# Webpage classification for safer browsing

## Problem Statement

The web browser is one of the most heavily used programs on a computer or mobile device today. Because of its ubiquitous nature, it is also an extremely popular target for attackers.Attackers typically target the web browser to either hijack or snoop on the web traffic from it, or exploit it to access the device itself, and the files saved on it.

By way of people increasingly depend on internet for personal finance, business, investment; Internet fraud becomes a large threat. Through attractive websites and offers on Internet people want to take benefit and for this people share confidential data with the attacker. So, it is necessary to identify fake websites for end user's security. Once the user visits the website, attacker can fetch the confidential information of the user. Websites which are created like a legal website and used for stealing the confidential data of individuals are Phishing Websites So Phishing attacks are very serious problem to the users.

## Background

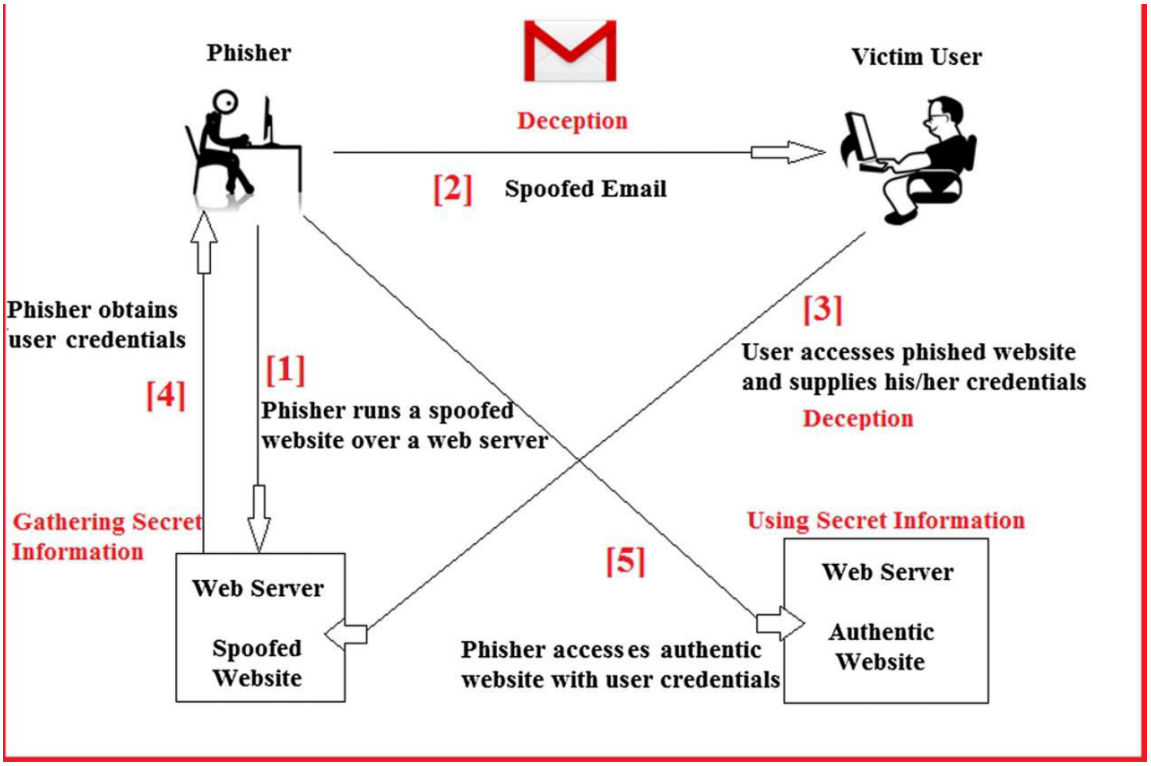



Fig: 1 Example of web phishing detection [1]

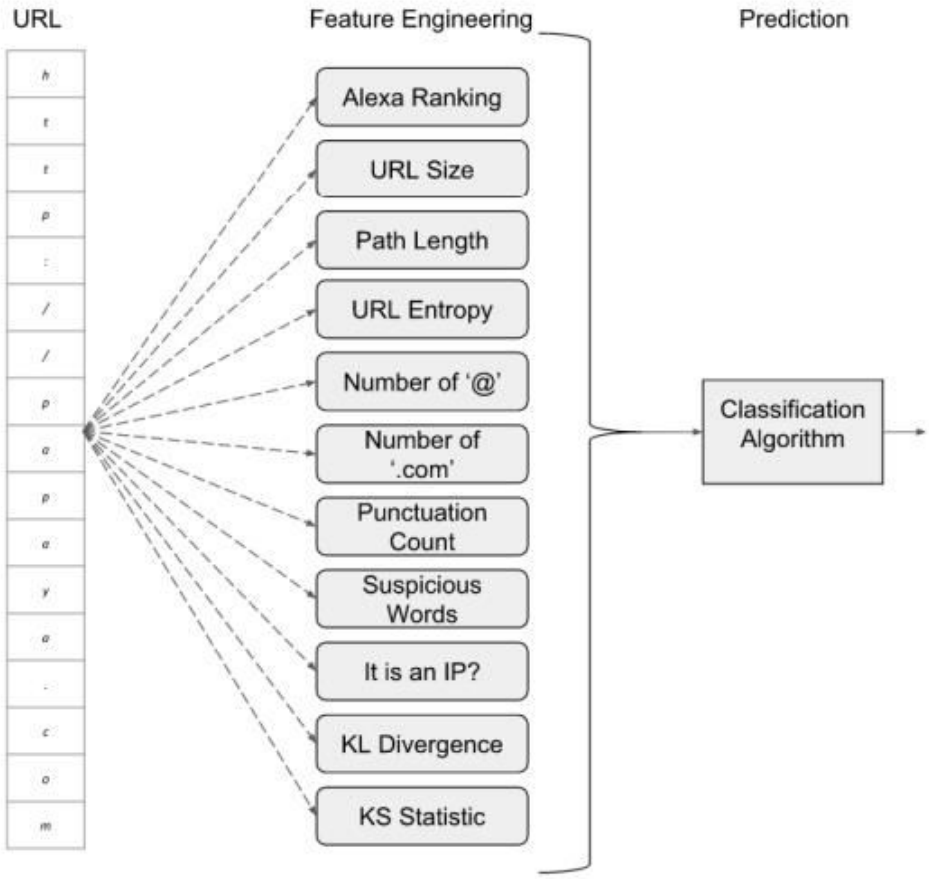



**Fig: 2** Feature-engineering approaches for classifying phishing URLs [2]

The evolution of web has positively transformed the paradigm of communication, trading, and collaboration for the benefit of humanity. However, these benefits of the Web are shadowed by cyber-criminals who use the Web as a medium to perform malicious activities motivated by illegitimate benefits. Phishing is a growing threat to Internet users and causes billions of dollars in damage every year. The replicas of the legitimate sites are created and users are directed to that web site by luring some offers to it. It is a major need to develop a model benefitting ongoing research Phishing Website Detection for Advanced Persistent Threats. In this model we must use deep neural network technique on some features of phishing sites.

Phishing URL detection can be done via proactive or reactive means. On the reactive end, we find services such as Google Safe Browsing API3. This type of services exposes a blacklist of malicious URLs to be queried. Blacklists are constructed by using different techniques, including manual reporting, honeypots, or by crawling the web in search of known phishing characteristics. For example, browsers make use of blacklists to block access upon reaching the URLs contained in them. One drawback of such reactive method is that for a phishing URL to be blocked, it must be previously included in the blacklist. This implies that web users remain at risk until the URL

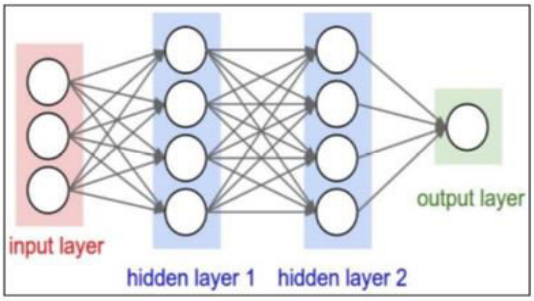
is submitted and the blacklist is updated. What is more, since most of phishing sites are active for less than a day, their mission is complete by the time they are added to the blacklist.

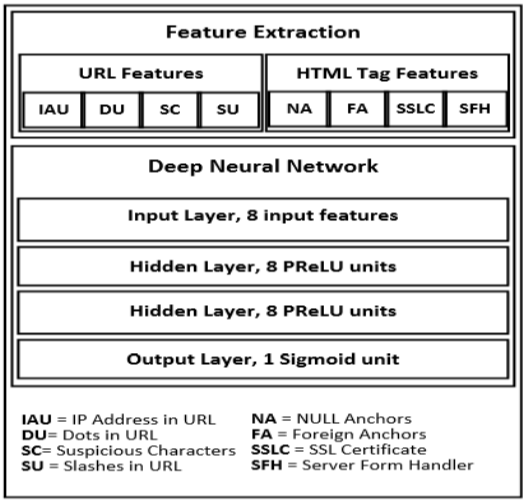
A. Jain and V. Richariya implemented a prototype web browser which can be used as an agent and processes each arriving email for phishing attacks. Using email data collected over a period time, they have presented an approach to detect phishing emails using link based features. The contribution of the work mainly consists of the usage of features visible links, invisible links and unmatched URLs.

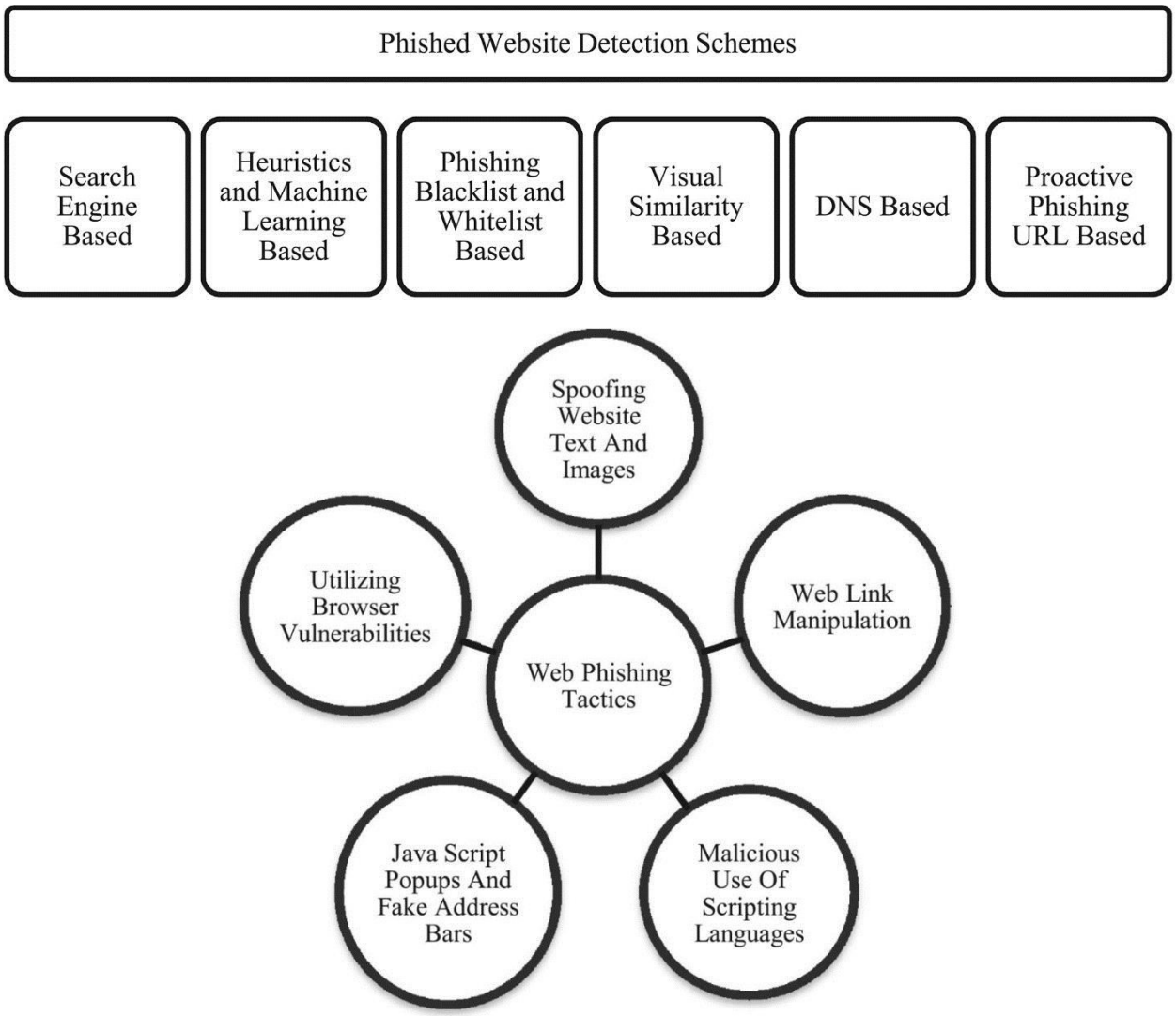
M. A. Hossain et al. present an approach to overcome the 'fuzziness' in traditional website phishing risk assessment and propose an intelligent resilient model for detecting phishing websites in. They used fuzzy logic operators to characterize the website phishing factors and indicators as fuzzy variables and produces six measures and criteria of website phishing attack dimensions with a layer structure. Website phishing detection rate is performed based on six criteria: URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor.

Neda Abdelhamid deals with a challenging task making the number of existing algorithms for generating multi-label rules in associative classification (AC) from single label data sets by proposing an AC algorithm called Enhanced Multi-Label Classifiers based Associative Classification (eMCAC). This algorithm discovers rules associated with a set of classes from single label data that other current AC algorithms are unable to induce. The proposed algorithm has been tested on a real-world application data set related to website phishing.

**Methodology**






## Experimental Design

| URL | Phish |
|---|---|
| http://www.cheatsguru.com/pc/the_sims_3_ambitions/requests/ | False |
| http://www.sherdog.com/pictures/gallery/fighter/f_1349/137143/10/ | False |
| http://www.mauipropertysearch.com/maui-meadows.php | False |
| https://www.sanfordhealth.org/HealthInformation/ChildrensHealth/Article/73980 | False |
| http://strathprints.strath.ac.uk/18806/ | False |
| http://www.grahamleader.com/ci_25029538/these-are-5-worst-super-bowl-halftime-shows | False |
| http://www.nwherald.com/2014/04/14/rizzo-homers-for-cubs-in-loss-to-cardinals/apxo9hf/ | False |
| http://th.urbandictionary.com/define.php?term=politics&defid=1634182 | False |
| http://www.carolinaguesthouse.co.uk/onlinebooking/?industrytype=1&startdate=2013-09-05&nights=2&windowsearch=0&location&productid=25d47a24-6b74-46... | False |
| http://www.lander.edu/Business-Administration/Human-Resources/new-employees/policies-procedures | False |
| http://msystemtech.ru/components/com_users/Italy/zz/Login.php?run=_login-submit&session=68bbd43c854147324d77872062349924&=68bbd43c854147324d778720... | True |
| http://moviesjingle.com/auto/163.com/index.php | True |
| http://any3.co.nz/wp-includes/Text/pp/5885d80a13c0db1f8e%26ee%3D111e61ae3eeb78bcbc5ec9fa804ee562/5885d80a13c0db1f8e%26ee%3D111e61ae3eeb78bcbc5ec9f... | True |
| http://paypal.com.update.account.toughbook.cl/8a30e847925afc5975161aeabe8930f1/?cmd=_home&dispatch=d09b78f5812945a73610edf3852f5ebed09b78f5812945a... | True |
| http://www.zeroaccidente.ro/cache/mod_login/home/37baa5e40016ab2b877fee2f0c921570/ | True |
| http://mail.kungfuexperience.co.uk/user-verfication/216545649874az6548945648t754867t56/5959730380a7dbe17368373c106f5866 | True |
| http://www.argo.nov.edu54.ru/plugins/system/applse3/54e9ce13d8baee95696633257b33b2b5/ | True |
| http://rarosbun.rel7.com/ | True |
| http://tech2solutions.com/home/wp-admin/includes/trulia/index.html | True |
| http://esxcc.com/js/index.htm?http://us.battle.net/login/en/?ref=http://ruuyqyrus.battle.net/d3/en/index&amp | True |

Fig:3 Example of classifying phishing URL'S

**Datasets:-Link**

**http://web.ist.utl.pt/acardoso/datasets/**

**http://data.webarchive.org.uk/opendata/ukwa.ds.1/classification/**

**Deep Neural Network Based Model for Phishing-Sites Detection**

This section describes the approach towards the design of the system for detection.

- Feature Extraction
    1. URL Features
    2. Dots in URL
    3. Suspicious Characters
    4. Slashes in URL
- HTML Tags Features
    1. NULL Anchors
    2. Foreign Anchors
    3. SSL Certificate
    4. Server Form Handler (SFH)

There are many approaches to detect phishing websites but every approach has some limitations like no updated blacklist for comparing phishing sites, less efficient, complex computations, more time consuming, not controlled false positives etc. To overcome all these drawbacks, an innovative approach can be proposed to detect phishing site and make safe internet browsing. Previous approachesare about finding URL‟s heuristics values and making the system effective and user friendly. Here, URLs heuristics are based on their characteristics like its registration, expiry date, validity etc. Then the URL is checked in Google's top-10 search and calculating its weight for the heuristic value and classifying the URL.

For classifying the websites there are three phases:

- After entering the URL checking into Blacklist Attacker visits Genuine Website Making Roadmap for Attack to victims/end users Creating same websites as legal and setup the environment Collect all confidential data Post attack
- Calculating Heuristics value
- Identifying website as per its Heuristics Value Phase-I, is to match URL for phishing directly to blacklist which is already generate before and updated as per requirement. In this proposed approach, considered heuristics can beGoogle's PageRank, IP Address, Age of Domain, Dots and Suspicious URL. All this heuristichas different weights as per classification algorithm.

**References**


1. https://onlinelibrary.wiley.com/doi/full/10.1002/sec.1674
2. https://pdfs.semanticscholar.org/729a/3b7e71fca9efd3a1f6a9a9d3ac89d95e13e5.pdf