**Initial Research**

**Data Sets**

* In the paper, *Machine learning based phishing detection from URLs* by Ozgur Koray Shingoz, they collected a data set of legitimate and phishing URLs from Phish Tank. They used Yandex Search API to collect legitimate URLs from the site. Dataset can be found [here](https://github.com/ebubekirbbr/pdd/blob/master/input/keywords.txt)
* In the paper, *Detecting Phishing Websites Using Machine Learning,* a dataset was built from 12000 phishing URLs from PhishTank, while there are 4000 legit URLs from 10 chosen users. Further reductions of dupicates and missing data turned the size of 6116.
* *PhishStorm: Detecting Phishing With Streaming Analytics:* Used 2 datasets. Created Phishing Dataset using PhishTank, downloading lists on a daily basis from Oct 11th to Nov 10 to obtain dataset of 53089 phishing urls. Legitimate data set, selected URLs from the Open Directory Project (DMOZ), directory of all legitimate websites. Picked out 48009 legit URLs.
* *Phishing Web Sites Features Classification Based on Extreme Learning Machine.* Took 11000 Urls from UC Irvine Machine Learning Repository Database
* *Phishing Emails Detection Using CS-SVM:* In this email phishing paper, used 2 datasets phishing emails and non-phishing emails. Phishing emails made up of 2 Corpora, 1203 old phishing emails by Jose Nazairo in 2005 and 181 up to date phishing emails. Non-phishing set made up of 20071 emails from public Enron email set by CALO project
* *PhishMon: A Machine Learning Framework for Detecting Phishing Webpages*: Phishing Dataset made by live monitoring PhishTank for new urls over the course of September 2017 to November 2017, retrieving 4,807 phishing URLs. For legit URLs, randomly selected 20000 urls from Alexa top one million domain name list, reducing it to 17508
* *An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features:* 3983 phishing sites and 4021 legitimate sites
* *Detecting Phishing Websites and Targets Based on URLs and Webpage Links:* Phishing data set of 2892 URLs obtained from PhishTank from January 8, 2019 to January 9, 2018. 3305 legit urls obtains from the top 1 million Alexa rankings and network security challenge. Dataset can be found [here](https://github.com/huapingz/deepforest)
* *Wide Scope and Fast Websites Phishing Detection Using URLs Lexical Features:* Phishing Dataset collected from Phishtank and OpenPhish. Collected 465461 URLs from PhishTank and 4627 from OpenPhish. Legit Dataset of URLs collected from dmoz.org and webcrawler.com. 10275 from dmoz, 10275 from webcrawler
* *Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks:* The datasets took sent to people at a tier-1 research institution. For their phishing emails, they had a dataset of 388, 264 emails that were caught by Sophos, an email filtering system, and then 37606 emails that were uncaught. And then there is a dataset of 158444 legit emails

**Features**

Lookup based features:

* DNS Lookup
* Whois database
* Age of Domain
* Page Rank (Value between 0 – 1, used in SEO, rank of Phishing sites is 0)
* Search Result
* Alexa Rank

URL Based Features:

* Protocol (https://)
* Subdomain (www)
* Second Level Domain (SLD) (xyz-company)
* Top Level Domain (TLD) (.com)
* Path (…/info/…)
* File name (index.html)
* IP Address

Word Based Features (NLP)

* Brand names
* Keywords
* Random words
* Suspicious Words (need to build a corpus)
* Word Lengths (Longest, Shortest, Average)
* Word Counts (adjacent, keyword, brand name, subdomains, random word counts, ect.)
* Typo-Squated Words – words that people might accidentally mispell (Google → Goggle)
* Special Characters
* Check URL on authority figures
* Puny Coded (More research on this needed)
* Term Frequency-Inverse Document Frequency (TF-IDF) (commonly used to fetch keywords)
* Length of URL (According to *Comparative Analysis of Features Based Machine*

*Learning Approaches for Phishing Detection,* 30.3% of phishing urls contain 80+ characters)

Webpage based features

* Hyperlinks
* Server Form Handler
* Website Traffic
* Amount of Form-based HTML elements (input, submits, form elements, ect.)

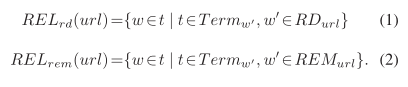
**PhishStorm: Detecting Phishing With Streaming Analytics:**

Split incoming URLs into two sets (using http://sezopoztos.com/paypalitlogin/us/webscr.html?cmd=\_login-run as an example:

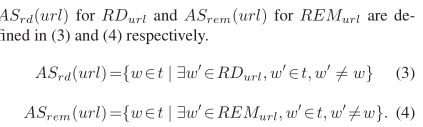
* Registered Domains RDURL = { sezopoztos,sezopoztos.com } (Typically contains everything before the TLLD)
* Remaining part (of a URL) REMURL = { paypal, it, login, us, web, src, html ,cmd, login,run }

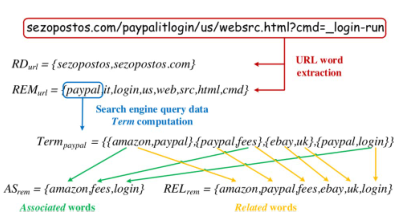
Also performed evaluation of URL using Search Engine Query Data to find if any keywords in URL is related to specific brands or can be mistaken by them. Used Google Trends and Yahoo Clues to retreieve related terms. Used these in conjunction to previous set to create new set of words

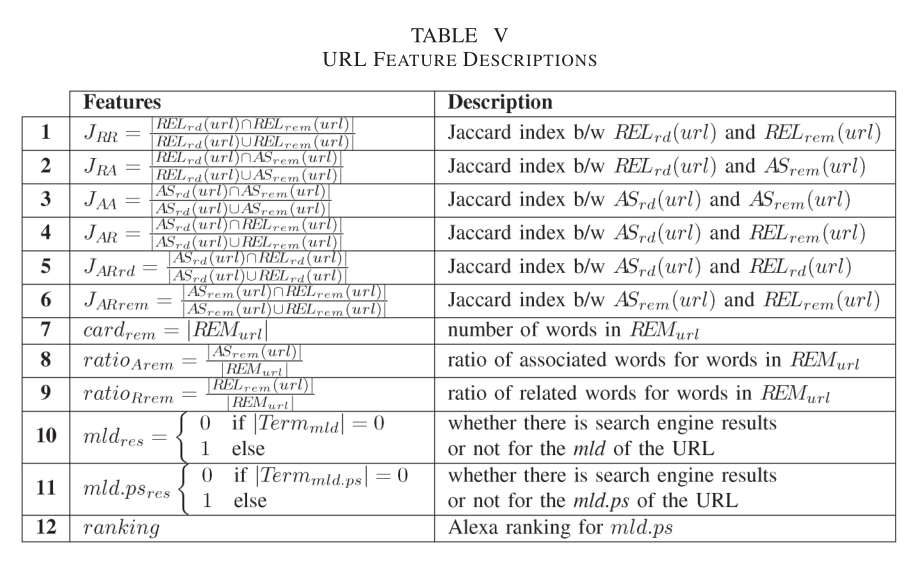
* Related words of a set (found through Google Trends/Yahoo Clues results) RELSet(URL)
  + (1) Related words in the set of registered domain
  + (2) Related words in set of remain words of URL



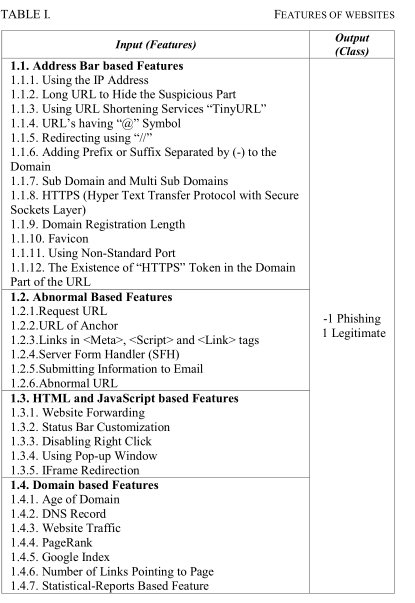
* Associated Words of a set ASSet(URL) (words that commonly appear with another)







**Phishing Web Sites Features Classification Based on Extreme Learning Machine Feature List:**



**Phishing Emails Detection Using CS-SVM:**

To broaden my horizons, I wanted to read about phishing emails and the potential features extracted from emails. URLs still primary concern, but can get more from email

* Header based Features
  + Sending time
  + Presence of Dark copy
  + # of Carbon Copies (CC)
  + Reply Email
  + Presence of terms in title (order, payment, RE-)
* Body based Features
  + Disparities between href attribute and LINK text
  + Presence of sensitive words in the LINK text
  + Presence of javascript
  + Specific Verbs
  + Use of “You” and “Your”
* URL Based Features
  + Presence of IP Adress
  + Number of URLs
  + Number of dots in domain name
  + Presence of specific characters (“@”, “-“)
  + Active Duration of domain names
  + Length of URL
  + DNS records
  + Alexa ranking
  + Registration Duration
  + Multiple protocols in domain
  + Domain similarity measures (Typosquating)
  + Url shortening
  + Disparities between DNS in URLs and DNS of sender

**PhishMon: A Machine Learning Framework for Detecting Phishing Webpages**

* HTTP Features
  + Header Field Names
  + # of header fields
  + # of Non standard header fields
* Code Complexity Features
  + Minified JS Code
  + # of External Script Blocks
  + # of Ineline Script blocks
  + # of DOM event handlers
  + JS library count
  + Landing Page Variants
  + Lines of Code (LOC)
  + URL Redirect
* Certificate Features
  + X.509 Certificate
  + Passing Browser Validation
  + Valid on multiple 2-level domain names
  + Extended Validation (EV) Certificate
  + Name of Issuer
  + Certificate Longevity Period
  + Certificate Age

**An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features:**

Length related

* Length of domain
* Length of file name part
* Path URL ration
* (the length of path divided the length of URL), and the length of longest token in the path part of URL
* Average length of words
* Longest word
* Shortest
* Standard division of word length
* Length of url

Counts

* Domain tokens
* Raw words in URL after parsing out special characters
* # of digits in the query part,
* Number of dots ‘.’
* # of digits in domain, subdomain and path
* # of special characters

Hyperlink/HTML info

* Total hyperlink features (typically, phishing sites have limited webpages or use hyperlink hidden techniques)
* # of internal and external hyperlinks
* Null hyperlinks
* Suspicious CSS
* Internal and External redirection hyperlinks
* Error based features
* Authenticity of login forms
* Existence of Favicon

Suspicious URL

* Brand check for the domain
* # of brand names
* Presence of ‘www’
* Presence of ‘com’
* Embedded domains (Putting a legit domain like [www.apple.com](http://www.apple.com) in the path)
* Is IP Address
* Check for characters ‘@’ and ‘-‘. (URL to the right of an @ symbol is the actual url, and the left is disregarded)
* Sensitive word check
* Top Level Domain check.

**Detecting Phishing Websites and Targets Based on URLs and Webpage Links:**

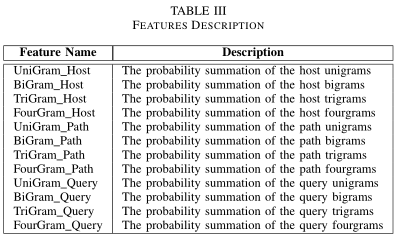
Statistical Features

* IP Address
* Suspicious Characters (‘@’, ‘&’, ‘-’, and ‘\_’, etc.)
* Network Protocol (HTTP vs HTTPS)
* Alexa Ranking
* Length of Entire URL/Host Name/Main Domain/
* # of dots in host name/URL path
* URL Token count
* Host name token count
* Search results

Lexical Features

* Split URL into 8 sets of keywords (host names, path, ect.)
* Calculate TF-IDF of keywords extracted from webpage content
* Do pairwise intersections on the 8 keyword sets, to obtain a 28-dimensional vector and treat it as a lexical feature

**Wide Scope and Fast Websites Phishing Detection Using URLs Lexical Features:**

* Split URL into parts
  + Hostname
  + Path
  + Query
* Here, they took the approach of building a Language Model (LM) from the collection of keywords in the dataset. The main function is to assign possibilities to a set of strings. The set of words are built out using N-grams 

**Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks:**

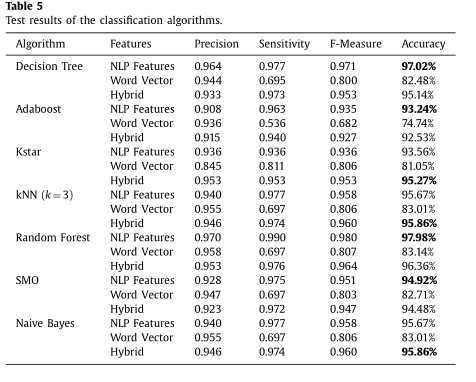
Another study based on Email phishing, focusing more on NLP

* Commonly Known phishing words + synonyms
  + Phishing emails typically use deceptive, attention grabbing words to force users into bad actions
  + Ex: “urgent”
* Words associated with the research institution
  + Name of city where institution located, name of departments, ect.
* Commonly occurring words from phishing corpus + synonyms
  + Built up through frequency analysis of phishing emails
* Proper nouns (Organization names)
  + Use trusted organization to hide nature of emails
* Structural Email Features
  + Embedding links
  + # of html tags
  + Use of images
  + URLLs
  + Encoding

**Classifiers**

**Classifiers used in *Machine learning based phishing detection from URLs*:** Naïve Bayes, Random Forest, kNN (n = 3), Adaboost, K-star, SMO, Decision Tree

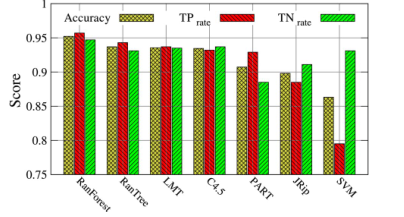
**Results in *Machine learning based phishing detection from URLs***



**Classifiers used in *Boosting the Phishing Detection Performance by Semantic Analysis*:** AdaBoost, Bagging Algorithm, Random Forest, SMO

**Classifiers in *Detecting Phishing Websites Using Machine Learning:*** Random Forest using 36 features.

**Classifiers in *PhishStorm: Detecting Phishing With Streaming Analytics:*** Random Tree, Random Forest, C4.5, LMT, PART, JRIP, SVM

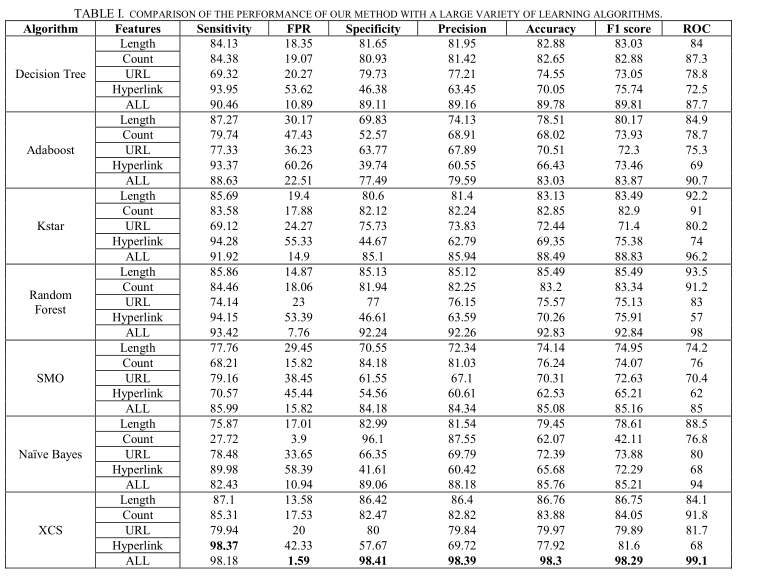


**Classifiers in Phishing Web Sites Features Classification Based on Extreme Learning Machine:** Extreme Learning Machine

**Classifiers in Phishing Emails Detection Using CS-SVM:** Used SVM variant called CS-SVM (Cuckoo Search SVM). It uses an algo called Cuckoo Search to optimize Parameter selection.

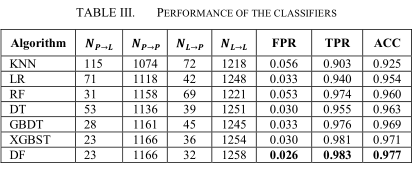
**Classifiers in PhishMon: A Machine Learning Framework for Detecting Phishing Webpages:** CART, KNN, AdaBoost, Random Forest

**Classifiers in An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features:** Decision Tree, Adaboost, Kstar, Random Forest, SMO, Naïve Bayes, XCS.

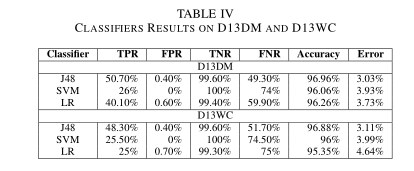


The main classifier analyzed in this paper was the XCS, a rule based learning system that evolves it rule set, adapting it’s ruleset depending on changes to it’s environment.

**Classifiers in Detecting Phishing Websites and Targets Based on URLs and Webpage Links:** KNN, Logistic Regression, Random Forest, Decision Tree, Gradient Boosting Decision Tree, XGBoost, Deep Forest



**Classifiers in Wide Scope and Fast Websites Phishing Detection Using URLs Lexical Features:** J48, Support Vector Machine and Logistic Regression



**Classifiers in Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks:** A Boosting-Based Classification was used called Random Under-Sampling Boost Algorithm. With the boosting algo, used Gaussian Naïve Bayes, decision trees, ect.

**Interesting Notes:**

*PhishStorm: Detecting Phishing With Streaming Analytics* suggested different ways of obfuscating a phishing URL

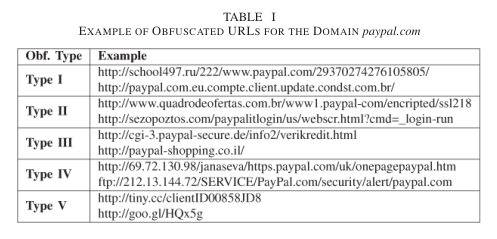
• Type I: URL obfuscation with other domain: The mld.ps is a real domain name, usually registered by the phisher, while the original domain being phished is part of the path, the query or the upper level domain.

• Type II: URL obfuscation with keywords: Again the mld.ps is a real domain name, and the brand being phished and related words are part of the path, the query or upper level domain.

• Type III: Typosquatting domains or long domains: the mld.ps of the URL is the domain being phished but misspelled, with letters or words missing or added, or the domain is pronounced the same way as the original but written differently. The targeted brand can also be combined with other words to create an unregistered domain.

• Type IV: URL obfuscation with IP address: the URL’s hostname is replaced by an IP address and the brand being phished is part of the path or the query.

• Type V: Obfuscation with URL shortener: A URL shortening service is used to hide the name of the real host. Such URLs are not meaningful and are mainly used in phishing attacks targeting services that use this kind of short URL, like Twitter.



**Papers Read (that I found pertinent information)**

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. doi: 10.1016/j.eswa.2018.09.029

Comparative Analysis of Features Based Machine Learning Approaches for Phishing Detection

PhishStorm: Detecting Phishing With Streaming Analytics

Phishing Web Sites Features Classification Based on Extreme Learning Machine

Phishing Emails Detection Using CS-SVM

PhishMon: A Machine Learning Framework for Detecting Phishing Webpages

An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features

Detecting Phishing Websites and Targets Based on URLs and Webpage Links

Wide Scope and Fast Websites Phishing Detection Using URLs Lexical Features

Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks