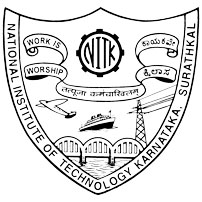
NATIONAL INSTITUTE OF TECHNOLOGY



IT302

WEB TECHNOLOGIES AND APPLICATIONS

PROJECT REPORT

SUBMITTED TO: Ms. Chaitra Bhat,

Department of Information Technology,

NITK

SUBMITTED BY: P Pavan [15IT224]

Lokaksha B M [15IT124]

Sachin K Rathod [15IT253]

Lakshmish M Gond [15IT215]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

TOPIC: Anomaly behavior analysis of website vulnerability and security

INTRODUCTION

The world wide web has grown exponentially over the previous decade in terms of its size that is currently over a billion sties, as well as the number of users. According to White Hat security's “2015 Website Security Statistics Report” more than 86% of all websites have one or more critical vulnerability and the likelihood of information leakage is 56%.

The main research goal of this paper is to overcome this challenge by presenting an online anomaly behavior analysis of websites (e.g., HTML files) to detect any malicious codes or pages that have been injected by web attacks. Our anomaly analysis approach utilizes feature selection, data mining, data analytics and statistical techniques to identify accurately the webpage contents that have been compromised or can be exploited by attacks such as phishing attacks, cross site scripting attacks, html injection attacks, malware insertion attacks, just to name a few.

We have validated our approach on more than 10,000 files and showed that our approach can detect malicious HTML files with a true positive rate of 99% and a false positive rate of 0.8% for abnormal files.

HTML ATTACK TYPES

**Hidden Iframes**-An Iframe (Inline Frame) is a way of loading one web page inside another, usually from a different server. It can be useful in creating online applications. Malware writers can make the included page hidden by making the Iframe as small as 0 pixels square, give them coordinates that make the frame off the page, or set their property to “hidden”. Then anything can be run inside the Iframe such as JavaScript or a remote malicious web page downloading code. The following example demonstrates hiding an Iframe by making it one pixel square:

**<iframe src=”http://www.xyz.com” width=”1px” height=”1px”>**

**Malicious Reference**-is a method used to link one page to a malicious page or to download a malicious file when clicked on. There are several tags that can be modified in either the HTML or CSS file that could cause these actions. The first method could be accomplished using the “meta” tag inside the HTML Head. This would look something like the following for malicious redirects:

**<meta content=”url=http://www.xyz.com” />**

**Malicious Scripts**-There are many ways to use ***JavaScript*** for malicious intent. The code in a ***JavaScript*** function is not executed when the function is defined. It is executed when the function is invoked. There are almost unlimited ways to use ***JavaScript*** maliciously and a few of them are in the following list:

 Read files from local drive

 Fill up a local drive

 Access or replace files on the local machine

 Close or open windows

 Launch an application

 Read browser history or cookies

 Exploit bugs in a browser

 Load another document or script from a different domain

WEB SCRAPING

Web scraping is data scraping used for extracting data from websites. We are using php and javascript for web scraping. We use the scraped data as features in the dataset for feature selection.

Web scraping a web page involves fetching it and extracting from it.Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list.

PYTHON CODE

from bs4 import BeautifulSoup as soup

import json

import requests

f=open('dataset1.csv','w')

f.write("a,meta,head,iframe,script,span,li,frame,p,div,input,img,select,form,option,link,base,param,embed,object,noscript,nav,Website\n")

urls={'Google':'https://www.google.co.in/','Microsoft':'https://www.microsoft.com/en-in','Nokia':'https://www.nokia.com/en\_int','Amazon':'https://www.amazon.in/','Twitter':'https://twitter.com/','Facebook':'https://www.facebook.com/','Youtube':'https://www.youtube.com/','Yahoo':'https://in.yahoo.com/?p=us','Apple':'https://www.apple.com/','Paytm':'https://paytm.com/','NITK':'http://www.nitk.ac.in/','Flipkart':'https://www.flipkart.com/','SBI':'https://www.onlinesbi.com/','Netflix':'https://www.netflix.com/in/','Myntra':'https://www.myntra.com/','Fmovies':'https://fmovies.to/','Indian Railways':'https://www.irctc.co.in/eticketing/loginHome.jsf','Hotstar':'http://www.hotstar.com/','IEEE':'http://ieeexplore.ieee.org/','EA Games':'https://www.ea.com/','Cricbuzz':'http://www.cricbuzz.com/','w3schools':'https://www.w3schools.com/','ZeeStudio':'http://www.ozee.com/zeekannada','Sony':'http://www.sony.co.in/'}

DATA-DRIVEN STATIC ANALYIS

FEATURE SELECTION METHODOLOGY

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. Here we are using feature selection for selecting the most relevant features in the html page that affect the security of the web page.

We have used Genetic Algorithm for Feature Selection.

GENETIC ALGORITHM

In a genetic algorithm, a [population](https://en.wikipedia.org/wiki/Population) of [candidate solutions](https://en.wikipedia.org/wiki/Candidate_solution) (called individuals, creatures, or [phenotypes](https://en.wikipedia.org/wiki/Phenotype)) to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties (its [chromosomes](https://en.wikipedia.org/wiki/Chromosome) or [genotype](https://en.wikipedia.org/wiki/Genotype)) which can be mutated and altered; traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

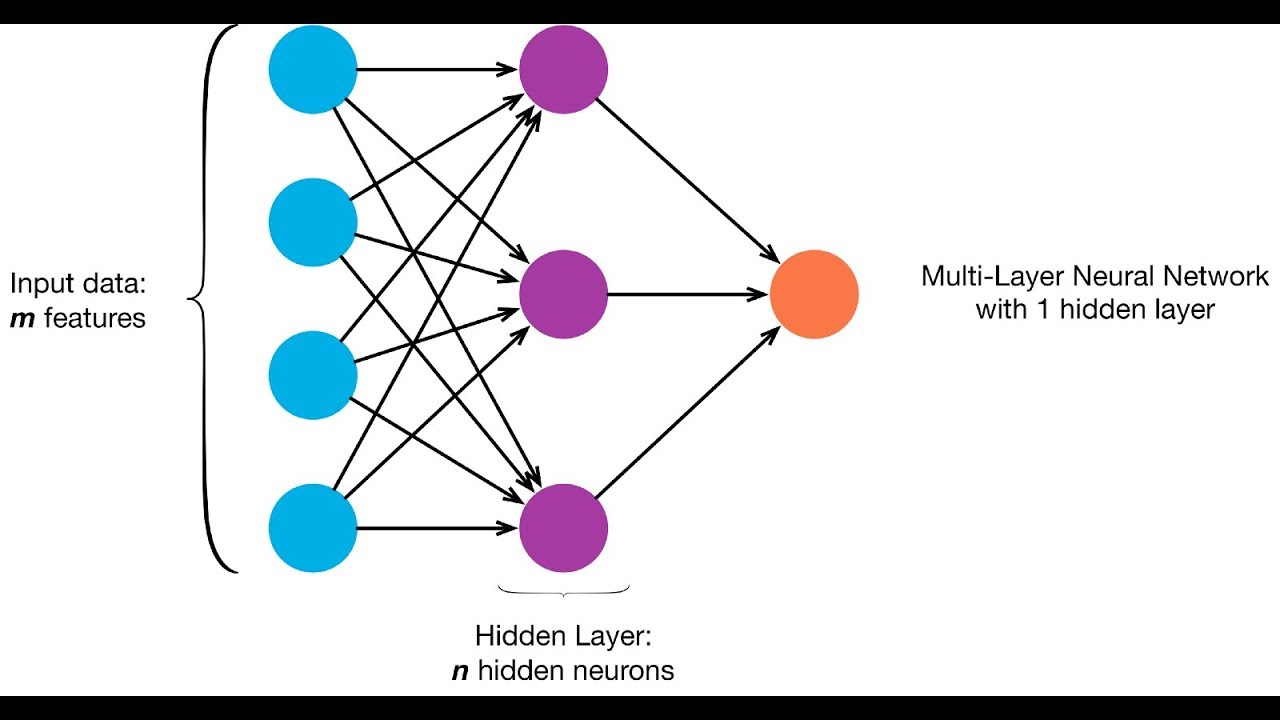
The evolution usually starts from a population of randomly generated individuals, and is an [iterative process](https://en.wikipedia.org/wiki/Iteration), with the population in each iteration called a *generation*. In each generation, the [fitness](https://en.wikipedia.org/wiki/Fitness_(biology)) of every individual in the population is evaluated; the fitness is usually the value of the [objective function](https://en.wikipedia.org/wiki/Objective_function) in the optimization problem being solved. The more fit individuals are [stochastically](https://en.wikipedia.org/wiki/Stochastics) selected from the current population, and each individual's genome is modified ([recombined](https://en.wikipedia.org/wiki/Crossover_(genetic_algorithm)) and possibly randomly mutated) to form a new generation. The new generation of candidate solutions is then used in the next iteration of the [algorithm](https://en.wikipedia.org/wiki/Algorithm). Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.

CLASSIFIERS

MULTI-LAYERED NEURAL NETWORK

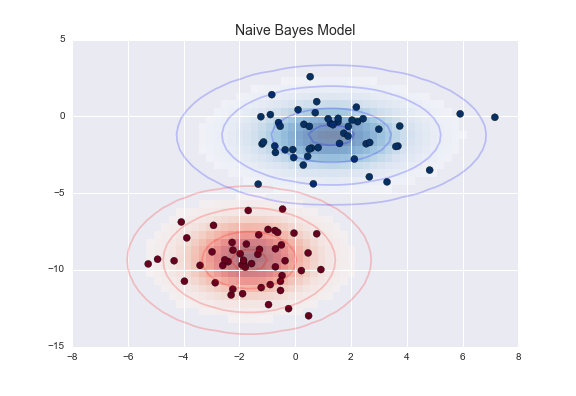
A multilayer neural network is a class of [feedforward](https://en.wikipedia.org/wiki/Feedforward_neural_network) [artificial neural network](https://en.wikipedia.org/wiki/Artificial_neural_network). An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear [activation function](https://en.wikipedia.org/wiki/Activation_function). MLP utilizes a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) technique called [backpropagation](https://en.wikipedia.org/wiki/Backpropagation) for training. Its multiple layers and non-linear activation distinguish MLP from a linear [perceptron](https://en.wikipedia.org/wiki/Perceptron). It can distinguish data that is not [linearly separable](https://en.wikipedia.org/wiki/Linear_separability).

If a multilayer perceptron has a linear [activation function](https://en.wikipedia.org/wiki/Activation_function) in all neurons, that is, a linear function that maps the [weighted inputs](https://en.wikipedia.org/wiki/Synaptic_weight) to the output of each neuron, then [linear algebra](https://en.wikipedia.org/wiki/Linear_algebra) shows that any number of layers can be reduced to a two-layer input-output model. In MLPs some neurons use a nonlinear activation function that was developed to model the frequency of [action potentials](https://en.wikipedia.org/wiki/Action_potentials), or firing, of biological neurons.



NAÏVE BAYE’S

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of [feature](https://en.wikipedia.org/wiki/Feature_vector) values, where the class labels are drawn from some finite set. It is not a single [algorithm](https://en.wikipedia.org/wiki/Algorithm) for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is [independent](https://en.wikipedia.org/wiki/Independence_(probability_theory)) of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible [correlations](https://en.wikipedia.org/wiki/Correlation_and_dependence) between the color, roundness, and diameter features.



CONCLUSION

This paper consists of a data-driven approach to detect malicious html files. The html files will be scraped for the data using python and then that data will be used for feature selection as well as for detecting vulnerability and security of the web page. Feature selection is done using Genetic Algorithm. After Genetic algorithm best set of features will be selected and the selected features will be classified using different classifier approach. Classifier used here are Naïve Bayes Classifier which uses conditional probability approach and Multi-Layer Neural Networks. The result of the classifier can be used to classify whether website is malicious or not.

REFERENCES

<http://ieeexplore.ieee.org/document/7945697/>