International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST229149

# Phishing Detector Extension Using Machine Learning

**Sushil Nagre, Sumeet Mathapati , Ketan Sutar , Rajkumar Suryawanshi**

Student, Department of Computer Engineering, Zeal College of Engineering & Research Pune, Maharashtra, India

**Article Info**
Volume 9, Issue 2
Page Number : 488-491

**Publication Issue**
March-April-2022

**Article History**
Accepted : 03 April 2022
Published : 20 April 2022

## ABSTRACT

The goal of our project is to implement a machine learning solution to the matter of detecting phishing and malicious web links. The tip results of our project are going to be a software package which uses machine learning algorithm to detect malicious URLs. Phishing is that the technique of extracting user credentials and sensitive data from users by masquerading as a real website. In phishing, the user is supplied with a mirror website which is clone of the legitimate one but with malicious code to extract and send user credentials to phishers. Phishing attacks can cause huge financial losses for patrons of banking and financial services. the normal 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 unproved and error to define the edge which is employed to classify malicious links from benign ones. the disadvantage to the current approach is poor accuracy and low adaptability to new phishing links. We attempt to use machine learning to beat these drawbacks by implementing some classification algorithms and comparing the performance of those algorithms on our dataset. we are going to test algorithms like Logistic Regression, SVM, Decision Trees and Neural Networks on a dataset of phishing links from UCI Machine Learning repository and pick the simplest model to develop a browser plugin, which might be published as a chrome extension.

## I. INTRODUCTION

Financial services such as banking are now readily available on the web which makes people's lives easier. It is therefore essential that the protection and security of such services be maintained. one of the biggest threats to web security is cybercrime. Phishing scams are a way to extract user information by pretending to be a real website or service over the net. There are a variety of cybercrime attacks such as Spear sensitive identity theft, targeted at specific individuals or companies, Clone sensitive identity theft is a form of sensitive identity theft when a compiled email or link is copied to a new mail with special attachments (possible) link, Whaling, etc.

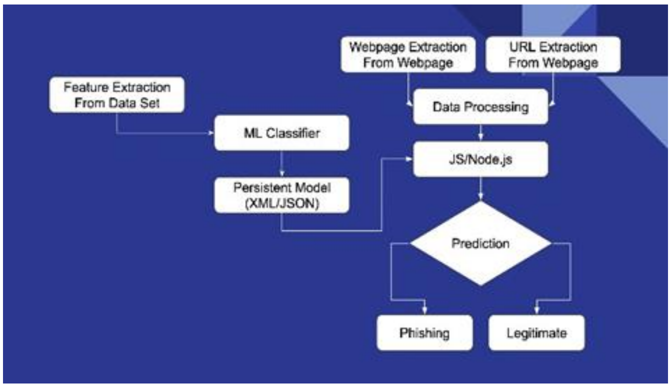
Theft of sensitive information can result in significant financial losses.

Similarly, the IRS has warned of an increase in attacks on sensitive identity theft with an increase of more than 400% on reported cases.

In our project, we will test three machine learning algorithms in the database of features that represent the features that are often associated with the pages of identity theft, select the most effective model supported by its functionality and build an online browser plugin that can ultimately be used by users. The project report is designed as follows; The Pre-Work section describes the common methods of crime detection and a few mechanical learning methods tried in recent years, the Proposed Approach section very well describes our approach and what will be the top product of our project, the Database section describes the data we use for our project and a list of for use in our project, the Machine Learning Algorithms section describes the various algorithms we have tested with our database in terms of their meanings, The Chrome Plugin implementation section explains the structure of our sensitive identity theft system and provides descriptions of various software modules within the system, the Outcomes section provides test results of us with graphical algorithms that arrange comparisons between three algorithms for features such as accuracy, sensitivity and falsehood. a good rating, and the concluding section summarizes the pr objective with a view to future developers.

## II. METHODS AND MATERIAL

We suggest using machine learning to overcome obstacles associated with common methods of detection of identity theft. The problem of detection of sensitive identity theft is a very relevant candidate for the use of machine learning solutions due to the easy availability of sufficient amounts of information on sensitive crime attack patterns. the basic idea is to use machine learning algorithms in the available databases of sensitive data theft pages to come up with a model that may tend to perform categories in real time if the provided web content may be a criminal webpage or official web page. We will produce a learned model into a software tool that can be easily used to complete users in order to combat criminal attempts to steal sensitive information. For this purpose we have chosen to use a machine learning algorithm from the beginning using JavaScript and building Chrome.


Webpage Extraction From Webpage
URL Extraction From Webpage
Feature Extraction From Data Set
Data Processing
ML Classifier
JS/Node.js
Persistent Model (XML/JSON)
Prediction
Phishing
Legitimate

## III. RESULT AND DISCUSSION

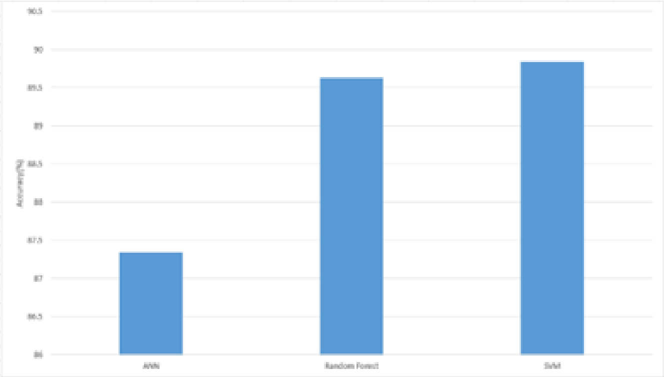
We have trained and tested supervised machine learning algorithms in training Database. The following algorithms are selected based on their performance in categories problems. The database was divided into training and the test was set at 7:3.

### A. Random Forest

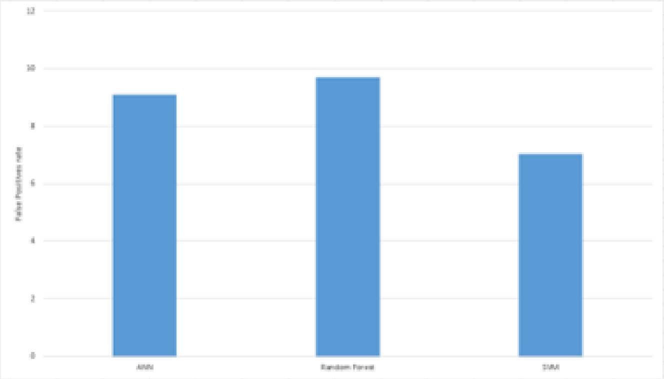
Random forests are subdivisions that include multiple tree predictions, where each tree is based on random vector values independently sampled. Moreover, all the trees in the forest are propagated in the same way. To construct a tree, we assume that n is the number of training observations and p is the number of variables (features) during the training set. identifying the selection node in the tree determines k «p as the number of variables to be selected. We selected a bootstrap sample from the n test within the training set and used the rest of the recognition to estimate the tree error within the test phase. Thus, we randomly select the k variable as a fence somewhere in the tree and calculate the most effective division supporting the k variable within the training set. Trees grow regularly and have never been pruned compared to other tree-straightening methods. Random forests can handle a large number of variables in a data set. Also, during the forest-building process they produce an unbiased internal measure of common error. in addition, they will properly balance missing data.

### B. SVM:

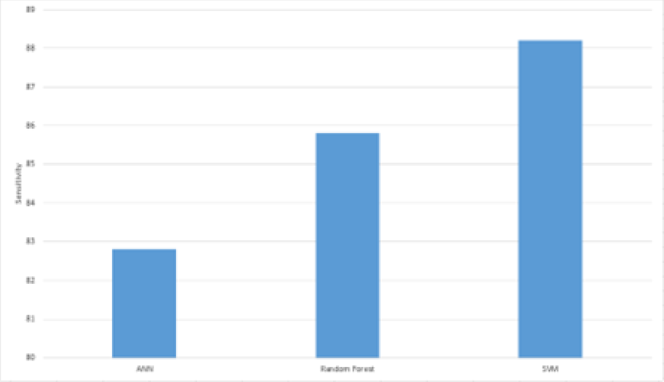
Support Vector Machine (SVM) is a discriminatory machine-controlled machine, which conforms to the design principle of separating a large aircraft with a high safety zone, called a margin, to reduce the risk of error predictors.



Accuracy of Classifiers



False Positive Rate of Classifiers



Sensitivity of Classifiers

## IV. CONCLUSION

So in summary, we have seen how serious cybercrime threatens web security and security and how dangerous identity theft is.

Figure 4: False level of class dividers

adoption is a very important problem area. review a few common methods of detection of identity theft; namely methods to evaluate the list of blocked and heuristic, and their problems. we tested three machine learning algorithms in the 'Phishing Websites Dataset' from the UCI Machine Learning Repository and reviewed their results. We then selected a simple algorithm that supports its functionality and built the Chrome extension to detect web content of sensitive identity theft. The extension allows easy use of our crime detection model to steal sensitive information for users. For future enhancements, we will build a system to detect sensitive identity theft as a fast web service that is able to integrate online learning so that new patterns of sensitive identity theft can be easily read and improve the accuracy of our models by extracting better features than the core problem. We have reviewed some of the most common ways to detect identity theft; namely methods to evaluate the list of blocked and heuristic, and their problems. We tested three machine learning algorithms in the 'Phishing Websites Dataset' from the UCI Machine Learning Repository and reviewed their results. We then selected the best algorithm based on its functionality and built the Chrome extension for accessing web pages to steal sensitive information.

The extension allows easy use of our crime detection model to steal sensitive information from end users. For future improvements, we aim to build a system to detect sensitive identity theft as a fast web service that will integrate online learning so that new patterns of sensitive identity theft can be easily learned and improve the accuracy of our models by delivering better features.

## V. REFERENCES

[1]. Microsoft, Microsoft Consumer safety report. Available at https://news.microsoft.com/en-sg/2014/02/11/microsoft-consumer-safety-index-reveals-impact-of-poor- online-safety-behaviours-in singapore/sm.001xdu50tlxsej410r11kqvks u4nz

[2]. E., B., K., T. (2015)., Phishing URL Detection: A Machine Learning and Web Mining-based Approach. International Journal of Computer Applications,123(13),46-50. doi:10.5120/ijca2015905665

[3]. Sci-kit learn, SVM library. Available: http://scikit-learn.org/stable/modules/svm.html

[4]. Aparna Mote, Utakarsha Musmade,Nikita Deshmukh,"Gross Domestic Product (GDP) Prediction: A Review", 2019 JETIR May 2019, Volume 6, Issue 5.

[5]. Sci-kit learn, Random forests library. Available: http://scikit-learn.org/stable/modules/Random foresets. html