IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary Research**
**Website: www.ijdacr.com (Volume 7, Issue 12, July 2019)**

# Phishing URL Detection using Bayesian Optimized Random Forest Classifier

Vijaypal Singh Rana
M. Tech. Scholar
Dept. of Computer Science and Engineering
Doon Institute of Engineering & Technology,
Dehradun, India
vpsrana82@gmail.com

Rahul Joshi
Assistant Professor
Dept. of Computer Science and Engineering
Doon Institute of Engineering & Technology,
Dehradun, India
joshirahul8271@gmail.com

*Abstract* – **Internet scams are numerous and varied. Anyone is likely to be the target of an attack while browsing the net. More and more crooks do not hesitate to use Social Engineering as a lever to acquire sensitive data unfairly by exploiting human flaws. Phishing is a Social Engineering technique used by these hackers. It is used to steal personal information in order to commit an identity theft without the knowledge of their victims. The persuasion power of these crooks is the keystone of a successful attack. This paper presents a model with the highest precision results which consists of Bayesian optimized support vector machine classifier. The performance of proposed framework is evaluated using accuracy, precision and sensitivity.**

*Keywords* – **Bayesian Optimization, Phishing URL, Random Forest Classifier.**

## I. Introduction

For centuries, scams of all kinds have been devised and orchestrated by unscrupulous people in order to deceive the trust of others and thus obtain goods fraudulently.

Going back to the sixteenth century, we see the appearance of a scam called the scam of "The Spanish prisoner". The scam meant extorting money from wealthy middle-class people by a combination of pretending that a handsome and wealthy Spanish princess was being held captive by the Turks and that ransom was required for her release.

At the end of the eighteenth century, another fraud based on the same psychological effects that it appeared, it is: "The letter of Jerusalem". The principle being that the scammer makes believe a victim, through a series of letters addressed to it that he has a fabulous treasure but for reasons beyond his control, he no longer has the opportunity to access. The scammer then calls on the goodness of his victim to seek his help to recover this treasure. This one, tempted by the prospect of this treasure, begins to pay money so that the rogue can recover it. Unfortunately for the victim, the scammer will always find an excuse to legitimize the impossibility to repatriate the jackpot while inciting the victim to pay a new sum of money to continue the quest for treasure.

At the end of the twentieth century, the appearance of the internet gave a descendant to this scam, called this time "Fraud 4-1-9". Based on the same principle as "The letter of Jerusalem", this fraud abusing the ingenuousness of its victims, is now operated by modern means of communication namely by email, mainly by email but sometimes also by SMS.

The common point of all these scams is their exploitation of the psychological flaws of the human being. A victim is manipulated by a scammer who abuses his trust and credulity to extract what he needs. The democratization of internet access has extended the possibilities of scams of all kinds on the web. Scams using social engineering have become numerous and varied.

We still see in many Internet users, a lack of specific knowledge on this subject, which can lead to the disclosure of personal information thus allowing malicious people to impersonate their identity in order to derive an advantage.

At present, no anti-virus is able to completely protect users against their own weaknesses. The common sense of everyone is the rule of gold to avoid being trapped.

Phishing fraud - that is, the theft of banking or personal information by phishing techniques and their conversion into money or goods and services - has been steadily increasing for several years and the phenomenon does not seem to have occurred. On the contrary, it has become a widespread practice among web crooks due to the increased use of social networks, e-commerce, mobile devices [1] [2] and cloud solutions to store and manage sensitive data. To convince oneself of this, simply type the terms "bank", "fraud", "Scam" and "phishing" in Google or Google Scholar. We get almost a million results in Google and 10,800 in Google Scholar. This is how important the topic is on the Internet and arouses the interest of researchers and organizations that fight against phishing. Among these organizations, there is the Anti-Phishing Working Group (APWG) which published in its 2017 report that more than 91% of all phishing attacks in 2016 targeted five types of industries in particular, financial institutions, cloud-based data hosts, web hosts, online payment services and e-commerce services [3]. This figure of 91% represents an average increase of 33% per type of industry compared to 2015. An increase which is, however, abnormally high for Canadian companies that have experienced, among the developed countries, the strongest phishing growth in 2016, nearly 237% according to the Phishlabs 2017 report [4], mainly in the financial institutions sector, where the 444% ceiling was reached. Trademarks targeted by phishing campaigns reached an average 2016 record of 380 per month, 13% higher than the previous year.

In addition to targeting businesses and trademarks, fraudsters target consumers who connect to the Internet.

Other statistics from the 2016 and 2017 APWG reports show that the number of detected websites that were dedicated to phishing attacks increased from 393K in 2014 to 1.22M in 2016, an increase of 310%. As for the number of domains where these sites reside, it would be 170K in 2016, which represents an increase of 23% compared to 2015.

Since March 2016, 93% of all phishing emails had a "ransomware" encryption system, according to a report published by Phishme Inc.[5]. Also, there is an increase in the types of attack targets. Attackers increasingly prefer to attack online payment systems like PayPal, Boleto, Bitcoin [6], and businesses that manage personal information.

Another, and not least, indicator of the extent of the phenomenon of phishing is the multiplicity, both in America and in the rest of the world, of national organizations and multinational coalitions of companies fighting this scourge. Their goal is to share information and know-how to reduce or even eliminate identity theft and fraud that result from the growing problem of phishing. These organizations include the FBI and NW3C partner Anti-Phishing Working Group, the Internet Crime Complaint Center (IC3), The Coalition on Online Identity Theft, the SCAMwatch website, The Federal Trade Commission of the United States, The 419 Coalition Website [7].

## II. Proposed Methodology

### A. Proposed Architecture

The classifier takes unclassified URLs as input, and returns a predicted binary class as output (either Phish or Benign). Our aim is to evaluate the effectiveness of URL features as discriminating features.

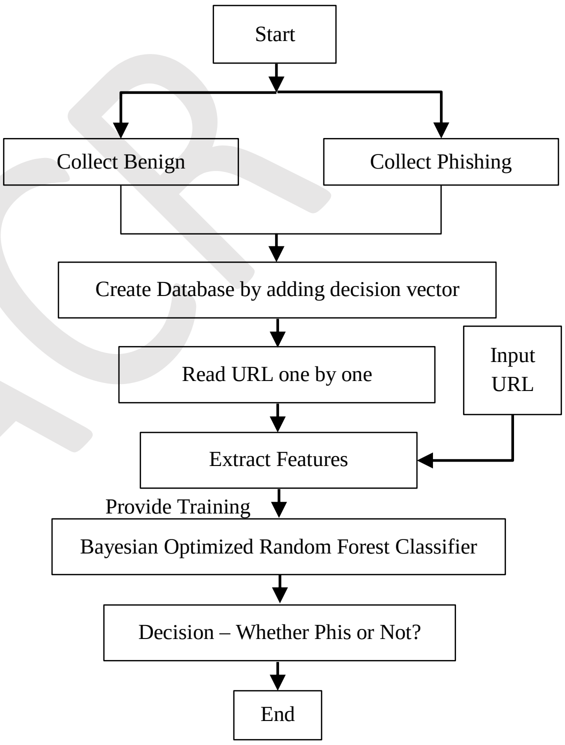
Start
Collect Benign
Collect Phishing
Create Database by adding decision vector
Read URL one by one
Input URL
Extract Features
Provide Training
Bayesian Optimized Random Forest Classifier
Decision – Whether Phis or Not?
End

Figure 1: Flow diagram of proposed approach

We started with collection of URLs and then after loading the URLs we started by reading URLs one by one for feature extraction. To facilitate feature extraction, each URL was split into three sections: protocol, domain, and path. All subsequent feature extraction was performed on these sub-regions. After collecting of URL features, the classifier's life initiates by a supervised learning phase. During this phase, the classifier is fed with pre-classified URL along with their pre-defined class. The classifier is then able to perceive a classification model. Once the learning phase is complete, the classifier is given unclassified URLs as input, and a predicted class is returned as output.

Architectures also hold room for checking a particular URL for Phishing. A random URL is provided to the trained classifier for recognizing the class (Phishing or Benign) of the given URL.

### B. Collection of URLs

For the base of URLs we used a public database used by OPENDNS, this database is published by their site phishtank.com [8] (phishtank.com is created by OPENDNS) which is a site where users can report suspicious sites. The database made public by the site is verified by experts who say that these addresses are actually phishing addresses. We used two versions with different release dates for more precision. For the base of the safe addresses we used, several sources like ALEXA.com [9], and Google ranking specialized in the statistics of the traffic on internet and which give the classification of the most popular sites periodically. For the number of instances of the base we have 4806: phishing addresses. 535: addresses.

### C. Lexical Feature Extraction

Lexical features are the textual properties of the URL itself, not the substance of the page it indicates. URLs are human-readable text strings that are parsed in a standard manner by customer projects. Through a multistep determination process, programs make an interpretation of each URL into guidelines that find the server facilitating the site and indicate where the site or asset is set on that host.

- IP Address
- Protocol
- Number of Dots and Slashes
- Suspicious Character @ and %40
- Multiple Occurrence (.com, https, http)
- Keyword Check
- Company Check

### D. Classification Algorithm

The input to the classifiers in MATLAB is two .txt files; newben.txt and newphis.txt. The classification algorithm considered for processing the feature set is:

#### 1) *Random Forest Classifier*

The random forest technique modifies the Bagging method applied here to trees by adding a de-correlation criterion between these trees. The idea behind this method is to reduce the correlation without increasing the variance too much. The principle is to randomly choose a subset of variables that will be considered at each level of choice of the best node of the tree.

Consider a training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, $a$ has the number of attributes of the examples of $X$. Also consider $S_t$ a bootstrap containing $m$ instances obtained by resampling with replacement of $S$. Let $\{h_1, \ldots h_t\}$ be set of $T$ decision trees. Each tree $h_t$ is built from $S_t$. For each node of the tree, the partitioning attribute is chosen by considering a number $f\,(f < a)$ of randomly selected attributes (among the attributes $a$). To classify a new instance $x$, the random forest classifier performs a uniformly weighted majority vote of classifiers in that set for instance $x$. The algorithm illustrates this principle.

***Algorithm:***

```
Input: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, the training set.
Input: $T$, the number of decision trees in the random forest.
  For $t = 1, \ldots, T$ do
    1. Generate a Bootstrap sample $S_t$ of size $m$ from $S$
    2. Create a decision tree $h_t$ from $S_t$ by recursively repeating for each node of the tree the following steps:
       a. Randomly select $f$ attributes among $a$ attributes.
       b. Choose the partitioning attribute among $f$
       c. Partition the node into two child nodes
  End for
Output: $H$, the random forest classifier
```

#### 2) *Bayesian Optimization of Random Forest Classifier*

A direction for Bayesian optimization is to optimize continuous and mixed (discrete and continuous) variables in solving problems with various types of data. The main objective of using Bayesian optimization here is to find the suitable value for each parameter of random forest algorithm. There are at least three important practical choices that we need to consider: the covariance functions, selection of its hyperparameters and the acquisition functions. A default choice of covariance function is to use squared exponential kernel. Automatic relevance determination (ARD) Matern 5/2 kernel is used for the same [10].

$$K_{M52}(x, x') = \theta_0\left(1 + \sqrt{5r^2(x, x')} + \frac{5}{3}r^2(x, x')\right)\exp\left\{-\sqrt{5r^2(x, x')}\right\} \quad (1)$$

The above kernel function itself has few parameters that needs to be managed (such as covariance amplitude $\theta_0$ and the observation noise $v$). It can be done by marginalize over hyperparameters and compute the integrated acquisition function.

## III. Simulation Results

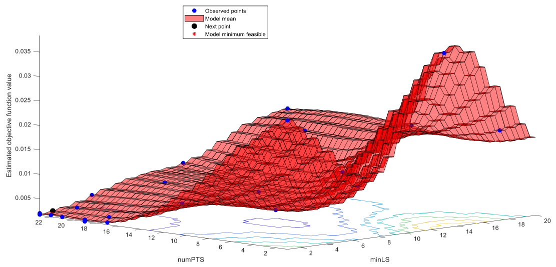

Figure 2: Objective function model

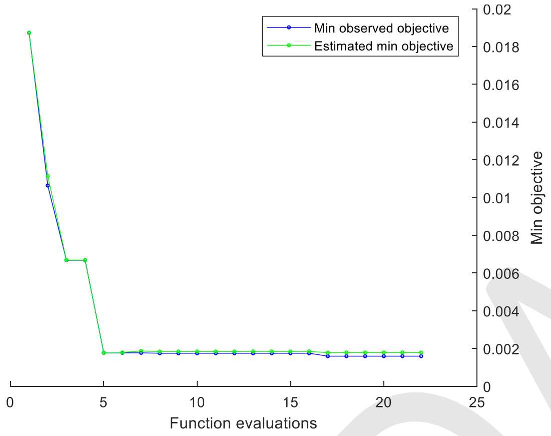

Figure 3: Min. objective vs. number of function evaluations

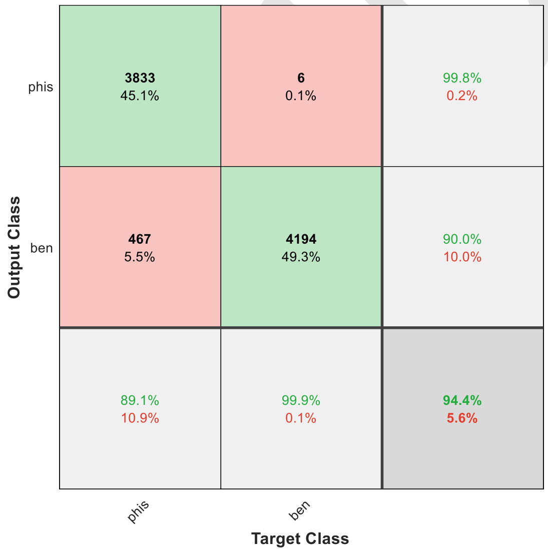

Figure 4: Confusion matrix plot for Random Forest classifier method

Here, TP=3833, TN=4194, FP=6 and FN=467

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3833+4194}{3833+4194+6+467} = 94.4\%$$

$$Precision = \frac{TP}{TP+FP} = \frac{3833}{3833+6} = 99.84\%$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{3833}{3833+467} = 89.13\%$$

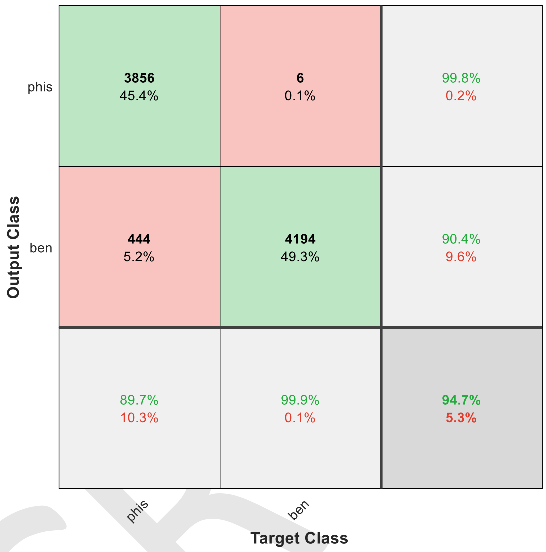

Figure 5: Confusion matrix plot for Bayesian optimized Random Forest classifier method

Here, TP=3856, TN=4194, FP=6 and FN=444

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3856+4194}{3856+4194+6+444} = 94.7\%$$

$$Precision = \frac{TP}{TP+FP} = \frac{3856}{3856+6} = 99.84\%$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{3856}{3856+444} = 89.7\%$$

## IV. Conclusion

The database "phishing web" offers a number and variety of attributes established by all the literature, however, the tests carried out show that of the 60-40 split case is presented in the simulation, nevertheless, it is proposed a possible consensus on the attributes that can come to clearly define a phishing URL. On the other hand, the amount of consigned attributes turns out to be an inconvenience due to the "curse of the dimension", since understanding and processing all these attributes translates into space, time and costs.

In this work, the gain that occurs when using classification techniques such as Bayesian optimized Random Forest classifier is revealed at the theoretical level, even though no technique is superior to the others in a general way, since they have limitations and own advantages that are coupled according to the model we are working with. A future study that uses specific data on the subject would probably help to understand the adoption of certain risk behaviours.

## Reference

[1] Kritzinger, Elmarie, and Sebastiaan H. von Solms. "Cyber security for home users: A new way of

protection through awareness enforcement." *Computers & Security* 29, no. 8 (2010): 840-847.

[2] Hille, Patrick, Gianfranco Walsh, and Mark Cleveland. "Consumer fear of online identity theft: Scale development and validation." *Journal of Interactive Marketing* 30 (2015): 1-19.

[3] Gupta, Brij B., Aakanksha Tewari, Ankit Kumar Jain, and Dharma P. Agrawal. "Fighting against phishing attacks: state of the art and future challenges." *Neural Computing and Applications* 28, no. 12 (2017): 3629-3654.

[4] Phishlabs. (2017). *2017 Phishing Trends and Intelligence Report: Hacking the Human*. Retrieved from: https://pages.phishlabs.com/rs/130-BFB-942/images/2017%20PhishLabs%20Phishing%20and%20Threat%20Intelligence%20Report.pdf

[5] PhishMe. (2016). *PhishMe Q1 2016 Malware Review*. Retrieved from https://phishme.com/project/phishme-q1-2016-malware-review/

[6] Jaeger, J.-M. D. (2016). Des pirates volent 72 millions de dollars à une plateforme de Bitcoin. Retrieved from http://www.lefigaro.fr/secteur/high-tech/2016/08/03/32001-20160803ARTFIG00143-des-pirates-volent-72-millions-de-dollars-a-une-plateforme-de- bitcoin.php

[7] Phishing-Initiative. (2017). Phishing Initiative. Retrieved from http://www.phishing-initiative.com/

[8] “PhishTank,” Online Available at: https://www.phishtank.com/

[9] “The Web Information Company,” Online Available at: www.alexa.com.

[10] Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. "Practical bayesian optimization of machine learning algorithms." In Advances in neural information processing systems, pp. 2951-2959. 2012.