**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 7.056**

***IJCSMC, Vol. 9, Issue. 4, April 2020, pg.71 – 75***

# Detection of Phishing and Suspicious URL Using Machine Learning

**Mr. Narayana K E[1]; Srinath R[2]; Srivaths G[3]; Varun S[4]**

[1]Prof, Dept. of CSE & Rajalakshmi Engineering College, Tamilnadu, India
[2]UG Student, CSE & Rajalakshmi Engineering College, Tamilnadu, India
[3]UG Student, CSE & Rajalakshmi Engineering College, Tamilnadu, India
[4]UG Student, CSE & Rajalakshmi Engineering College, Tamilnadu, India
[1] narayana.ke@rajalakshmi.edu.in; [2] srinathraghavendran3@gmail.com;
[3] srivaths1998@gmail.com; [4] varunsubramanyam13@gmail.com

***Abstract—Phishing is a fraudulent process or an attempt to steal one's personal information. Phishing usually occurs via email or by portraying website as a legitimate one. In order to stop the phishing attempt we have to find or recognize the phish. The solution to the problem requires Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we measured and compared the performance of the regressor in terms of accuracy and with the help of values generated from given conditions are predicted.We provided an accuracy of 79% and combination of 17 features.***

***Keywords— Phishing website detection, Random Forest Regressor, Train test Split, Machine Learning, Cross Validation.***

## I. INTRODUCTION

Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that the end goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers.

In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavours in messages and identifying phishing substance on sites, phishes think of new and half breed strategies to go around the accessible programming and systems.

Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to be originating from honest to goodness sources like money related foundations, ecommerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email. The misleading sites are intended to emulate the look of a genuine organization site page.

The employing so as to phishing invader's trap clients diverse social building strategies, for example, debilitating to suspend client accounts on the off chance that they don't finish the account upgrade process, give other data to approve their records or a few different motivations to get the clients to visit their satirize page.

*© 2020, IJCSMC All Rights Reserved* 

Supervised learning (Regression Technique) accommodates a vastly improved precision while unsupervised learning accommodates a quick and dependable way to deal with infer information from a dataset. That's why we used supervised learning in our work.

## II. RELATED WORKS

Amani Alswailem, Bashayr Alabdullah andNorah Alrumayh,2019[2] used to detect phishing websites based on random forest technique.They had combination of 36 features to classify the datasets and the given url.The 36 features can be categorized into three categories.They are features from url,features from page content and features from page rank.The url and page rank features can be extracted from URL where page content features can be extracted from DOM.By splitting the datasets into train data and test data they classified the given url with the help of trained data model.

Oby James, Ciza Thomas and Sandhya L,2013[6] proposed that domain names and phishing url have different length compared to legitimate URL.It is used to detect phishing websites based on Naive Baye's classifier or SVM or Regression classifier or KNN.The host based page based and lexical page based property were applied to the url to form a featured valued datasets.The feature value for a phishing website is 0 and feature for a benign dataset is 0.The dataset is splitted into train and test files.The classifier is decided by the WEKA and using MATLA.By using any one of the classification algorithm it can classify the given input url as phishing or not based on the training dataset files.

Ozgur Koray Sahingoz, Saide Işılay Baykal and Deniz Bulut[4] used 37,175 phishing and 36,400 legitimate websites and the model is trained using Artificial Neural Network(ANN) and Deep Neural Network(DNN).The websites are classified with the help of 16 features combination.In case of ANN the dataset is trained with the help of RELU and TANH activation function with one hidden layer and 20 neurons.In case of DNN the activation function is RELU and 40 neurons with 91% accuracy.Once the training is done,the given url is cross validated and the dataset is splitted into tainining and test data.Then the URL is classified using SVM to predict the output.

## III.DATASET AND REPROCESSING

The method in this paper is Random Forest Regressor where the dataset from [1], it consists of both phishing and legitimate website. It is separated into training data and test data. The dataset consists of a total of 11,056 page features with 6,157 legitimate website and 4,898 phishing website.

Each entity in dataset is broadly classified based on these 17 features.Based on the 17 features the result obtained for each entity is stored. It is easy for our model Random Forest Regressor(RFR) to train the model and it is easily used for regression.

## IV.METHODOLOGY AND EXPERIMENTS

### *A. Splitting dataset using Train Test split:*

The preprocessing is done by splitting dataset into train and test data..The dataset is splitted with the help of train test split from sklearn package.The dataset contains 6,157 legitimate website and 4,898 phishing website.In our case we provide the test_size parameter as 0.2.So it splits the 11,056 data into 8,845 training data (i.e 80%) and 2,211 test data(i.e 20%).

*© 2020, IJCSMC All Rights Reserved* 

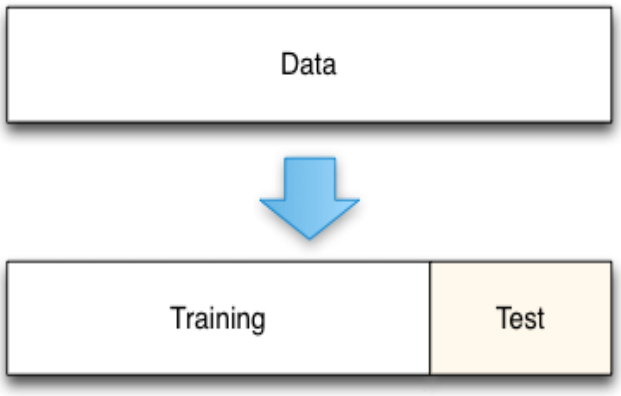


Fig.1 Train test Split

*B. Train model using Random Forset Regressor:*

Once the dataset is splitted using traintest split the train data is provided to Random Forest Regressor(RFR) and the model is trained.The Random Forest Regressor (RFR) is imported from sklearn package.The model is trained with the help of decision trees.The training data is splitted and provided to each decision tree for training.In our case,we provided the number of decision tree as 10.So each 10 decision tree is trained of different training data inputs.Once the model is trained we provide the test data to the model.So the test data is splitted and provided to 10 decision trees.Each decision tree provides a continuous value for the test data based on the model trained.The continuous value from each decision tree is taken and average is calculated.From the average it is predicted that it provides correct value for a test data or not.

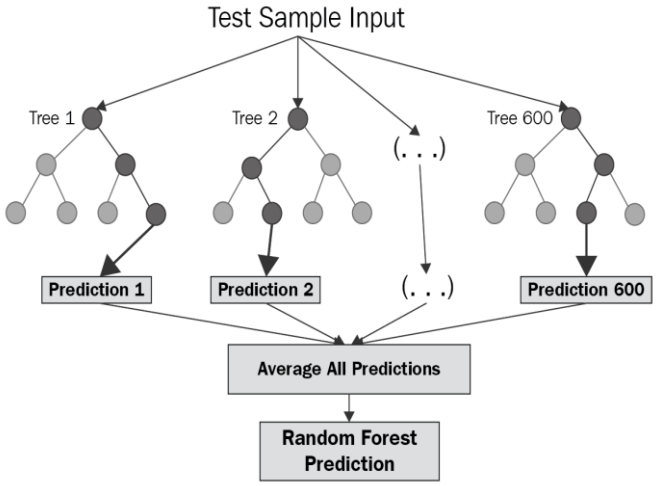


Fig. 2 Random Forest Regressor(RFR) process

*C. Cross Validation:*

After training the model,the dataset is cross validated to check the accuracy.The cross validation is imported from sklearn package.The dataset is splitted based on the k-fold value specified in cross validation.In our case,we provide the k-fold value is 5.So the dataset is splitted into 5 parts.

Once the dataset is splitted into five parts.Any four of the five part is taken and it is used to train the model.The remaining one part is used as test data.Once the model is trained the test data is used to predict the value.By taking anyone of the four part in previous train data as test data and other three as train data.This process is continued untill all part is treated as test data.Each value observed during the process is used to calculate the accuracy.

*© 2020, IJCSMC All Rights Reserved* 

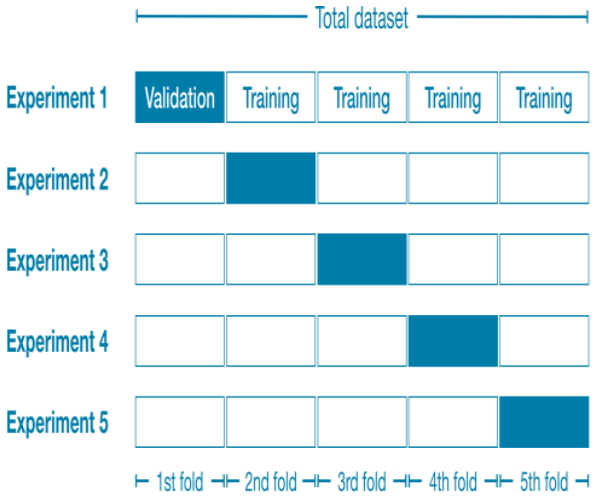


Fig. 3 Cross validation process

## V. FINAL RESULT

The values generated for the given URL from the 17 condition is taken as test data and provided to the model to predict the output.Each decison tree in the model takes the test data which will provides some continuous value.The value from each decision tree is taken and averaged to get some final value.

If the average value from the model lies between in the range of -1 to -0.1 or less than -1 then it is classified as Phishing Website.

If the average value from the model lies between in the range of 0 to 0.9 then it is called as Suspicious Website.

If the average value from the model is greater than or equal to 1 then it is a Legitimate Website.

- Percentage of Finding Legitimate Website correctly: 90.2%
- Percentage of finding Phishing Website: 80.51%
- Percentage of finding Suspicious Website: 70%

The algorithm has an accuracy of 79% by using Random Forest regressor and combination of 17 features

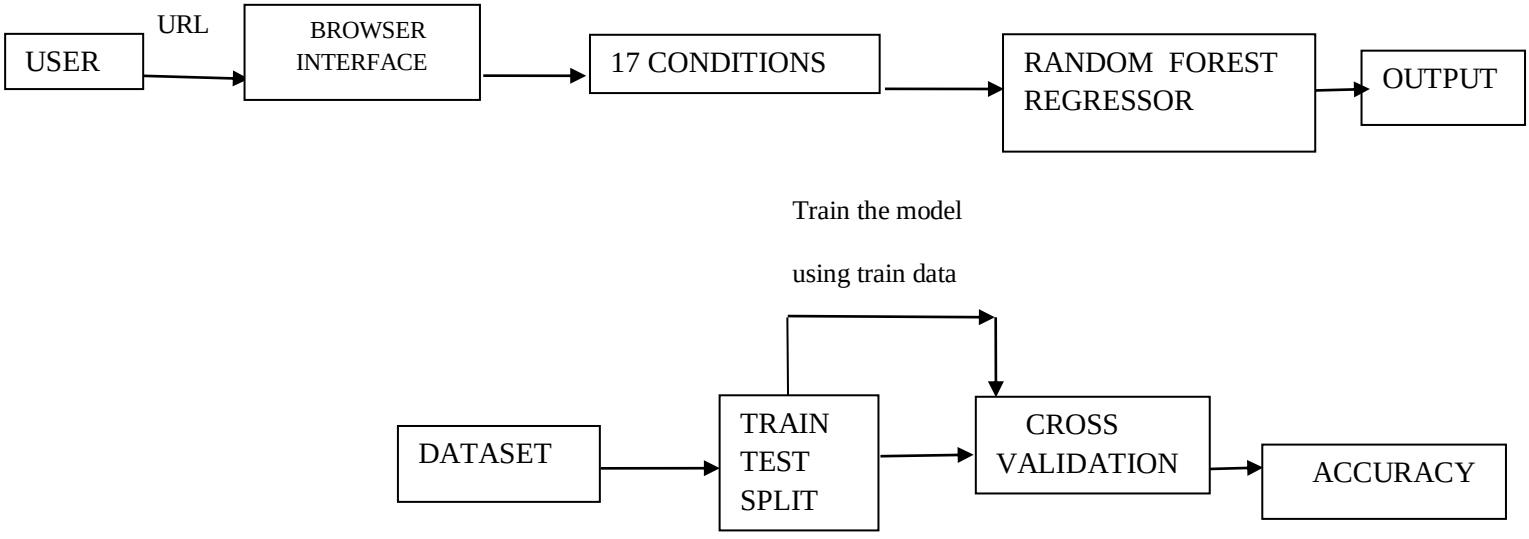


Fig 4 System Architecture

*© 2020, IJCSMC All Rights Reserved* 

## VI. CONCLUSIONS

This paper presents identification of phishing websites. It also identifies suspicious websites and they are predicted with an accuracy of 79%.

| Name | Specification | Model |
|---|---|---|
| CPU | Intel | i5 5200U |
| GPU | Intel HD graphics | 5500 |
| RAM | DDR3 | 8GB |
| OS | Windows | 10 |

# REFERENCES

[1] Datasets from Phising website dataset at kaggle. https://www.kaggle.com/akashkr/phishing-website-dataset
[2] (Amani Alswailem,,Bashayr Alabdullah and Norah Alrumayh,2019), "Detecting Phishing Websites Using Machine Learning".
[3] (Kholoud Althobaiti,Ghaidaa Rummani,Kami Vaniea,2019), "A Review of Human- and Computer-Facing URL Phishing Features".
[4] (Ozgur Koray Sahingoz, Saide Işılay Baykal and Deniz Bulut), "Phishing Detection from Urls by Using Neural Networks".
[5] (Ebubekir Buber, Önder Demir and Ozgur Koray Sahingoz,2017), "Feature Selections for the Machine Learning based Detection of Phishing Websites".
[6] (Joby James, Ciza Thomas and Sandhya L,2013), "Detection of Phishing URLs Using Machine Learning Techniques".
[7] (Vaibhav Patil Pritesh Thakkar Pritesh Thakkar Tushar Bhat,2018), "Detection and Prevention of Phishing Websites using Machine Learning Approach".
[8] (Amirreza Niakanlahiji,Bei-Tseng Chu and Ehab Al-Shaer,2018), "PhishMon: A Machine Learning Framework for Detecting Phishing Webpages".
[9] (Sonam Saxena,Amit Shrivastava and Vijay Birchha,2019), "A Proposal on Phishing URL Classification for Web Security".
[10] (Hetal Rahul Rajpura and Hiteishi Diwanji,2013), "Enhancement of Fake Website Detection Techniques Using Feature Selection and Filtering Algorithms".

*© 2020, IJCSMC All Rights Reserved* 