from which 2553 are correctly classified and 211 are not correctly classified, mean absolute error is 0.1526, kappa statistics is 0.8447 are outputs. Also, in "Fig. 4" we are visualizing the different parameters in a bar chart that can show the accuracy of that algorithm in more precisely. As in description, there is there a series of the bar has shown in a bar chart. Series1 shows the bar of the Fraud websites, series2 show the bar of the Legitimate websites and series3 shows the weighted average of these parameters that are defined in the bar chart.

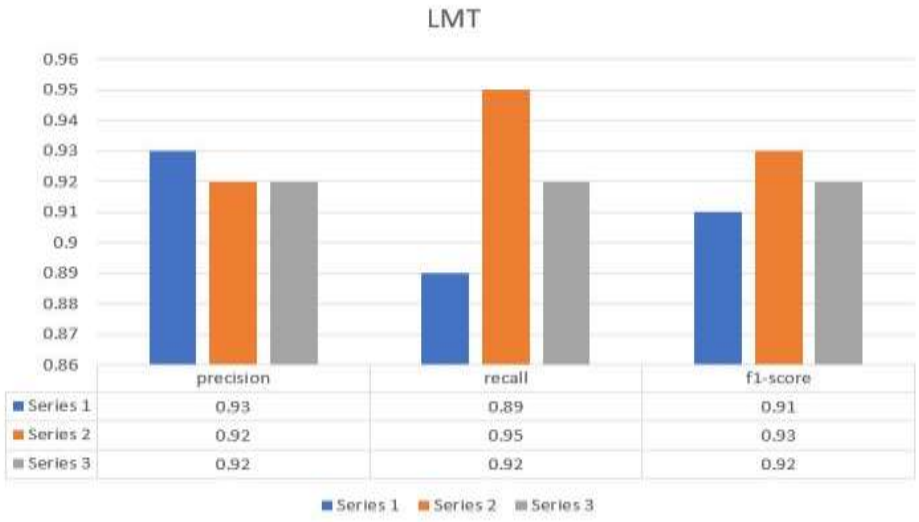

Fig-4: Bar chart Representation of the LMT algorithm

Table-2: Statistical Information Of LMT Algorithm

Correctly Classified Instances	2553	92.3661%
Incorrectly Classified Instances	211	7.6338%
Kappa Statistics	0.8447	
Mean Absolute Error	0.1526	
Root Mean Squared Error	0.5525	
Total Number of Instances	2764	

C. Naïve Bayes

We use the Naïve Bayes algorithm in our application to analyse the legitimacy of the websites. In the result we are extracting some statistical information about the algorithm that shows different parameters to describe the accuracy of the algorithm as shown in "Table 3" classification accuracy achieved shows that 90.0868\% out of total 2764 instances from which 2490 are correctly classified and 274 are not correctly classified, mean absolute error is 0.1982, kappa statistics is 0.7992 are outputs. Also, in "Fig. 5" we are visualizing the different parameters in a bar chart that can show the accuracy of that algorithm in more precisely. As in description, there is there a series of the bar has shown in a bar chart. Series1 shows the bar of the Fraud websites, series2 show the bar of the Legitimate websites and series3 shows the weighted average of these parameters that are defined in the bar chart.

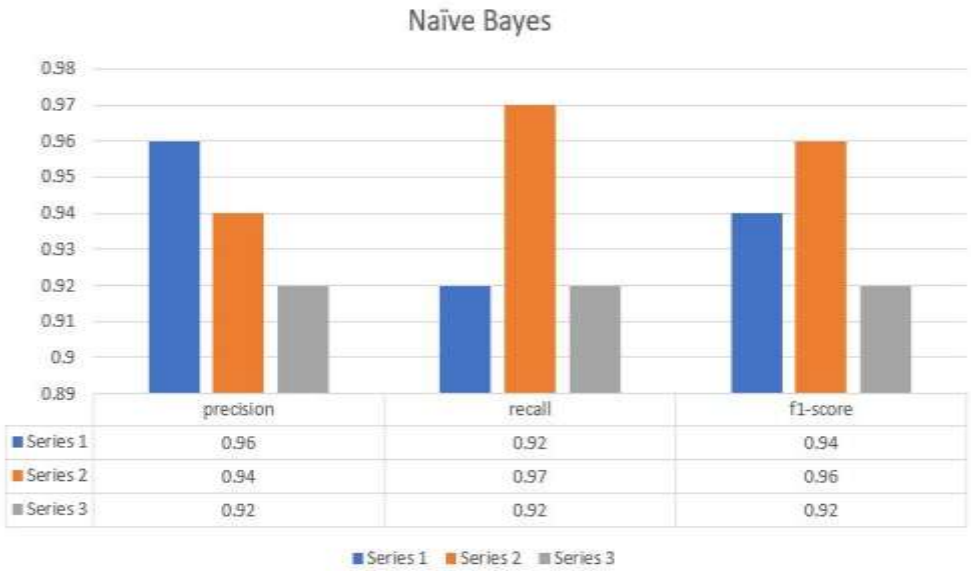

Fig-5: Bar chart Representation of the Naïve Bayes algorithm

Table-3: Statistical Information Of Naïve Bayes Algorithm

Correctly Classified Instances	2490	90.0868%
Incorrectly Classified Instances	247	9.9131%
Kappa Statistics	0.7992	
Mean Absolute Error	0.1982	
Root Mean Squared Error	0.6297	
Total Number of Instances	2764	

D. J48

We use the J48 algorithm in our application to analyze the legitimacy of the websites. In the result we are extracting some statistical information about the algorithm that shows different parameters to describe the accuracy of the algorithm as shown in "Table 4" classification accuracy achieved shows that 95.8755% out of total 2764 instances from which 2650 are correctly classified and 114 are not correctly classified, mean absolute error is 0.0824, kappa statistics is 0.9164 are outputs. Also, in "Fig. 6" we are visualizing the different parameters in a bar chart that can show the accuracy of that algorithm in more precisely. As in description, there is there a series of the bar has shown in a bar chart. Series1 shows the bar of the Fraud websites, series2 show the bar of the Legitimate websites and series3 shows the weighted average of these parameters that are defined in the bar chart.

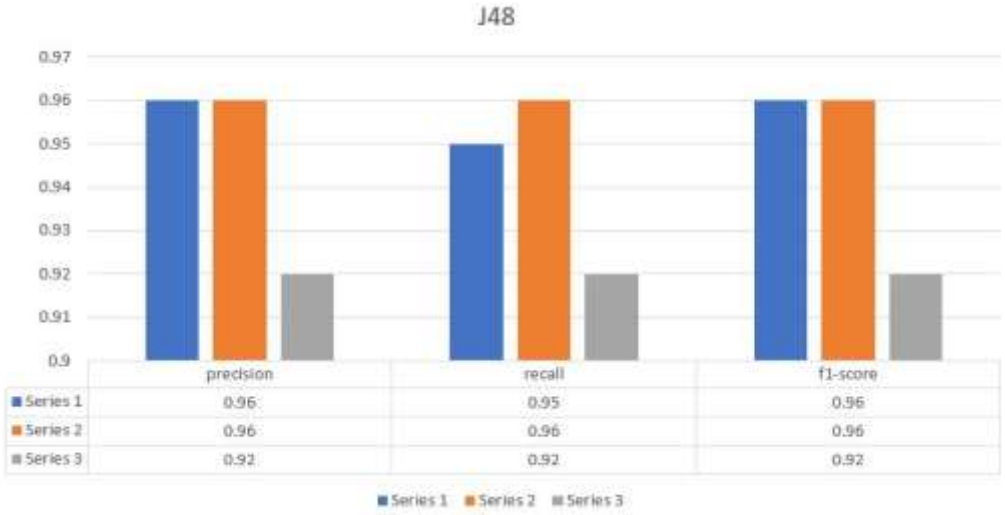

Fig-6: Bar chart Representation of the J48 algorithm

Table-4: Statistical Information Of J48 Algorithm

Correctly Classified Instances	2650	95.8755%
Incorrectly Classified Instances	114	4.1244%
Kappa Statistics	0.9164	
Mean Absolute Error	0.0607	
Root Mean Squared Error	0.3486	
Total Number of Instances	2764	

E. SVM

We use the SVM algorithm in our application to analyse the legitimacy of the websites. In the result we are extracting some statistical information about the algorithm that shows different parameters to describe the accuracy of the algorithm as shown in "Table 5" classification accuracy achieved shows that 95.0434% out of total 2764 instances from which 2627 are correctly classified and 137 are not correctly classified, mean absolute error is 0.0991, kappa statistics is 0.8991 are outputs. Also, in "Fig. 7" we are visualizing the different parameters in a bar chart that can show the accuracy of that algorithm in more precisely. As in description, there is there a series of the bar has shown in a bar chart. Series1 shows the bar of the Fraud websites, series2 show the bar of the Legitimate websites and series3 shows the weighted average of these parameters that are defined in the bar chart.

Table-5: Statistical Information Of SVM Algorithm

Correctly Classified Instances	2627	95.0434%
Incorrectly Classified Instances	137	4.9565%
Kappa Statistics	0.8991	
Mean Absolute Error	0.0991	

Root Mean Squared Error	0.4452
Total Number of Instances	2764

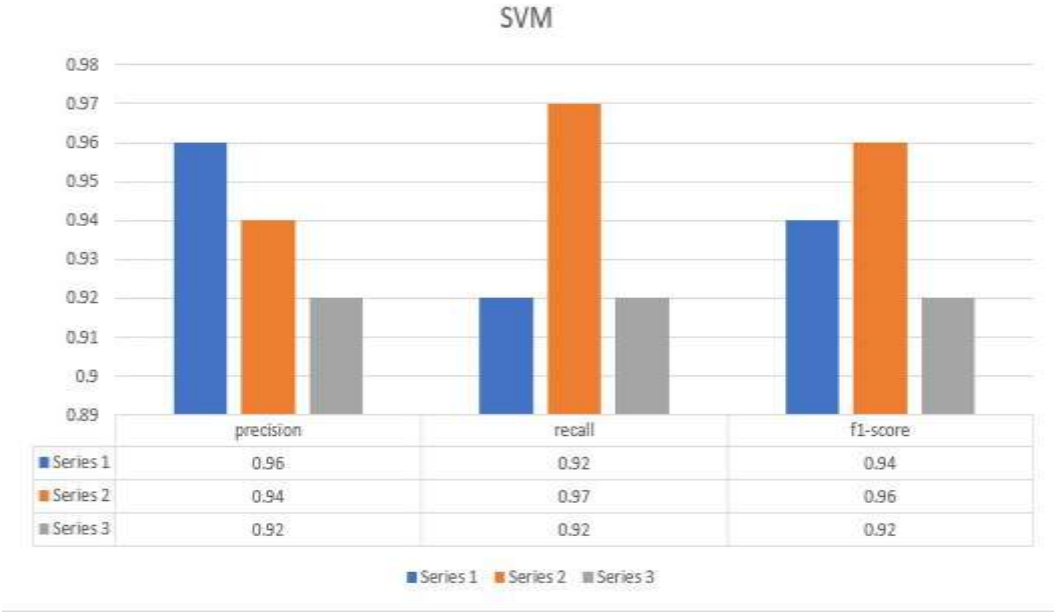

Fig-7: Bar chart Representation of the SVM algorithm

VI. CONCLUSION

Phishing has become a modern-day threat to this rapidly growing world of technology. Today, every country is aiming for cashless transactions, online business, etc. to upgrade with the world. However, phishing is becoming an obstacle to this progress. Phishers are targeting the E-Banking and E-Commerce websites the most, because of its increasing use among the people today.

In this project, we designed a phishing website analyzer which predicts if the visited webpage is a phishy website or not. The website legitimacy is actively predicted by comparing the results produced by five different algorithms. The phishing website analyzer will choose the algorithm that predicts the outcome most accurately. Based on these results, the given webpage will be classified as safe or otherwise.

As a part of future work, other website data collection algorithms like WHOIS could be used to get automated parameters of the website. PhishTank can be implemented in the project for real-time dataset update. It can collect the datasets from the website and provide them to the project using API's.

VII. REFERENCES

[1] D. R. Ibrahim and A. H. Hadi, "Phishing Websites Prediction Using Classification Techniques," in 2017 International Conference on New Trends in Computing Sciences (ICTCS), 2017, pp. 133-137: IEEE..

[2] Rana M. Amir Latif, Muhammad Umer, Tayyaba Tariq, Muhammad Farhan, Osama Rizwan and Ghazanfar Ali, "A Smart Methodology for Analyzing Secure E-Banking and E-Commerce Websites", in Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 8th –12th January, 2019.

[3] Chou N., Ledesma R., Teraguchi Y., and Mitchell J., "Client-Side Defense Against Web-Based Identity Theft," in Proceedings of the 11th Annual Network and Distributed System Security Symposium, San Diego, pp. 1-16, 2004.

[4] Fette I., Sadeh N., and Tomasic A., "Learning to Detect Phishing Emails," in Proceedings of the 16th International Conference on World Wide Web, Banff, pp. 649-656, 2007.

[5] Vamsee Muppavarapu, Archanaa Rajendran, and Shriram Vasudevan, "Phishing Detection using RDF and Random Forests", The International Arab Journal of Information Technology, Vol. 15, No. 5, September 2018.

[6] Zhang Y., Hong J., and Cranor L., "Cantina: A Content-Based Approach To Detecting Phishing Web Sites", in Proceedings of the 16thInternational Conference on World Wide Web, Banff, pp. 639-648, 2007.

[7] Xiang G., Hong J., Rose C., and Cranor L., "CANTINA+: A Feature-Rich Machine Learning Framework For Detecting Phishing Web Sites," ACM Transactions on Information and System Security, vol. 14, no. 2, pp. 1-32, 2011.

[8] Pan Y. and Ding X., "Anomaly Based Web Phishing Page Detection," in Proceedings of 22nd Annual Computer Security Applications Conference, Miami Beach, pp. 381-392, 2006.

[9] Prakash P., Kumar M., Kompella R., and Gupta M., "Phishnet: Predictive Blacklisting to Detect Phishing Attacks," in Proceedings IEEE INFOCOM, San Diego, pp. 1-5, 2010.

[10] M. Haque, "Sentiment analysis by using fuzzy logic," arXiv preprint arXiv:1403.3185, 2014.