Detection of Phishing Websites using an Efficient Feature-Based Machine Learning Framework

V. V. Ramalingam, Paras Yadav, Prakhar Srivastava

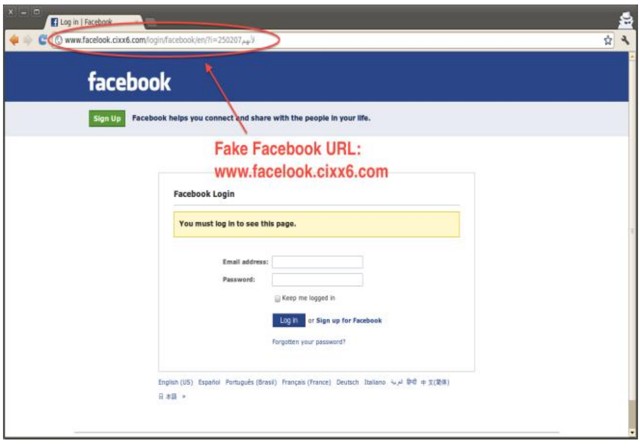
Abstract: ***Phishing is a cyber-attack which is socially engineered to trick naive online users into revealing sensitive information such as user data, login credentials, social security number, banking information etc. Attackers fool the Internet users by posing as a legitimate webpage to retrieve personal information. This can also be done by sending emails posing as reputable companies or businesses. Phishing exploits several vulnerabilities effectively and there is no one solution which protects users from all vulnerabilities. A classification/prediction model is designed based on heuristic features that are extracted from website domain, URL, web protocol, source code to eliminate the drawbacks of existing anti-phishing techniques. In the model we combine some existing solutions such as blacklisting and whitelisting, heuristics and visual-based similarity which provides higher level security. We use the model with different Machine Learning Algorithms, namely Logistic Regression, Decision Trees, K-Nearest Neighbours and Random Forests, and compare the results to find the most efficient machine learning framework.***

Keywords: ***Machine Learning, Blacklist, Whitelist, Cyber-attacks, Logistic Regression, K-Nearest Neighbours***

I. INTRODUCTION

The age of digitization is here. The internet has proven to be a boon to people throughout the world. Each task ranging from buying groceries to handling bank statements can be done with just a few clicks. Governments are encouraging their people to join the movement be a part of the era of digitization which has motivated the masses to be more and more active on the Internet. Each year, millions of people who didn't have access to the internet join this network which has provided a huge customer base for business to flourish and economies to skyrocket. A lot of these people are unversed to the threats that they are exposed to once they surf the internet. This forms a huge population of vulnerable people that can be targeted by an army of cyber criminals waiting to hunt down pregnable internet users. Tech companies have invested millions of dollars to ensure the security of their users to prevent them from being prey to cyber-attacks by creating smart anti-viruses, impenetrable firewalls, green padlocks etc. due to which a considerable decline in cybercrime can be seen. It has also eliminated certain kinds of cyber-attacks.

The elimination of certain Kinds of cyber-attacks saw an increase in the number of users being targeted by phishing attacks. The difficulty in carrying out attacks from a technical aspect encouraged attackers to exploit the social aspect of the internet. The goal of a safe and terror free internet has encouraged us to develop a real-time phishing detection system that may block all doors for cyber-crime. The project provides a compact and suitable measure to detect phishing websites and to alert the users of to protect themselves from being tricked into revealing their information to cyber criminals.



Facebook helps you connect and share with the people in your life.

You must log in to see this page.

Facebook Phishing Scam URL

II. STATE OF THE ART

In an attempt to minimize phishing attacks and to increase the accuracy of phishing detection, several techniques have been developed over the years. Some of these techniques are:

A. Blacklist

Blacklist is essentially a large database of URLs and web pages which are known to be used by hackers to steal the User's confidential information or to install malicious malwares on the User's computer. Blacklists act as the first line of defence and are extremely valuable sources used by cybersecurity applications to prevent users from accessing phishing web pages. They are also used in email filters to detect phishing emails or spam. Despite being very useful, Blacklists are not always effective and have a very low success rate in real world applications, this is because hackers keep using new URLs and web pages over time. To improve the effectiveness of blacklists it is very important that anti-phishing organisations and blacklist publishers keep updating the blacklist database in frequent intervals and are able to keep up with the attackers.

Revised Manuscript Received on February 15, 2020.

V. V. Ramalingam, Department of Computer Applications from Bharadhidasan University (2000), M.Phil Degree in Computer Science from Periyar University (2007).

Paras Yadav, Department of Computer Science and Engineering, B.Tech student in SRM Institute of Science and Technology.

Prakhar Srivastava, Department of Computer Science and Engineering, B.Tech student in SRM Institute of Science and Technology.

B. Whitelist

Websites can be categorized into two categories i.e. legitimate websites and phishing websites. A list of all known legitimate websites is created which indicate that the websites given in the whitelist are safe to use and does not pose any threats to the users. In other words, a whitelist is a list of Websites approved for authorized access or privileged membership to enter a specific area in the computing world. Whitelisting prevents access to websites outside of the whitelist. It blocks unknown and suspicious programs. This prevents threats like malware from entering your system or network. Whitelisting makes it possible to undergo IP address whitelisting which restricts the browser to from accessing websites not present in the whitelist. Whitelisting can be considered as the opposite of blacklisting and is a useful tool to secure users from various phishing attacks.

C. Baitalarm

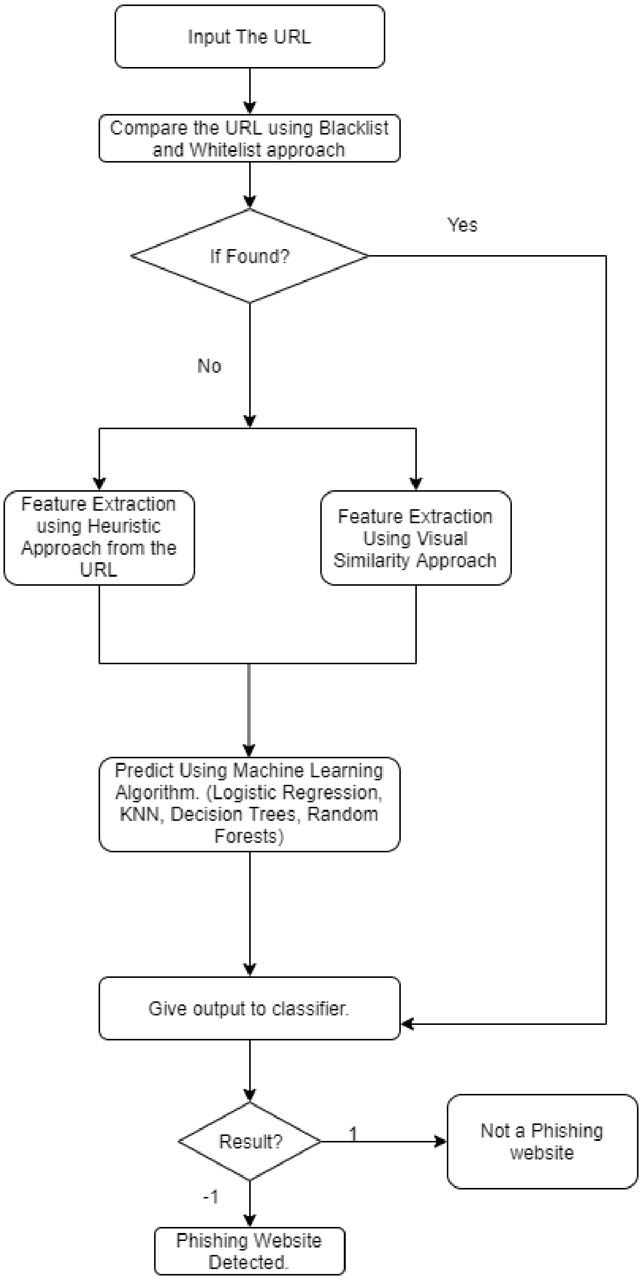
Baitalarm is a software which relies on a visual similarity approach in order to detect phishing web pages. It looks for similarities in the structure of web pages. A web page's appearance is made up of two components, namely, its content and its layout. The most conventional method of specifying the layout of web pages is by using Cascading style Sheets (CSS), Baitalarm exploits this aspect of web pages and uses various algorithms to detect similarities between the various elements present in the CSS of the web pages. Every time a user loads a new web page, the browser requests for its CSS file from the server. The algorithm compares the content and layout specifications of two web pages and generates a similarity score as output.

D. Goldphish

Goldphish is another system which employs a visual similarity approach to identify whether a web page is a phishing web page or not. It uses the images of web pages to perform content analysis and detect phishing websites. At the core of Goldphish is the idea that the logos and symbols used by most well-known organizations or brands have text in them which is particular to that company. It is very difficult to modify these logos without alerting the target. Goldphish uses an Optical Character Recognition software (OCR) to extract text from the images. The extracted text is then fed into a search engine such as Google. In the result produced by the search engine, we can be sure that one of the first four links will belong to legitimate websites, this is because such web sites have a high page rank score. On the other hand temporarily created phishing websites will have a very low page rank score as they have only been online for a limited amount of time and hence will not be indexed in the top search results. Goldphish is now rarely used by organizations because it faces several issues. Firstly it increases the amount of time needed to load a web page, secondly it is also susceptible to attacks on Google's PageRank and Search algorithm

III. METHODOLOGY

Our proposed architecture builds on top of the existing, state of the art Blacklist-Whitelist technique to detect phishing websites. It can also be used in addition to Baitalarm and Goldphish techniques to perform a more comprehensive analysis of the website which can lead to better results and provide better protection to users from such malicious attacks.


Input The URL
Compare the URL using Blacklist and Whitelist approach
If Found?
Yes
No
Feature Extraction using Heuristic Approach from the URL
Feature Extraction Using Visual Similarity Approach
Predict Using Machine Learning Algorithm. (Logistic Regression, KNN, Decision Trees, Random Forests)
Give output to classifier.
Result?
1
Not a Phishing website
-1
Phishing Website Detected.

In our model, every time the user wants to visit a website, the URL of the website is first checked against a database of known dangerous phishing websites (Blacklist) and a database of URLs which are known to be safe (Whitelist). If the input URL is found in either of the list then we can immediately report the result to the user and block the web page if it is a known phishing website. The Blacklist and Whitelist are continuously updated as more phishing pages are discovered to save time and computing resources in the future.

We cannot always rely on the Blacklist because most of the time hackers employ Single-Use Phishing URLs to target users and hence these URLs will never be present in the Blacklist. Therefore if a URL is not found in the Blacklist, then we move on to the next stage of the model. Here we will use various machine learning algorithms to predict whether the web page is a phishing web page.

Retrieval Number: C5909029320/2020©BEIESP
DOI: 10.35940/ijeat.C5909.029320

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Before we can use the machine learning algorithms to make predictions, we first need to perform feature extraction from the URL. We use heuristic techniques to extract 30 features from the URL such as, URL length, presence of @ symbols, port number, URL prefix and suffix, Google Index of the page, whether it has any sub domains, presence of redirecting tags, age of the domain etc.

To make predictions we have evaluated the performance of 5 different Machine Learning Algorithms, namely, Logistic Regression, K-Nearest Neighbours, Support Vector Machines, Decision Trees and Random Forests.

A. Logistic Regression.

In Logistic Regression we use statistical analysis for predictive analytics and modelling to find the relationship between the dependent and independent variables by estimating probabilities using the logistic regression equation. The algorithm learns from the training data and assigns weights to all the individual features. When we encounter a new example the learnt weights are multiplied with the respective features and summed. The result is then fed into the Sigmoid Function to get the probability.

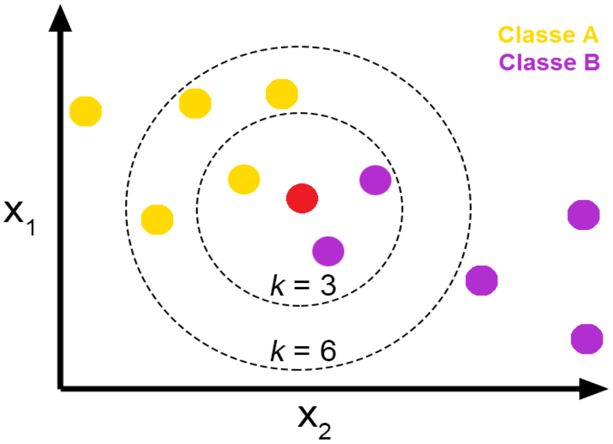
Here, we use Logistic Regression to assign the URL to one of two classes, Phishing (1) and Non-Phishing (0). We have set the probability threshold to 0.5. If the probability predicted by Logistic Regression is more than 0.5 then we assign the URL to the Phishing Class.

B. K-Nearest Neighbours

K-Nearest Neighbours Algorithm (KNN) is a nonparametric statistical algorithm used for solving regression and classification problems. The algorithm does not make any assumptions about the distribution of the data and is independent of it.. Here we have used KNN as a classification algorithm to distinguish between Phishing and Non-Phishing websites.

The KNN algorithm assumes that similar things will have similar features and hence will appear in close proximity to one another. For classification problems, the output of the algorithm is a class membership. The class of an object is assigned by determining the class of its K closest neighbours.The training data for KNN are multi-dimensional feature vectors, each of which has a class assigned to it. While classifying a new object, we take its unlabeled feature vector and calculate the distance of the vector from every other feature vector in the existing dataset.

The most commonly used distance metric is the Euclidean Distance, which is also what we have used here. We choose the Euclidean distance because it is the fastest to calculate. With a large number of training examples it is very important to be able to calculate the result fast to avoid delays in loading the webpage.



Once all the distances have been calculated, we take a majority vote of the K closest neighbours to assign to the unlabeled feature vector.

The selection of the parameter K depends on the data. Small values of the parameter are very susceptible to noise and the predictions become less stable. Larger values reduce the effect of noise, and give better predictions until a certain point after which it starts making the boundaries between classes become less distinct and the accuracy falls. In our attempts we found that we get the best accuracy when the value of the parameter K is 4.

C. Recursive Feature Elimination(RFE)

A person visits hundreds of web pages every day. If for every web page the user loaded, there was a delay of a few seconds to determine if the web page was a phishing page or not, it would have an adverse impact on the user experience and users would stop using the service leaving them vulnerable. It is very important that during runtime, the model produces results quickly. This is why we have chosen to use Recursive Feature Elimination to find the optimal features which have the highest impact on the result.

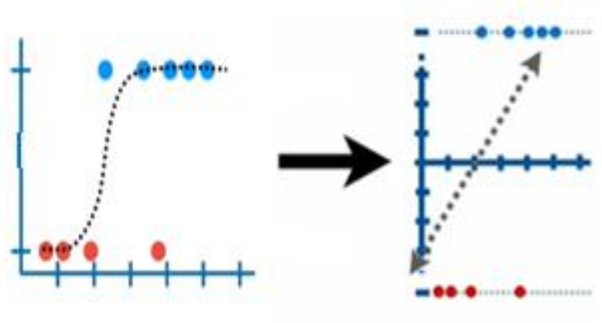
RFE aims to remove dependencies and collinearity that may exist among the features in the model by recursively eliminating a small set of features in every iteration. We have chosen the 3 as the number of features we want to keep for optimal performance.

IV. MATH

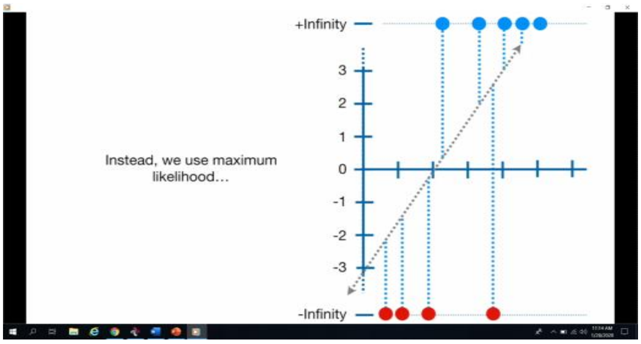
1.Logistic regression is a statistical model that uses a logistic function to make binary classifications (problems involving two class values). Logistic regression predicts discrete values rather than predicting continuous values. For this it relies on a S-shaped logistic function whose values range between 0 to 1, which is unlike linear models, which try to fit a straight line to the data. In our case, 0 refers to a legitimate website and 1 refers to a phishing website. Probability is calculated for each data point. For each data point, the probability is converted into log(odds) which maps the y-axis from a scale of 0 to 1(probability) to a scale of -∞ to +∞ using the logit function (1).

Odds = no. of times event occurring ÷ no. of times event not occurring

This converts the S-shaped function into a linear model.

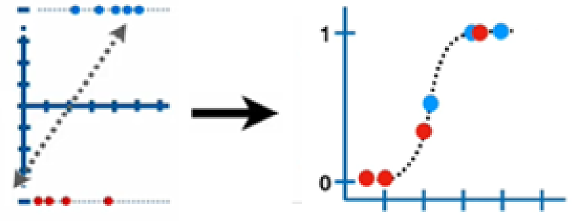


Logistic regression cannot use sum of squared residuals(R^2) As the data points are pushed to -∞(legitime website) and +∞(phishing website). So the distance of data points from the fit line will also be infinite. Thus , another approach is used which is known as maximum likelihood. The goal of maximum likelihood is to find an optimal way to fit distribution to the data. The data points are projected from -∞ and +∞ onto the fit line and the value corresponding to the y-axis is measured.



Instead, we use maximum likelihood...

These values are in terms of log(odds) and probabilities are calculated using the reverse logit function(2). The Probabilities are mapped onto the S-shaped curve.



To find the likelihood of the fit line, the product of log of probabilities of all data points is calculated. This is the likelihood of the given fit line. This process is repeated again and again until an optimal fit line (i.e with maximum likelihood value) is found. The fit line is rotated and the likelihood of the new fit line is calculated. The algorithm that rotates the fit line is smart enough to rotate it in a way that increases likelihood. Thus, optimal solutions can be found after a few rotations only. The value of probabilities are mapped onto the S-shaped curve which is used to make binary classification.

V. HELPFUL HINTS

A. Abbreviations and Acronyms

- ML: Machine Learning
- RMSE: Root Mean Squared Error
- LR : Logistic Regression
- RFE: Recursive Feature Elimination
- URL : Uniform Resource Locator
- MSE: Mean Squared Error
- OCR: Optical Character Recognition
- MAE: Mean Absolute Error

B. Equation

(1) Logit Function:

$$\mathrm{Log(odds)} = \log((p) \div (1-p))$$

Where p = probability of event occurring

(2)Inverse Logit Function:

$$p=(e^{\log\ (odds)}) \div (1+ e^{\log\ (odds)})$$

Where P=probability of event occurring.

(3)Euclidean Distance:

$$dist\ (A, B) = \sqrt{\frac{\sum_{i=1}^{m} (x_i - y_i)^2}{m}}$$

C. Figures and Tables

Table- I: Name of the Table that justify the values

S.No.	*Algorithm*	*Accuracy*
1	Logistic Regression	91.60
2	KNN	93.7

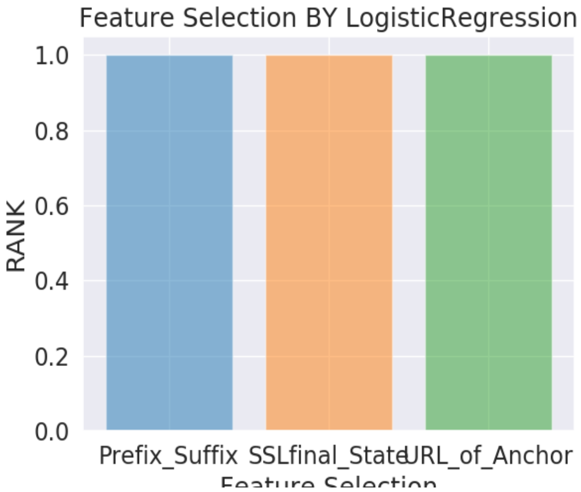
The figure, graph, chart can be written as per given below schedule.

VI. RESULTS AND DISCUSSION

The output and the accuracy of the various algorithms can be seen below. For testing the accuracy of the model we followed a Train/Test split of 75% Training Data and 25% testing data.

A. Logistic Regression.

When we applied RFE along with a logistic regression classifier, the RFE selected the following 3 features as being most relevant in predicting the result.

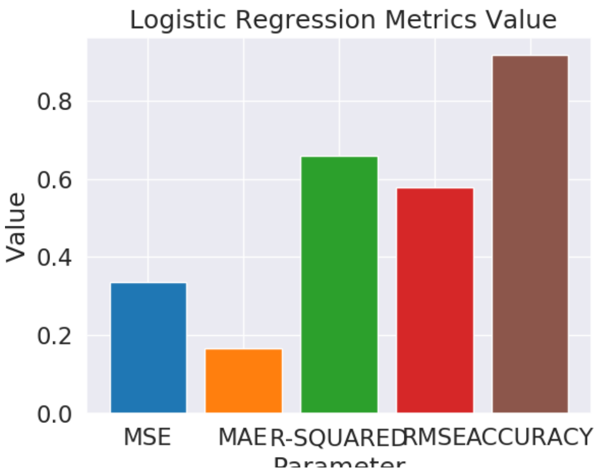


The feature *Prefix_Suffix*, indicates the presence of prefixes and suffixes separated by hyphens(-), to the domain name. The feature *SSLFinal_State*, indicates whether the Security certificate issued to the server is by a trusted source or not.

Retrieval Number: C5909029320/2020©BEIESP
DOI: 10.35940/ijeat.C5909.029320

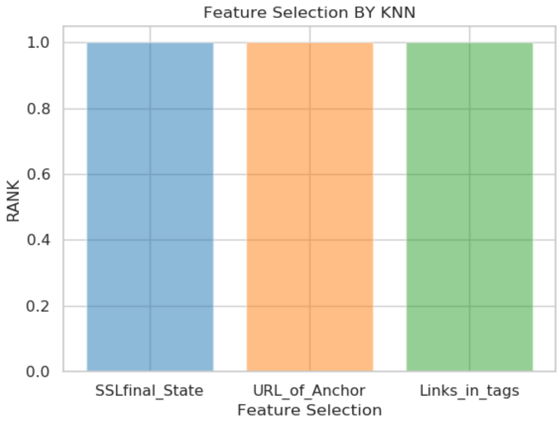
The feature *URL_of_Anchor* is set to 0 if the anchor in the URL tag does not link to any web page.

The accuracy of the Logistic Regression algorithm as measured by various metrics is given below.



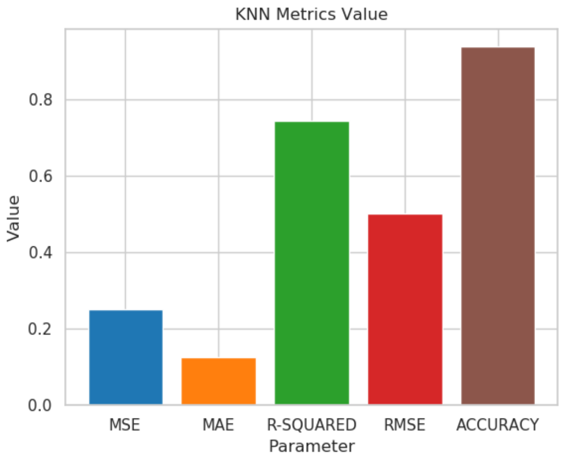
B. K-Nearest Neighbours.

When we applied RFE to the KNN algorithm, it returned the following 3 features.



In case of KNN, instead of checking for the presence of Prefixes and Suffixes in the URL, the algorithm selected the *Links_in_tags* feature, which indicates whether the URL includes metadata about the HTML in tags. Phishing websites usually do not have these tags.

The accuracy of KNN algorithm as measured by various metrics is given below.



VII. CONCLUSION

We explored the drawback of the current state of the art systems in detecting phishing attacks. It is evident that these systems in their present state do not protect users from one of the most common forms of attack on the internet.

Machine Learning is the perfect candidate suited to identifying phishing web pages as it can automatically learn to identify phishing websites.

Here we have compared the efficiency of two different machine learning algorithms in identifying the URL of a malicious website. As we can see, in our experiment, Logistic Regression gave an accuracy of 91.60%, whereas, K-Nearest Neighbours gave an accuracy of 93.7% in detecting phishing web pages. Based on our results, we would recommend using the KNN algorithm to identify phishing websites. Our results also show that using RFE for feature elimination not only improves performance at run time but also improves accuracy of the model by eliminating redundant features.

Future researchers can examine the performance of more advanced machine learning algorithms and also see how the efficiency of algorithms changes as larger amounts of training data is fed into the algorithms.

REFERENCES

1. Ankit Kumar Jain and B. B. Gupta, "Phishing Detection Analysis of Visual Similarity Based Approaches", Hindawi 2017.
2. A. Mishra and B. B. Gupta, "Hybrid Solution to Detect and Filter Zero-day Phishing Attacks", ERCICA 2014.
3. Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang, "Bait Alarm Detecting Phishing Sites Using Similarity in Fundamental Visual Features", INCS 2013.
4. Eric Medvet, Engin Kirda and Christopher Kruegel, "Visual-Similarity-Based Phishing Detection", ACM 2015.
5. Matthew Dunlop, Stephen Groat, and David Shelly, "GoldPhish Using Images for Content-Based Phishing analysis", IEEE 2010.
6. Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu,"Textual and Visual Content-Based Anti-Phishing A Bayesian Approach", IEEE 2011
7. Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner; "Lexical Feature Based Phishing URL Detection Using Online Learning", Department of Computer and Information Sciences The University of Alabama at Birmingham, Alabama, 2016
8. Pawan Prakash, Manish Kumar, Ramana Rao Kompella, MinaxiGupta,Purdue University, Indiana University "PhishNet: Predictive Blacklisting to Detect Phishing Attacks".
9. The Anti-Phishing Working Group, DNS Policy Committee;" Issues in Using DNS Whois Data for Phishing Site Take Down",The Anti-Phishing Working Group Memorandum, 2011.
10. Mohsen Sharifi and Seyed Hossein Siadati "A Phishing Sites Blacklist Generator".
11. JungMin Kang and DoHoon Lee "Advanced White List Approach forPreventing Access to Phishing Sites".
12. Joby James,SandhyaL,Ciza Thomas "Detection of phishing websites using Machine learning techniques", 2013 International Conference on Control Communication and Computing (ICCC).

AUTHORS PROFILE

Dr. V. V. Ramalingam, received Post graduate degree in Master of Computer Applications from Bharadhidasan University (2000), M.Phil Degree in Computer Science from Periyar University (2007) and M.Tech Degree in Computer Science from S.R.M University (2012) and completed his Ph.D in Bharathiar University (2017), Coimbatore. He also published more than twenty papers in International Journals and Conferences.

His research interest is in the area of Data Mining using Machine Learning Approach.

ParasYadav, Currently a final year Computer Science and Engineering, B.Tech student in SRM Institute of Science and Technology. He is a hard-working student and has immense interest in the field of research. Research in Machine Learning field really excites him and wishes to do more such works and gain further knowledge and recognition relating to the subject in future.

Prakhar Srivastava, Currently a final year Computer Science and Engineering, B.Tech student in SRM Institute of Science and Technology. He is a meticulous student and has immense interest in the field of research. Research in Machine Learning field really excites him and wishes to do more such works and gain further knowledge and recognition relating to the subject in future.