A Course on Social Engineering
Phishing & URL Analysis

Phishing Detection through URLs

Akbar Namin
Texas Tech University
Spring 2021

Computer Attacks

- Classified into three types [9]:
  - Physical attacks: Attacks against physical infrastructure (e.g., devices)
  - Syntactic attacks: Target the operating logic of computers and networks (e.g., software vulnerabilities)
  - Semantic attacks: Target people's vulnerabilities (e.g., interpretation of computer messages)
- Phishing is of type of semantic attacks.
  - People tend to believe information they read, without validating the credibility of the content they received

Phishing Attacks

- The most common attacks in cyber security
- Targeting the weakest element in cyber security chain
- A user is tricked into revealing sensitive and private information

Phishing Attacks

- Phishing motives:
  - Financial gain: Stealing credentials for financial gain
  - Identity hiding: Selling stolen identities to others (potentially criminals) to hide their real identities and conduct illegal activity.
  - Fame and notoriety: Demonstrating the hacking capabilities and peer recognition
  - Espionage
  - Stealing technologies
  - Governmental backed-up

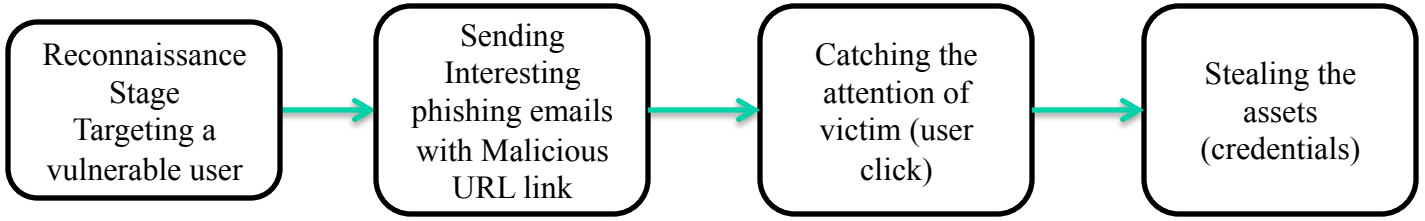
Phishing Attacks – Various Types

- Various types of phishing attacks [9]:
  - Website
    - Used to steal user's credential data
    - Begins by creating website that imitates the appearance of a legitimate Website to deceive users
  - Email
    - Use spam, fake websites to trick users into revealing personal information.
  - Web Trojans
    - The attacker collects the user's credentials locally and transfer them to the main phishers.
  - Content-injection phishing
    - Attacker replaces part of a legitimate site with fake content to deceive users so user give up their personal information to attacker.

Phishing Attacks

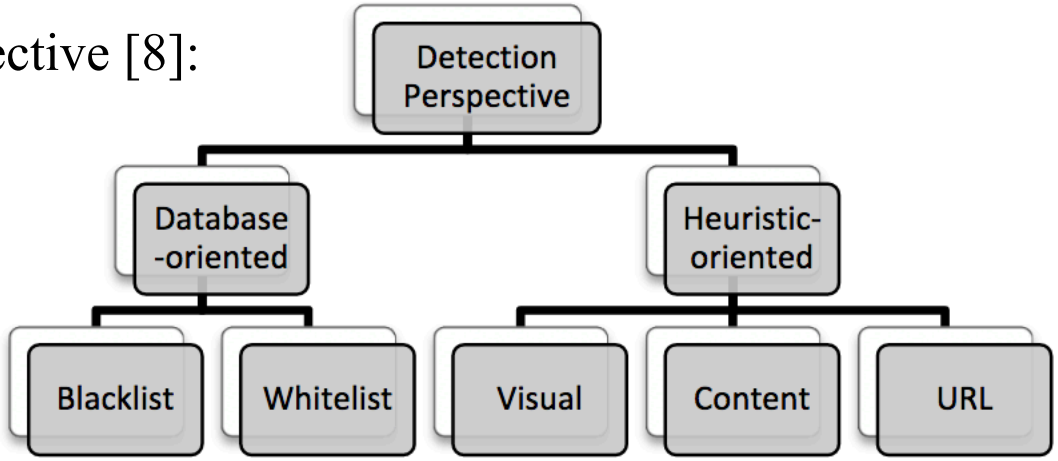
1. A fake website is created by the attacker
2. The website mimics the original Websites (e.g., PayPal's Website)
3. Similar appearance and visual similarities to the genuine website
4. A spoofed link embedded in a phishing email is sent to the users by the attacker
5. The goal is to lure the victim to click on the spoofed link and redirecting to a fake web page
6. The user's vulnerabilities is exploited

Phishing Attacks

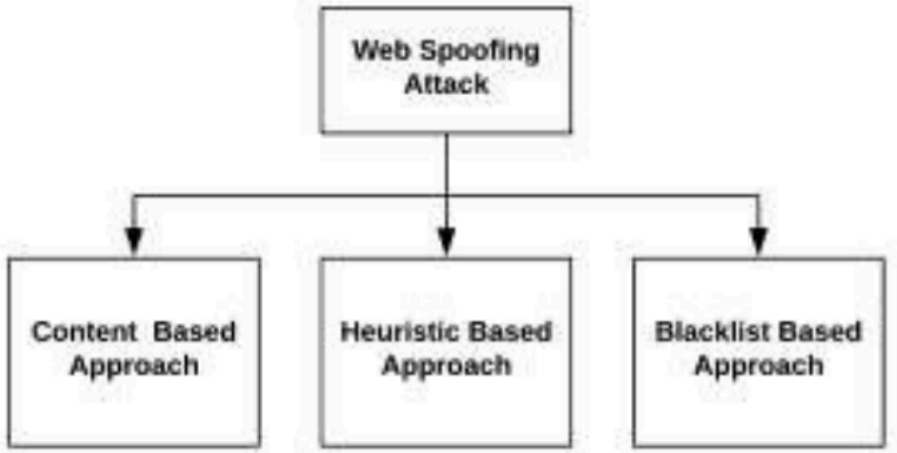


Phishing Attacks - Prevention

- Phishing detection perspective [8]:



- Phishing of Website detection [9]:



Phishing Attacks – Different Forms

- Different forms of phishing attacks [9]:
  - Fake URLs
  - Misspelled URLs
  - Anchor text
    - The links within the webpage points to a domain different from the domain specified in the URL address bar
  - Fake SSL lock
    - It is easy now for the attacker to obtain SSL certificates for their malicious sites and fool users
  - URL manipulating using java script
    - The attacker inserts a string to be used in the webpage and treated by the user‟s browser as an executable code. When the browser loads the page, the malicious script executes without the user consents

Phishing Attacks - Prevention

- Blacklisting through features offered by Web browsers
  - Pros
    - Easy to implement
  - Cons
    - Ineffective for detecting newly setup phishing Websites (unable to detect zero-day phishing Websites)
- Tool supports
  - Google Safe Browsing
    - Can be integrated with several mainstream browsers such as Chrome, Safari, Firefox
  - Microsoft's SmartScreen
    - For IE browser

Phishing Attacks - Prevention

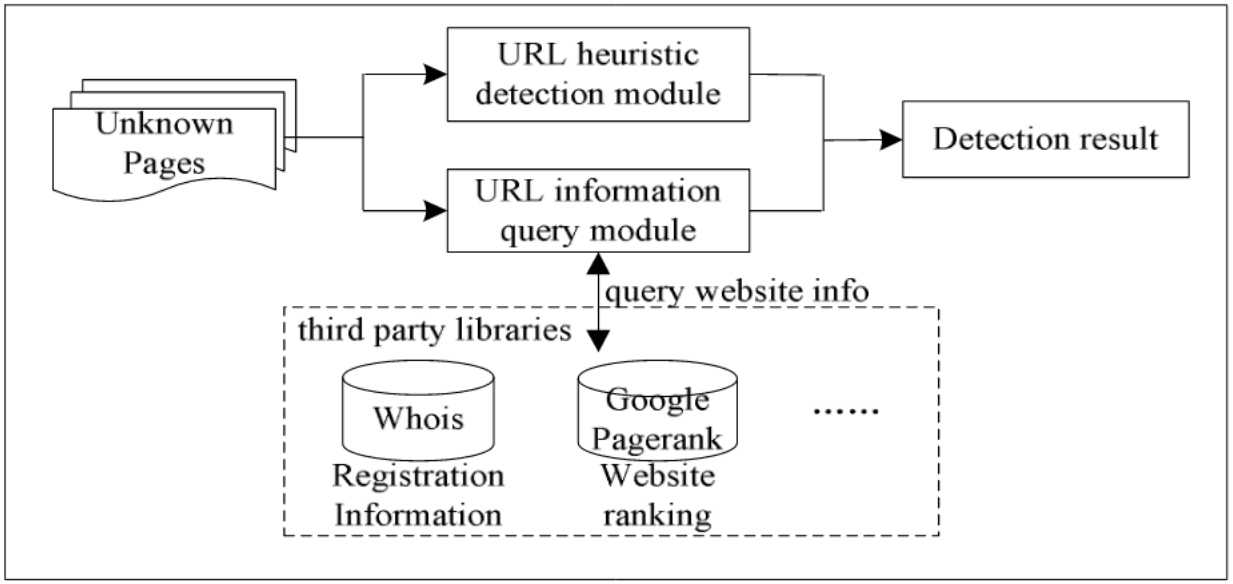
- Visual-Similarity-Based Approach [7]
  - Block-level similarity
    - The weighted average of the visual similarities of all matched block pairs between two pages.
    - Text blocks: colors, border style and alignment, etc.,
    - Image blocks: alternative text, dominant color, and image size, etc.
  - Layout similarity
    - The ratio of the weighted number of matched blocks to the total number of blocks in the true webpage.
    - Two blocks are considered matched if they both exhibit high visual similarity and satisfy the same constraints with corresponding matched blocks.
    - Layout similarity of two webpages: the ratio of the weighted number of matched blocks to the total number of blocks in the true webpage,
  - Overall-style similarity
    - Style consistency observed all over the pages of an online banking/trading system (color, fonts, etc).

Phishing Attacks - Detection

- Heuristics-based approaches
- Utilize the exploits common patterns, found in previously reported phishing attacks, to detect new ones
- Examples of patterns:
  - If the URL contains any IP address
  - If the domain name does not contain any official and correct organization name
  - The occurrence of any hyphen in the primary domain
  - The existence of a password fields in Web pages.
- Cons:
  - The proper formulation and weighing of heuristics can result in misclassification

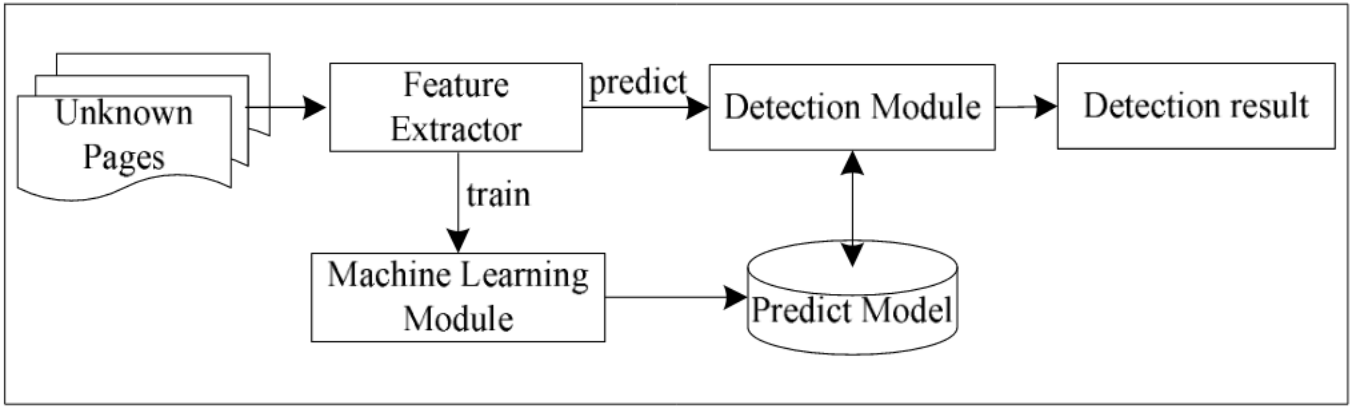
Phishing Attacks - Detection

- Heuristics-based approaches
- Heuristic Detection Model Based on URL and DNS features [11]



Phishing Attacks - Detection

- Machine Learning-based approaches
- Improvements over heuristic-based approaches
  - Tuning weights of heuristics automatically using training examples
    - The training examples may be URLs or contents of Web pages, or both
- Cons of using content of Web pages
  - The page should be loaded causing malicious actions already to be taken place
- The general formulation:
  - A binary classification problem in which URLs are classified into phishing and legitimate
- Content-based Machine Learning Detection Model [11]:



Phishing Attacks - Detection

- URL-based phishing attacks: embedding sensitive words/characters in link:
  - Mimic similar but misspelling words.
  - Contain special characters for redirecting.
  - Use shortened URLs.
  - Use sensitive keywords which seem reliable.
  - Add a malicious file in the link
- URL analysis-based approaches
  - Static analysis
  - Pattern recognition
  - Machine learning classification
  - Bag-of-words
- Benefit: Can detect if a URL link is a phishing attack even before clicking and visiting the link (on-the-fly detection)

Phishing Attacks - Detection

- A URL consists of three major parts:
  - Protocol (e.g., https://)
  - Hostname (e.g., secure.bankofamerica.com) : the server's name the resource is hosted on
  - Path (e.g., /login/sign-in/signOnV2Screen.go?msg=InvalidCredentials_2_Remaining…)
    - Consists of: directory, file name and arguments

https://secure.bankofamerica.com/login/sign-in/signOnV2Screen.go?msg=InvalidCredentials_2_Remaining&request_locale=en_us&lpOlbResetErrorCounter=0

Phishing Attacks - Detection

- URL features
  - Lexical: features that are extractable quickly from the URL string
    - E.g., the length of the URL, the number of dots in the URL
  - External: features that extractable through direct queries to remote servers
    - E.g. whois lookup, DNS resolution

Phishing URL Types

- Obfuscating types
  - Obfuscating the Host with an IP address
    - Hostname is replaced with an IP address, and the organization being phished is placed in the path
    - http://173.193.212.4/secure.bankofamerica/signin
  - Obfuscating the Host with another Domain
    - Host contains a valid looking domain name, and the path contains the organization being phished
    - http://123letmegetyou.ru/secure.bankofamerica/signin
  - Obfuscating with large host names
    - The organization being phished is in the host but appends a large string of words and domains after the host name
    - http://bankofamerica.com.customerservice.info//secure.bankofamerica/signin
  - Domain unknown or misspelled
    - No apparent relationship to the organization being phished or the domain name is misspelled
    - http://www.paypaI.com : the use of “I” instead of “l”

Phishing URL Types

- Obfuscating types
  - URL Shorteners
    - bit.ly, x.co, goo.gl, tiny.cc
    - Suitable for twitter with limited characters
    - http://bit.ly/2wbw48P instead of https://f5.com/labs/articles/threat-intelligence/cyber-security/russian-hackers-face-to-face
  - URL Doppelgangers
    - Example: the use of https://www.paypaI.com instead of https://www.paypal.com
    - The last character is replaced with “I” instead of “l”
  - URL Redirects
    - Hijacking redirecting mechanism to a phishing website
    - Example: instead of redirecting to http://courses.com/redirect.php?url=http://coolcoursesite.com to http://courses.com/redirect.php?url=http:////bitly.com/98K8eH

Phishing URL – Feature Analysis

- Web page-based:
  - Page rank: a numerical metric between [0, 1]
    - Representing the relative importance of the page: higher more important the page
    - Phishing Websites (pages) have low ranks
  - Short life
    - Phishing pages have short lives and thus are not indexed properly
  - Measuring the quality of Websites (pages)
    - Phishing emails have low scores for quality
    - Google provides recommendation for quality measuring
    - Google Webmasters: http://www.google.com/support/webmasters/bin/answer.py?answer=35769

Phishing URL – Feature Analysis

- Domain-based:
  - Only one feature: whether or not the URL's domain is listed as a Whitelist domains or not.

Phishing URL – Feature Analysis

- Typed-based:
  - Obfuscating the Host with an IP address
    - None of the whitelisted websites have their host names obfuscated with IP address
  - Obfuscating the Host with another Domain
    - Check if any of the organizations are found in the URL's path but not in its host.
  - Obfuscating with large host names
    - Determine the number of characters present after an organization in the hostname.
    - The maximum number of characters in a white list URL are around 14 characters

Phishing URL – Feature Analysis

- Word-based:
  - Phishing URLs often contain several suggestive word tokens
  - Examples of words: **webscr**, **secure**, **banking**, **ebayisapi**, **account**, **confirm**, **login** and **signin**.
    - Source: A Framework for Detection and Measurement of Phishing Attacks, G. S., Provos, N., Chew, M., Rubin, A.D.

Feature	Distribution of Feature Presence	
	White List (%)	Black List (%)
confirm	0.23	4.25
account	1.5	4.9
banking	0.87	7.95
secure	0.16	9.88
ebayisapi	1.5	13.9
webscr	0.32	14.2
login	2.61	21.53
signin	0.95	23.29

Phishing URL – Feature Analysis

- Common URL-based features [6]:

No	Feature Name	Description
1	IP address	Check if IP address is presented in existing domains
2	Avg. words length	Count average length of meaningful words in entire domain name
3	exe/zip	Check if exe/zip is present in URL
4	No of dots	Count # of dots in URL
5	Special symbols	Count special symbols in URL
6	URL length	Count # of characters in URL
7	Top-level domain (TLD) feature	Validate TLD-based features [39][40][44]
8	"http" count	Count # of "http" in URL
9	Brand name	Extract brand name in URL domain
10	"//" redirection	Check if "//" is included in URL path
11	Domain separated by "-"	Check if "-" is included in domain name
12	Multi-sub domain	Check how many # of multi-subdomains are included in URL
13	Suspicious words	Check if suspicious words are included in URL
14	Digits in domain	# of digits in domain
15	Character entropy	Calculate character distribution in entire

Phishing URL – Some Stats

- Distribution of Phishing by Organization

Organization	Unique Phishing URL	Potential Success Rate (%)
Ebay	231	14.8
Paypal	211	7.6
Fifth Third Bank	61	0
Unknown	60	8.2
Bank Of Scotland	32	8.6
Volksbank	29	0
Wells Fargo	29	0
Bank of America	28	2
Sparkasse	14	3
Private Banking	13	0
HSBC	7	0
Chase	5	3
Amazon	4	4
Banamex	4	0
Barclays	4	0

Phishing URL – Some Stats

- Geographical Distributions of Hosting Phishing Websites

Country	Phishing URLs(%)
United States	70.3
Sao Tome and Principe	6
Belize	4.5
China	2.9
Germany	1.3
Taiwan	1.2
United Kingdom	1
Russian Federation	1
Romania	0.9

Phishing URL – Feature Analysis

- Word-based:
  - URL
  - Page Based, Domain Based, Type Based and Word Based features.

Phishing URL – Detection Techniques

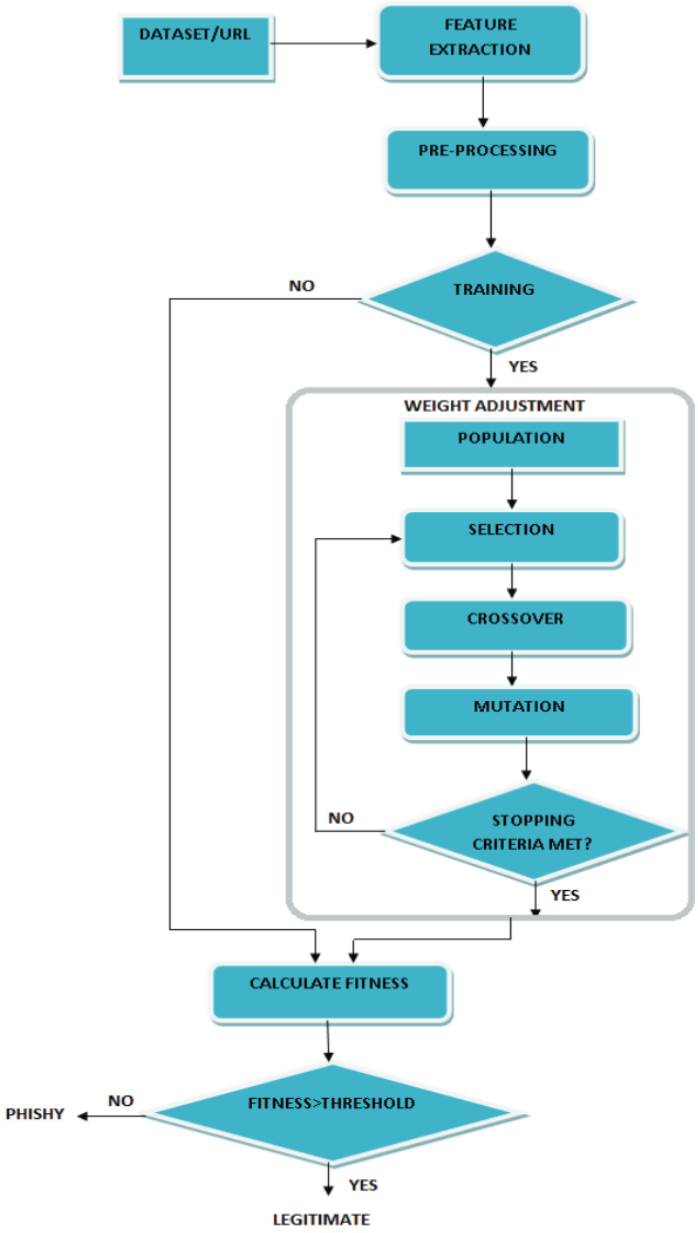
- Website Phishing Detection Techniques [9] :
  - Genetic algorithm based approaches
  - Based on associative classification data mining
  - Intelligent detection and categorization model
  - Heuristics anti-phishing detection-based approaches
  - Neuro-Fuzzy algorithm
  - Based on machine learning classifiers

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques:
  - Four phases of Genetic algorithm based approaches [10]:
    1. Phase 1: Feature Extraction
    2. Phase 2: Pre-processing:
       - The label of each feature is classified into phishing, legitimate or suspicious class.
    3. Phase 3: Weight adjustment:
       - Find the best weights that can classify the website accurately
       - Genetic algorithm is used for weight adjustment.
    4. Phase 4: Results:
       - The best weights derived in third phase are used to calculate the fitness of the URLs in data set and classified into legitimate or phishy by comparing the fitness with the threshold value.

Phishing URL – Detection Techniques

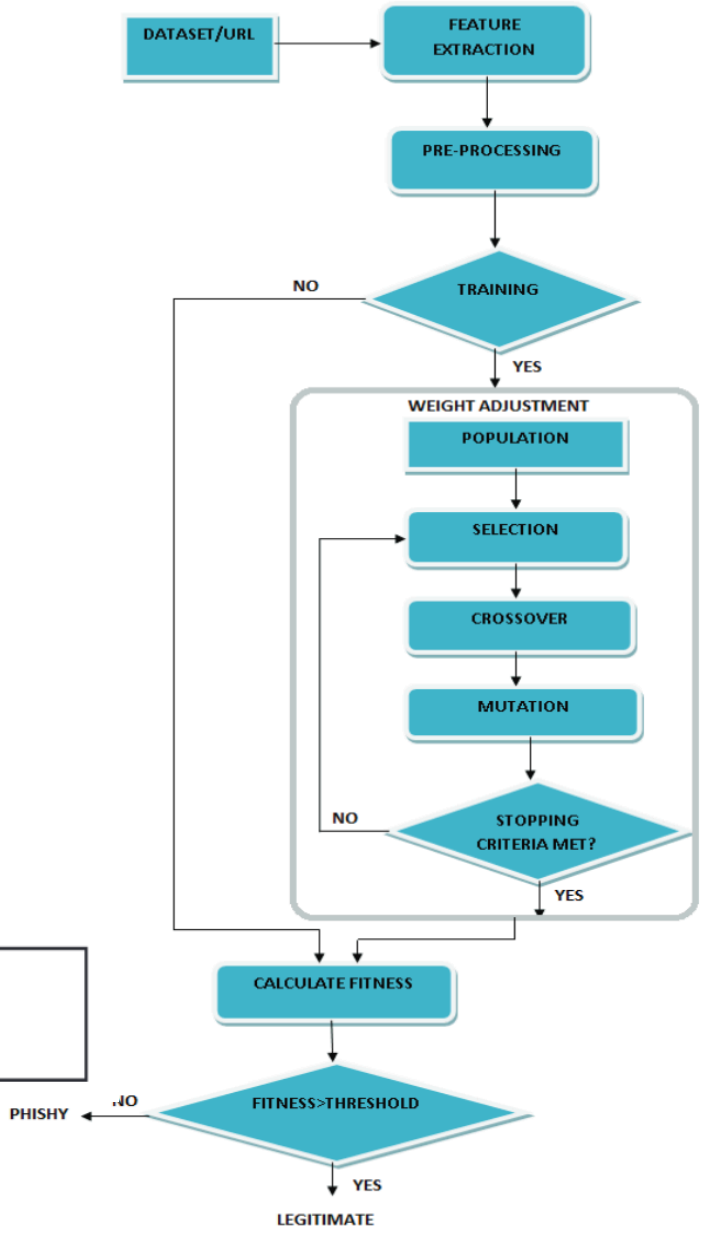
- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I - Feature Extraction:
    - 10 features are extracted from each URL.
    - The features are divided into two genes: Gene1 and Gene2.
    - Features in Gene1: Page Rank, Alexa Rank, Age, DNS Records and Abnormal URL.
    - Features in Gene2: Long URL, Prefix Suffix, Sub Domains, Http/Https and IP Address



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Page rank: The value of a webpage depends on the page rank value and the number of links going out of the webpages, pointing to it.
    - phishing webpages generally have no webpages pointing to it so the page rank value of the phishing webpage is N/A i.e. not available.
    - For legitimate webpages the page rank value lies between 0 and 10. So this feature can be very helpful for detecting phishing websites.
    - Rule for page rank value is :

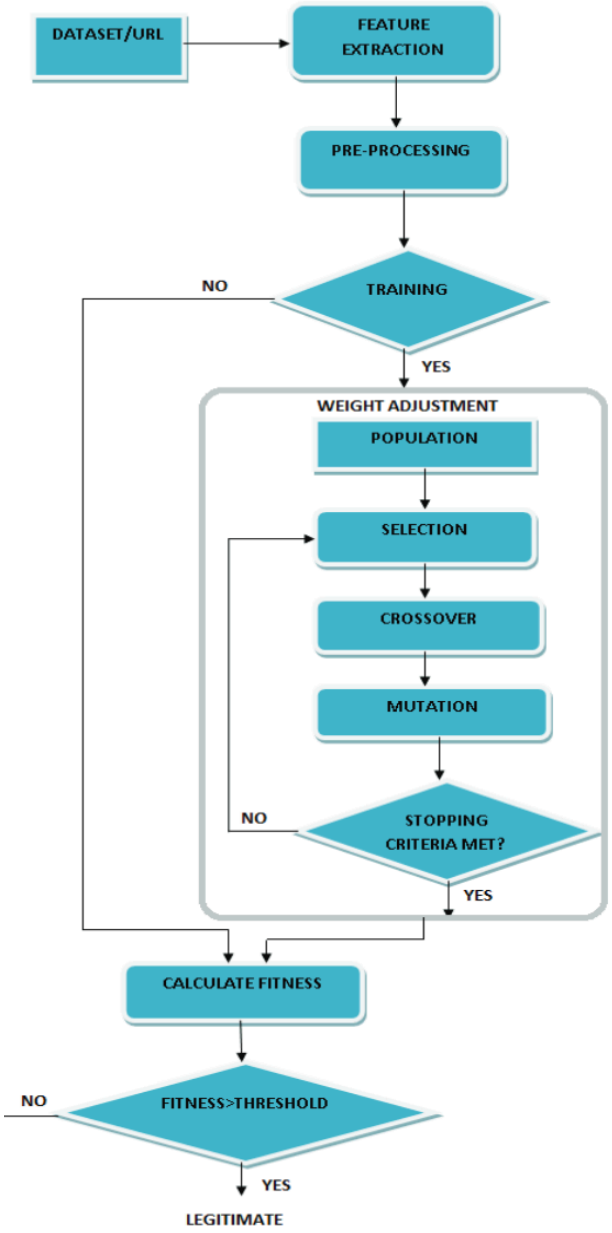
Rule: If page rank is N/A → Phishy
else → Legitimate



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Alexa rank: assigned to a website on the basis of website traffic.
    - measures the popularity of the website by determining the number of visitors and the number of pages they visit.
    - The life time of phishing websites is generally very short, so they may not be recognized by Alexa database.
    - Legitimate websites are mostly ranked among the top 150000
    - rule for Alexa rank is :

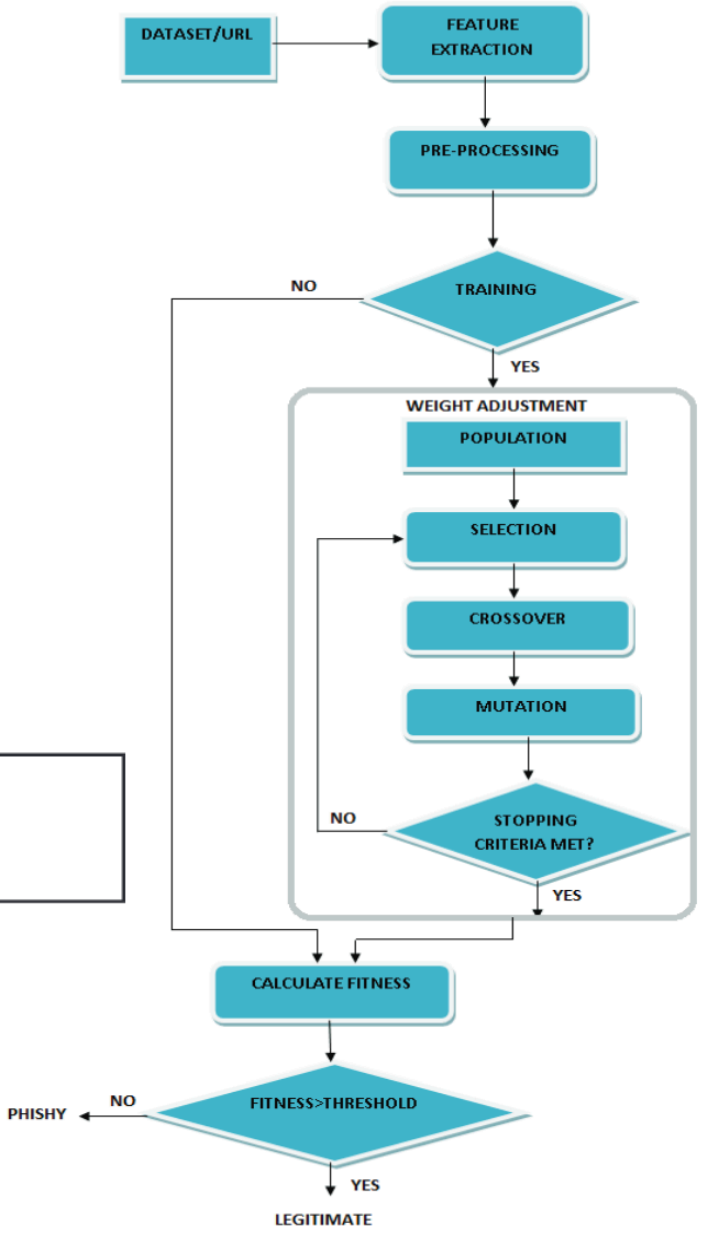
Rule: If alexa rank <=150000 → Legitimate
else if alexa rank >150000 → Suspicious
else → Phishy



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Age of Domain: The domains on which the phishing websites are hosted generally have age less than 6 months.
    - This feature can be extracted from whois database
    - The rule for age is :

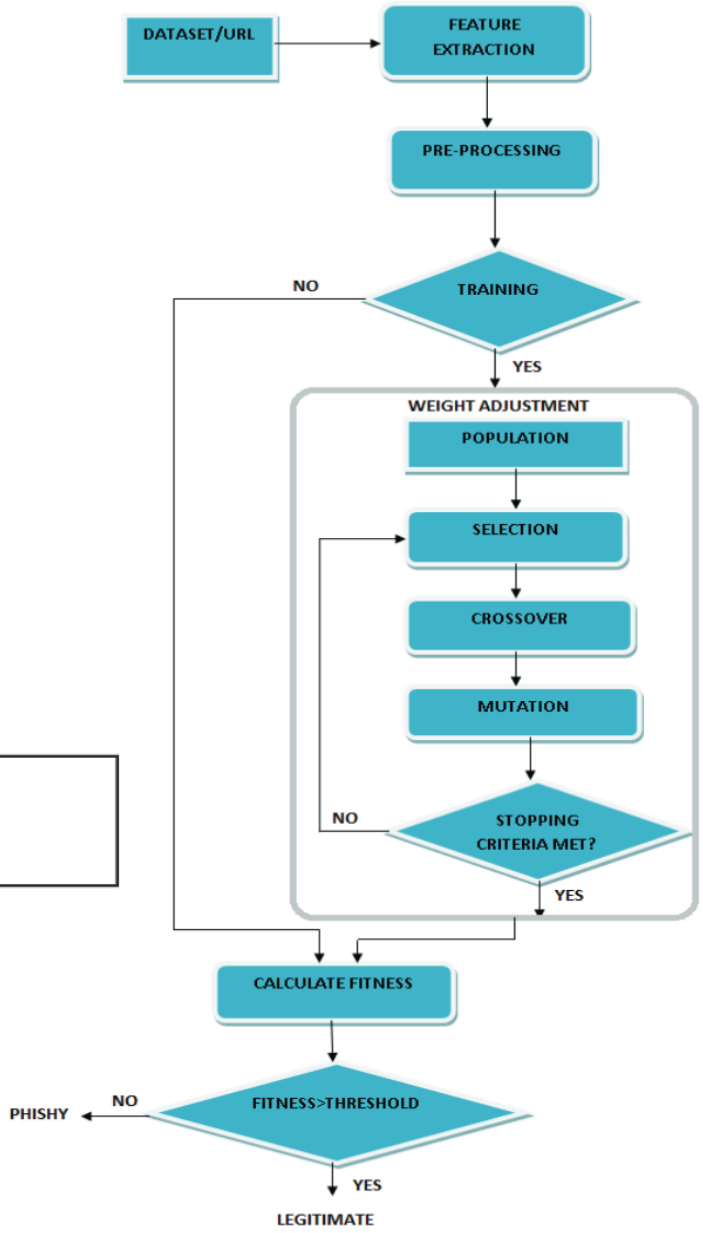
Rule: age ≤ 6 months → Phishy
else → Legitimate



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - DNS Record: For Phishing Website DNS Records are generally not found.
    - If the DNS Records exists then it is classified as legitimate otherwise it is classified as phishy.
    - Rule for DNS Records is :

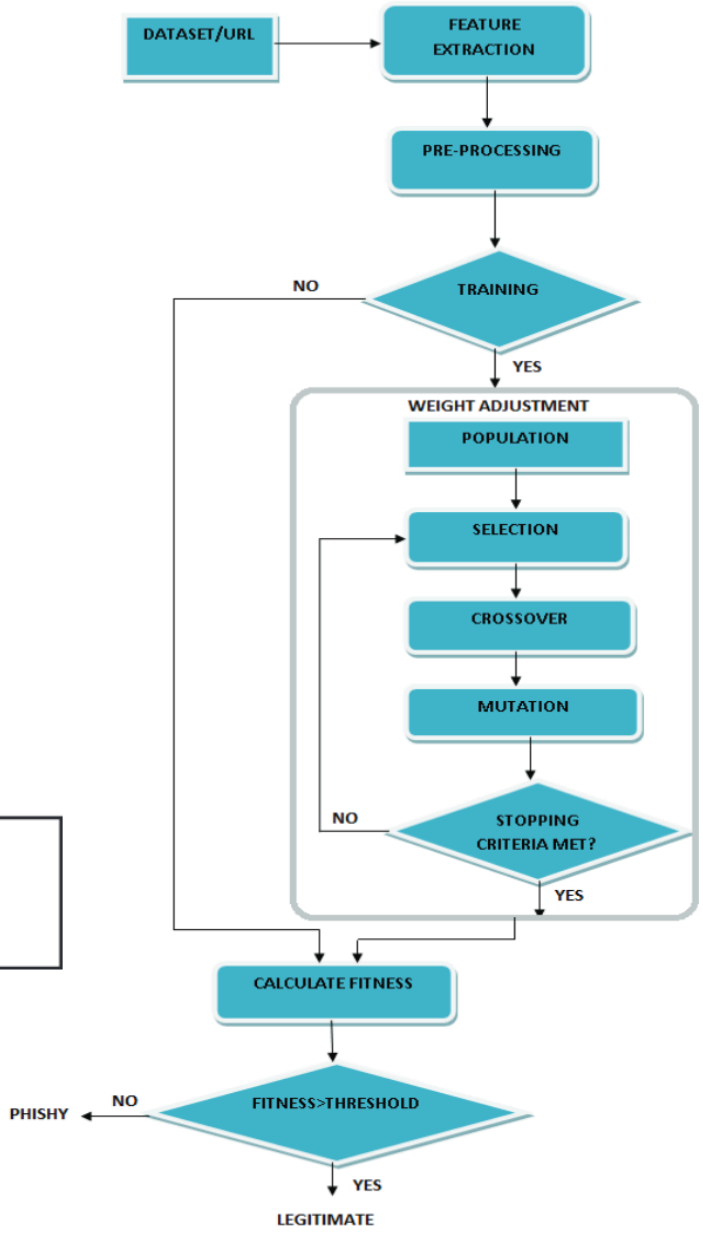
Rule: No DNS record → Phishy
else → Legitimate



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Abnormal URL: It can be extracted from whois database.
    - For Phishing websites, the whois server may not be specified so we cannot extract the information about the domain.
    - If whois server is not specified then it is assumed abnormal URL.
    - Rule for abnormal URL is:

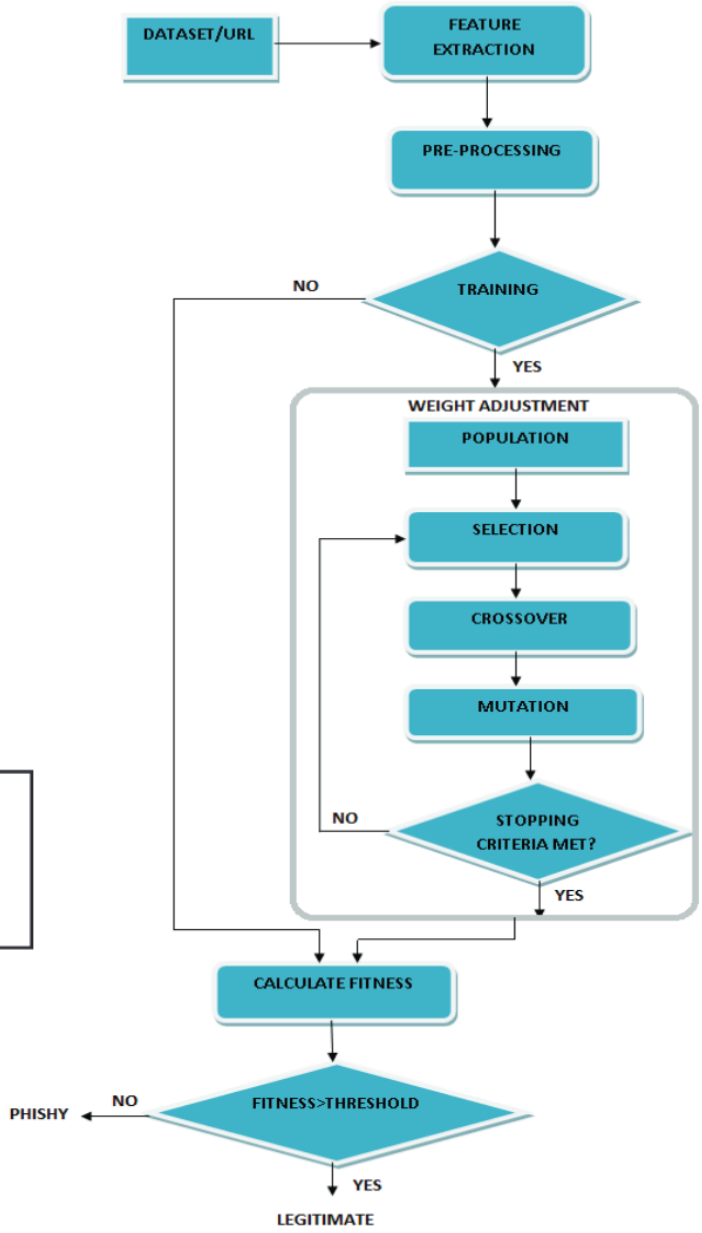
Rule: Whois Server not specified → Phishy
else → Legitimate



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Long URL: The Phishers may use very long URL's for hiding the suspicious part.
    - Heuristically, if the length of the URL is greater than or equal 54 characters then the URL is classified as phishing.
    - Rule for Long URL is :

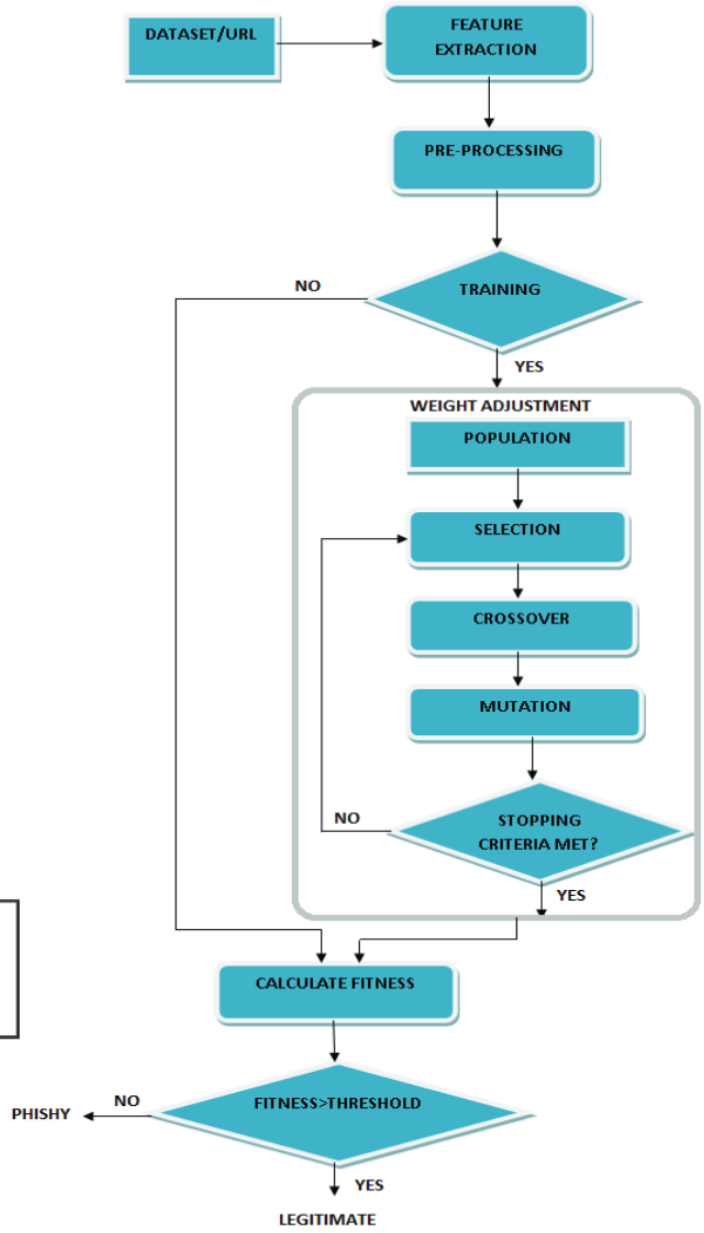
Rule: If URL length < 54 → Legitimate
URL length ≥ 54 and ≤ 75 → Suspicious
else → Phishy



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Prefix_Suffix: The Phishers try to deceive the users by adding prefix suffix using '-' to the domain names.
    - With this the users may feel they are dealing with legitimate site.
    - In legitimate domain names '-' character is rarely used.
    - If '-' is available in the domain name then it can be classified as Phishy.
    - Rule for prefix_suffix is:

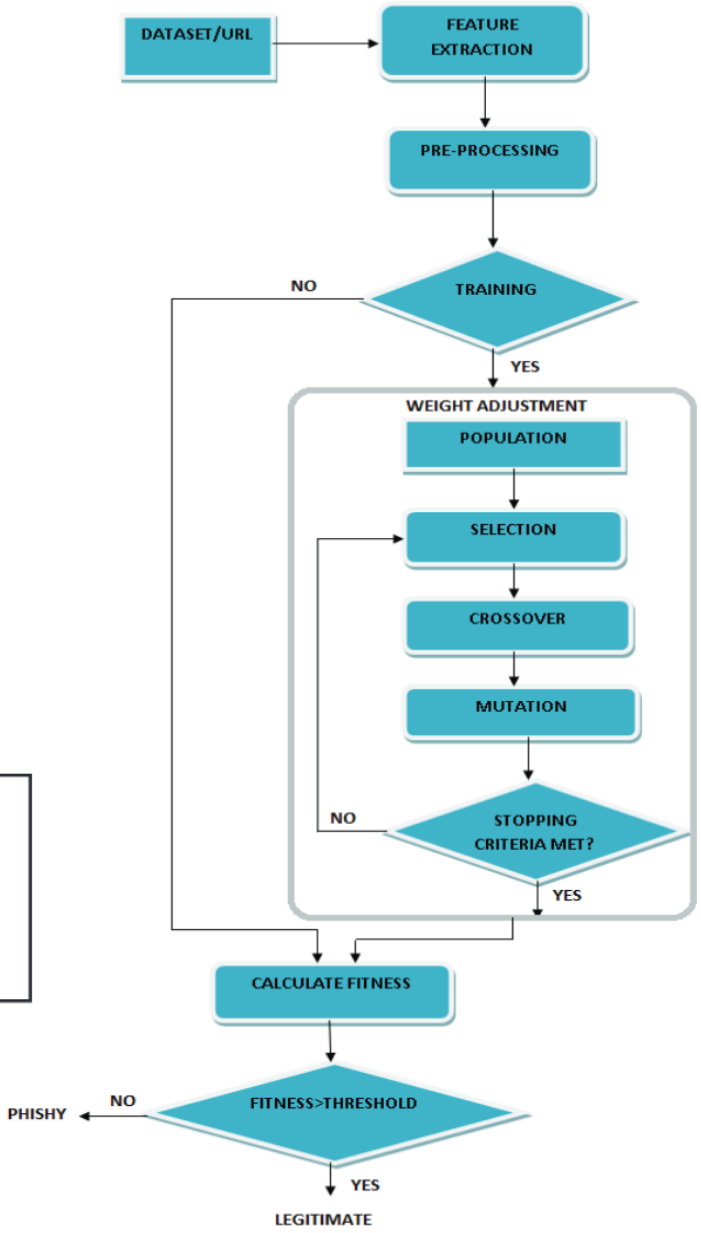
Rule: If domain part has '—' → Phishy
else → Legitimate



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Sub Domains: Phishing websites generally include many sub domains.
    - Ignoring www. the legitimate URL's generally have two dots in the URL.
    - If the URL contains many sub domains then it can be classified as Phishy.
    - Rule for Sub Domains is :

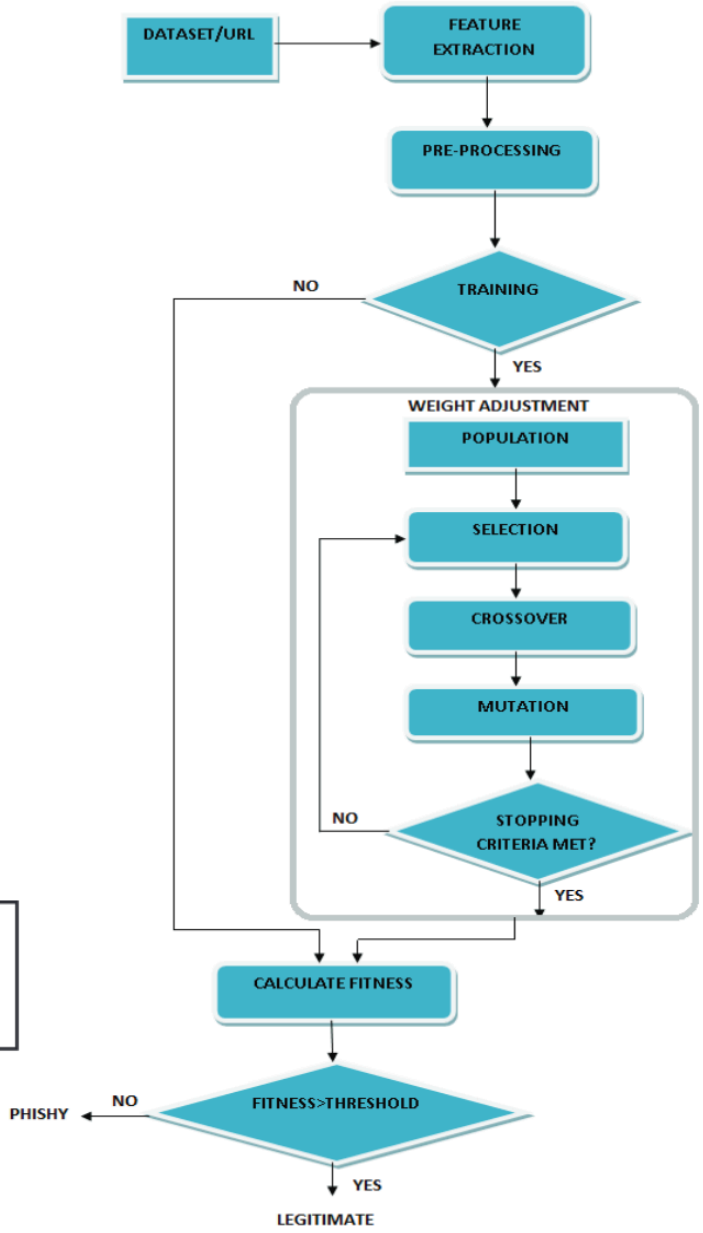
Rule: If dots in domain < 3 → Legitimate
else if =3 → Suspicious
else → Phishy



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - Http/Https: The Legitimate webpages that take input from the user use https protocol for security purpose.
    - The phishing webpages mostly uses http protocol.
    - if the webpage uses https protocol then it can be classified as legitimate otherwise phishy.
    - (having said that) Nowadays, it is possible to obtain fake security certificate for a webstie.
    - Rule for Http/Https is :

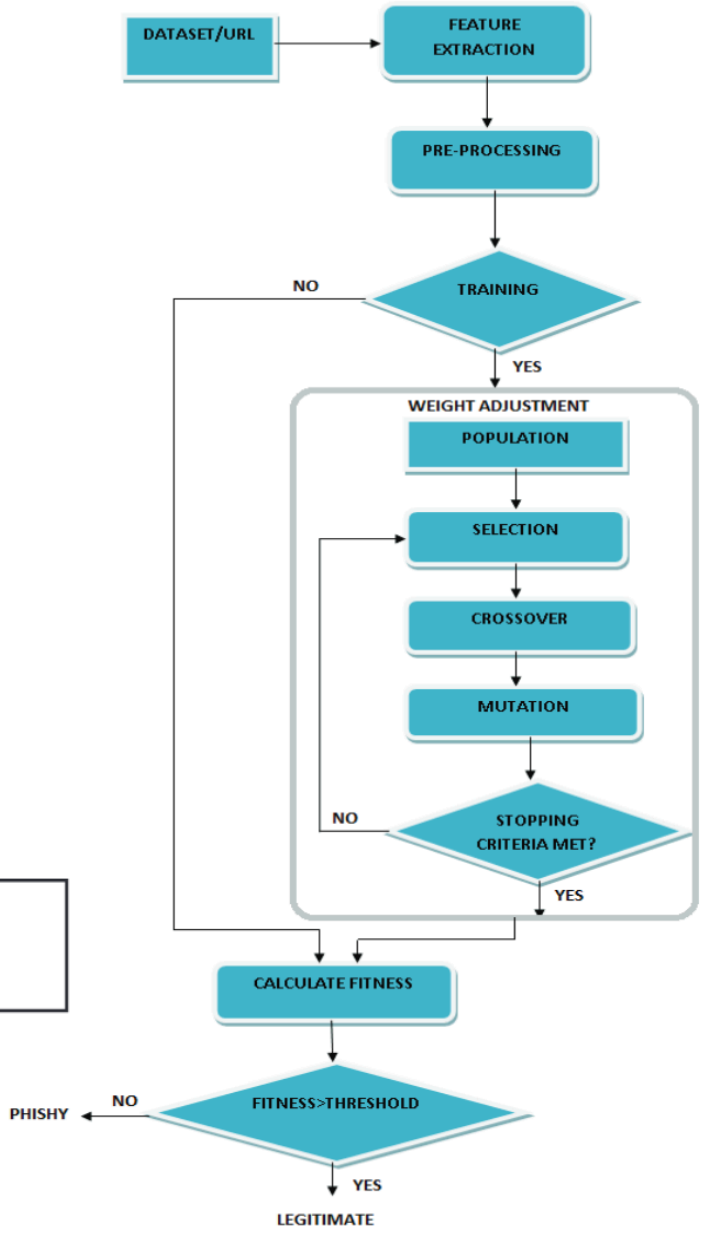
Rule: uses https → Legitimate
else → Phishy



Phishing URL – Detection Techniques

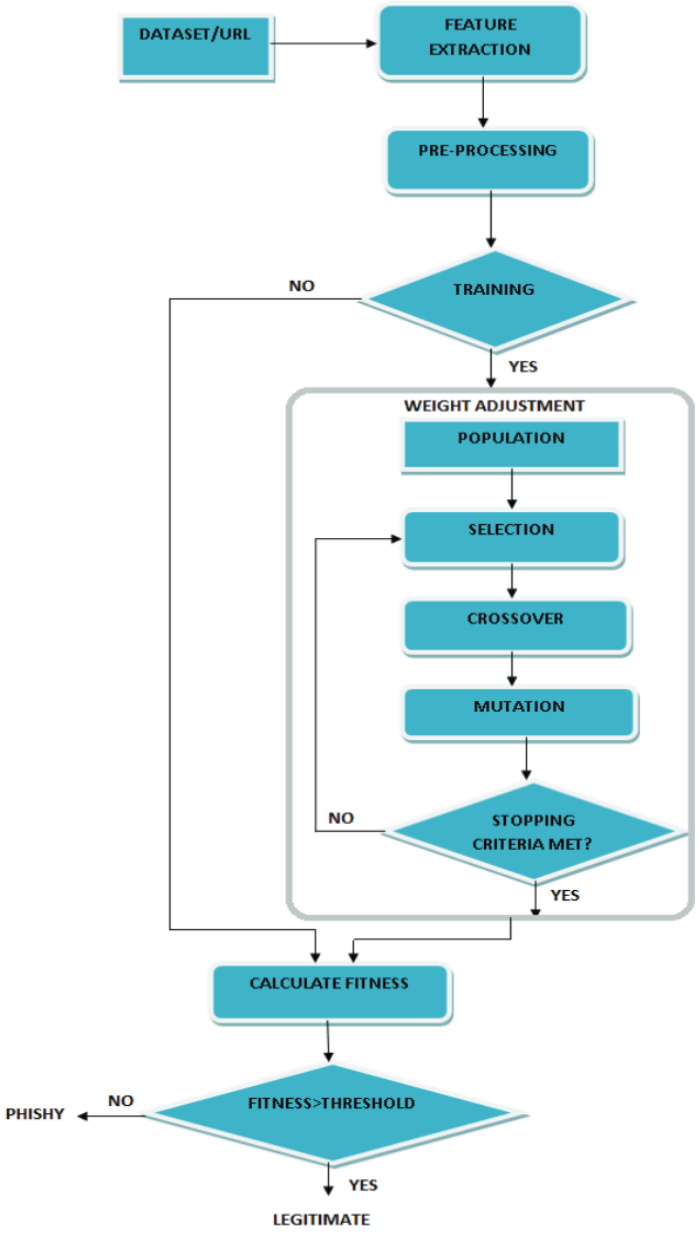
- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase I- Feature Extractions and rules:
    - IP Address: Users can remember the domain names of the websites.
    - It is difficult for them to remember IP addresses.
    - Phishers may use IP Address instead of domain names.
    - If the URL contains IP Address then one can be sure that it is a phishy website.
    - The rule for IP Address is

Rule: If IP address exists in URL → Phishy
else → Legitimate



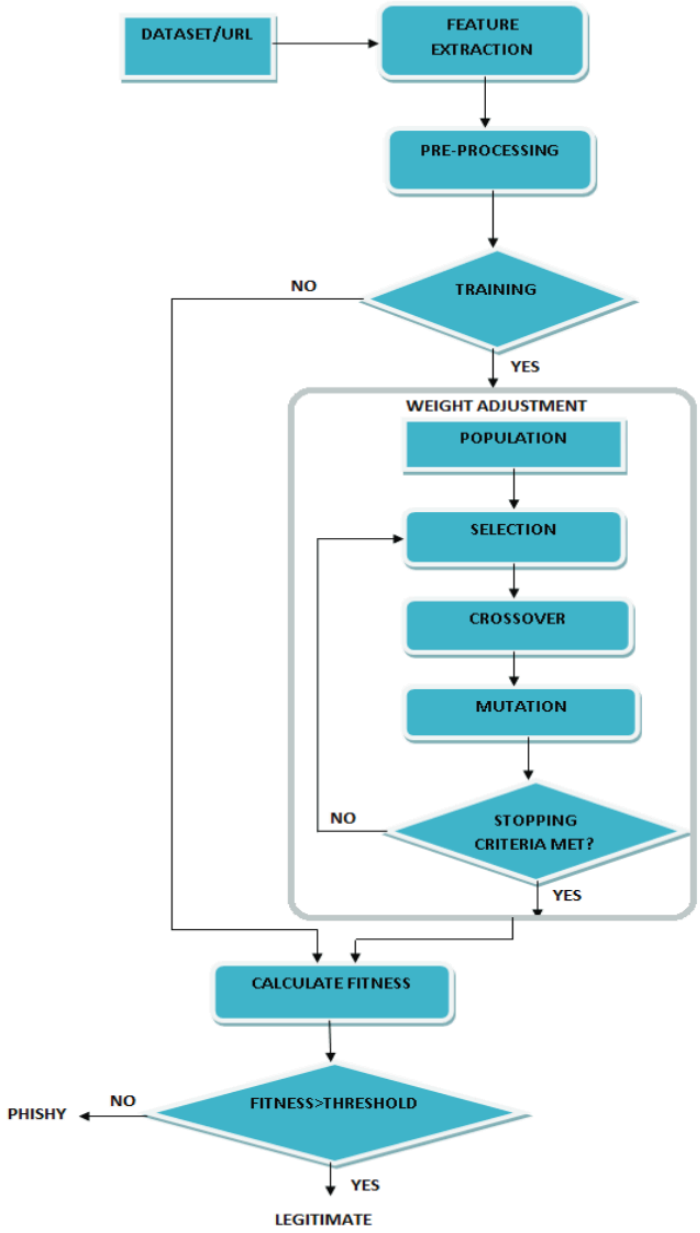
Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase II- Pre-processing:
    - The value of each feature is categorized into "phishing", "legitimate" or "suspicious" class.
    - Phishing, Suspicious and Legitimate classes are represented by -1, 0 and 1, respectively.



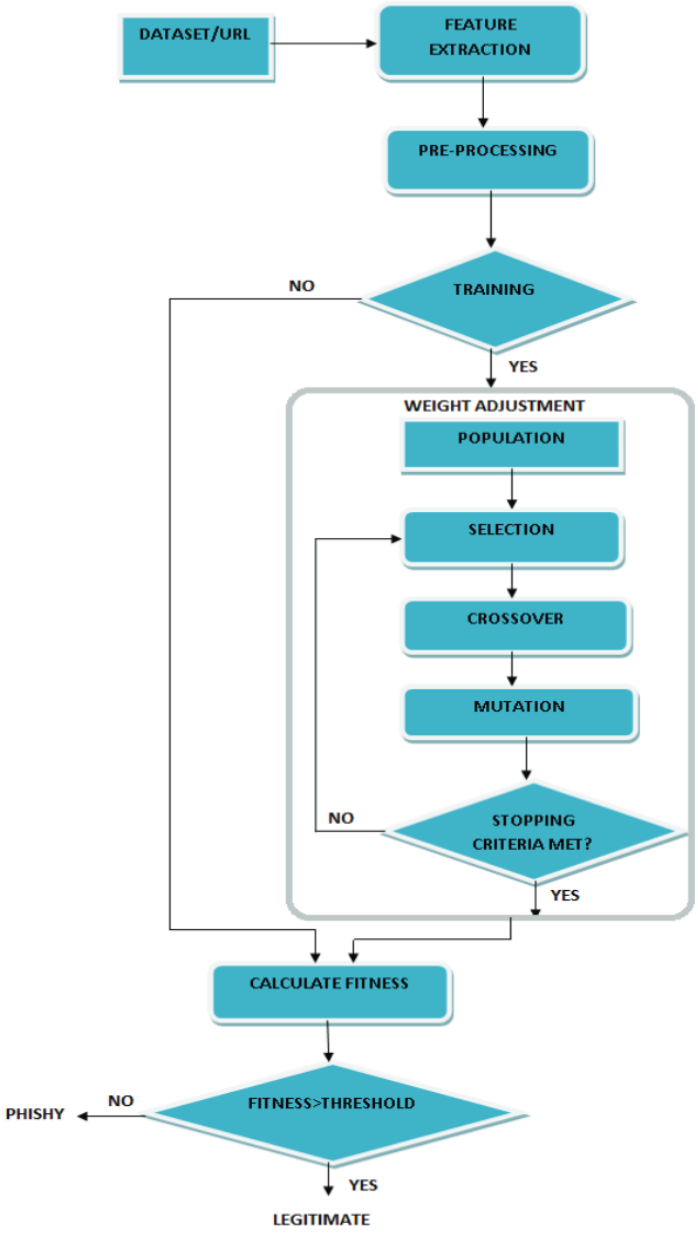
Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase III- Weight Adjustment:
    - Goal is to find the best weights that can classify the websites accurately.
    - Genetic algorithm is used for weight adjustment.



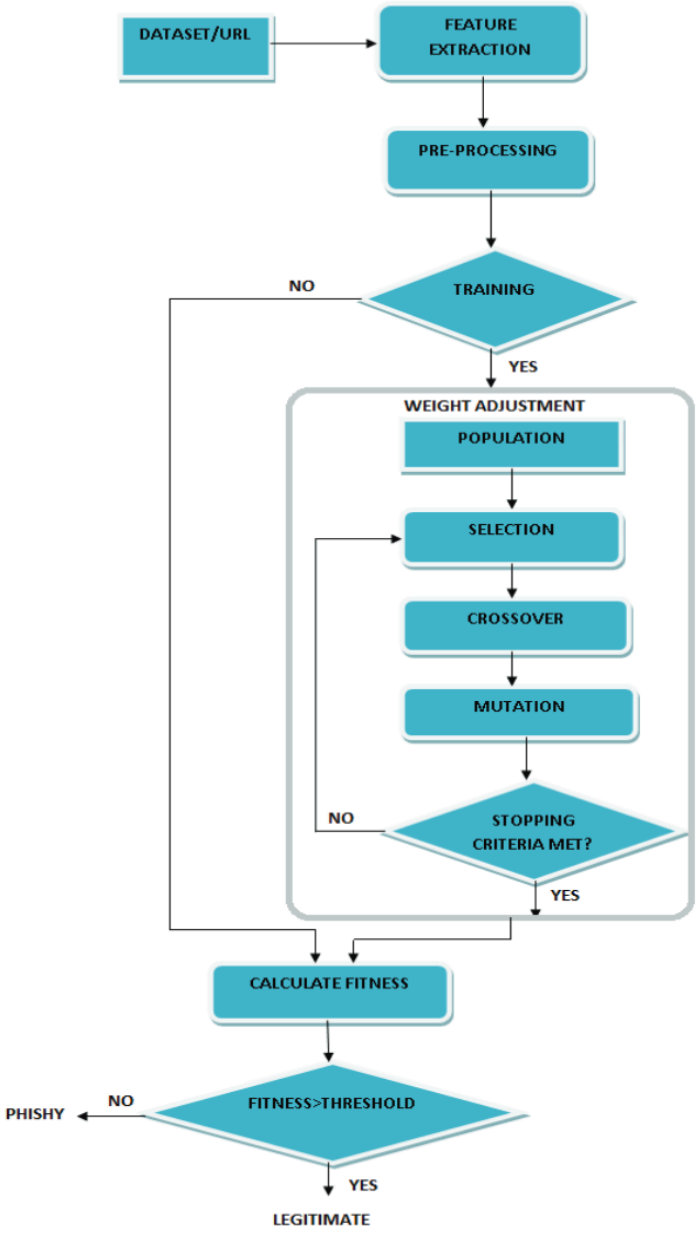
Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase III- Weight Adjustment:
    - Take information gain of each feature as its initial weight.
    - Using information gain as the initial weight enables assigning maximum weight to the feature that provides maximum information for classification.
    - The information gain is calculated from the training dataset.
    - Genetic algorithms help optimize the initial weights until the best optimal results achieved



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques :
  - Genetic algorithm [10]:
  - Phase IV- Results:
    - The best weights obtained in the third phase are used to calculate the fitness of the URLs in dataset
    - Each record in dataset is classified into phishy or legitimate by comparing the fitness with the threshold value.



Phishing URL – Detection Techniques

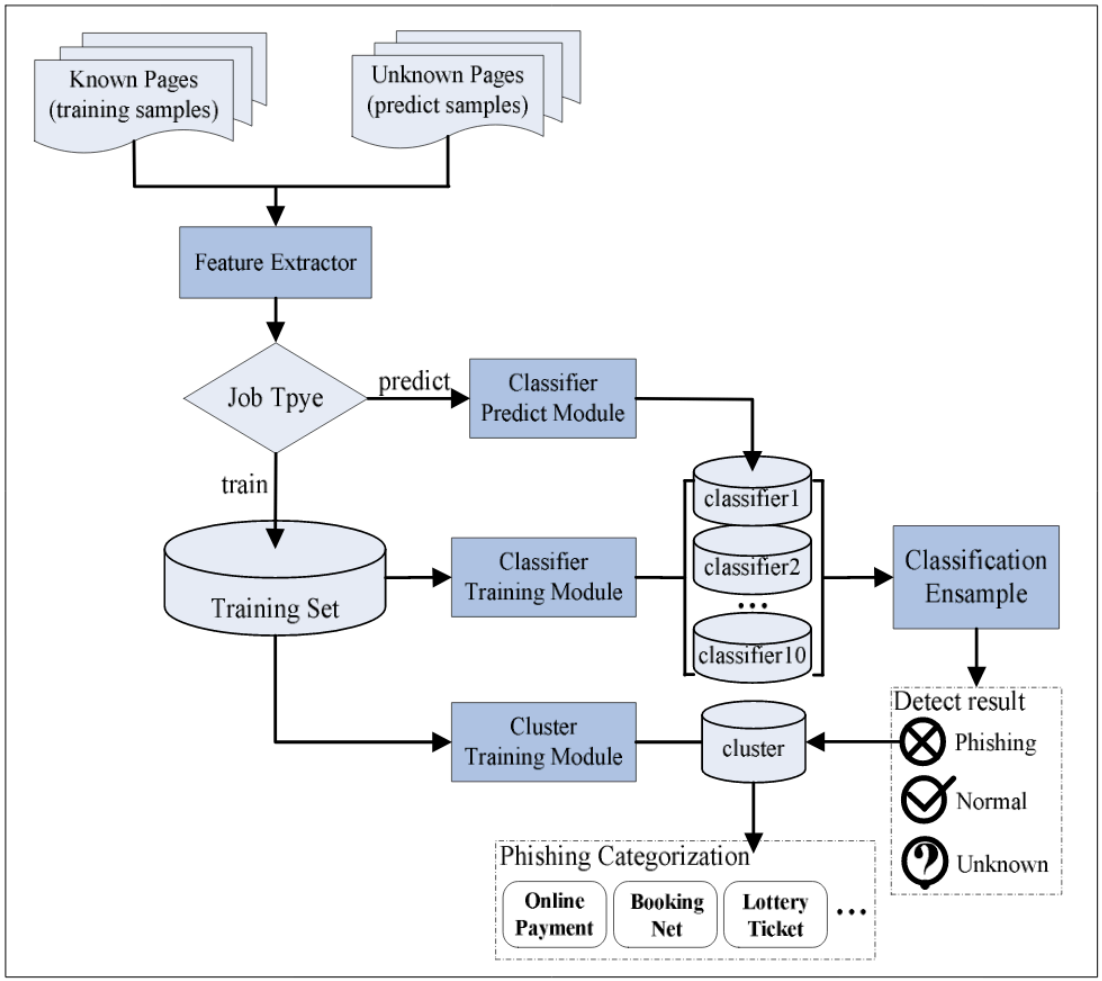
- Website Phishing Detection Techniques [9] :
  - Based on associative classification (AC) data mining
  - Performed in three phases.
    - Phase 1: search for hidden correlations among the attribute in the training data set and generate them as “Class Association Rule” in “If-Then” format.
    - Phase 2: rank and prune procedures start operating thresholds like confidence and support. The output of phase 2 is the set of Class Association Rule which represents the classifier.
    - Phase 3: the classifier derived gets evaluated on the test data to measure its effectiveness in forecasting the class of test data and the output of the last phase is the accuracy or error-rate of the classifier.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Intelligent detection and categorization model [11]:
    - Feature extractor:
      - Extract the terms from the webpages and then converts the terms to a group of 32-bit global IDs as the feature of the data collection
      - For training samples these integer vectors are transformed into term frequency features and collected in the database.
    - Classifier training module:
      - Ten heterogeneous classifiers are built based on the characteristic of different features; improved NBC (Naive Bayes Classifier) and SVM (Support Vector Machine) algorithm can be employed for the training.
    - Ensemble classification module:
      - Used to combine all the prediction results from heterogeneous classifiers.
      - Expected to have a better detection performance than individual classifier.
    - Cluster training module:
      - Hierarchical clustering algorithm is applied on the term frequency vectors with the TF-IDF weighting scheme.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Intelligent detection and categorization model [11]:



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Intelligent detection and categorization model [11]:
    - Feature extraction:

ID	Feature Name	Description
1	Title	Title of the webpage.
2	H1-H6	Content in the <h1> to <h6> tags.
3	Keyword	Keyword information in the Meta tag.
4	Description	Page description in the Meta tag.
5	Copyright	Copyright info in the Meta tag.
6	Link text	Corresponding text of the link.
7	Frame	Url address of the Frame.
8	Img	Url address of the Image.
9	Alt	Description text of the image.
10	String	All the other visible string of the page.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Intelligent detection and categorization model [11]:
    - Feature extraction:
    - Not all features are meaningful for phishing detection
    - There might be some redundant features
    - The model with redundant and meaningless terms might not be accurate
    - Use: Max Relevant Algorithm to determine the relevancy of features to phishing

$$Max\_Relevant(a_i f) = \sum\sum p(a_i, f) \log \frac{p(a_i, f)}{p(a_i) p(f)}$$

where a_i is the word with label i, f is the class label, $p(a_i)$ and $p(f)$ represent the frequencies of a_i and f respectively in the training samples, $p(a_i,f)$ is the frequency that a_i and f coexist.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Intelligent detection and categorization model [11]:
    - Cluster training module
    - A cluster is a collection of phishing websites that share some common traits between them and are "dissimilar" to the phishing websites belonging to other cluster
    - Hierarchical clustering algorithm

Input: The data set S.
Output: The best K clusters.

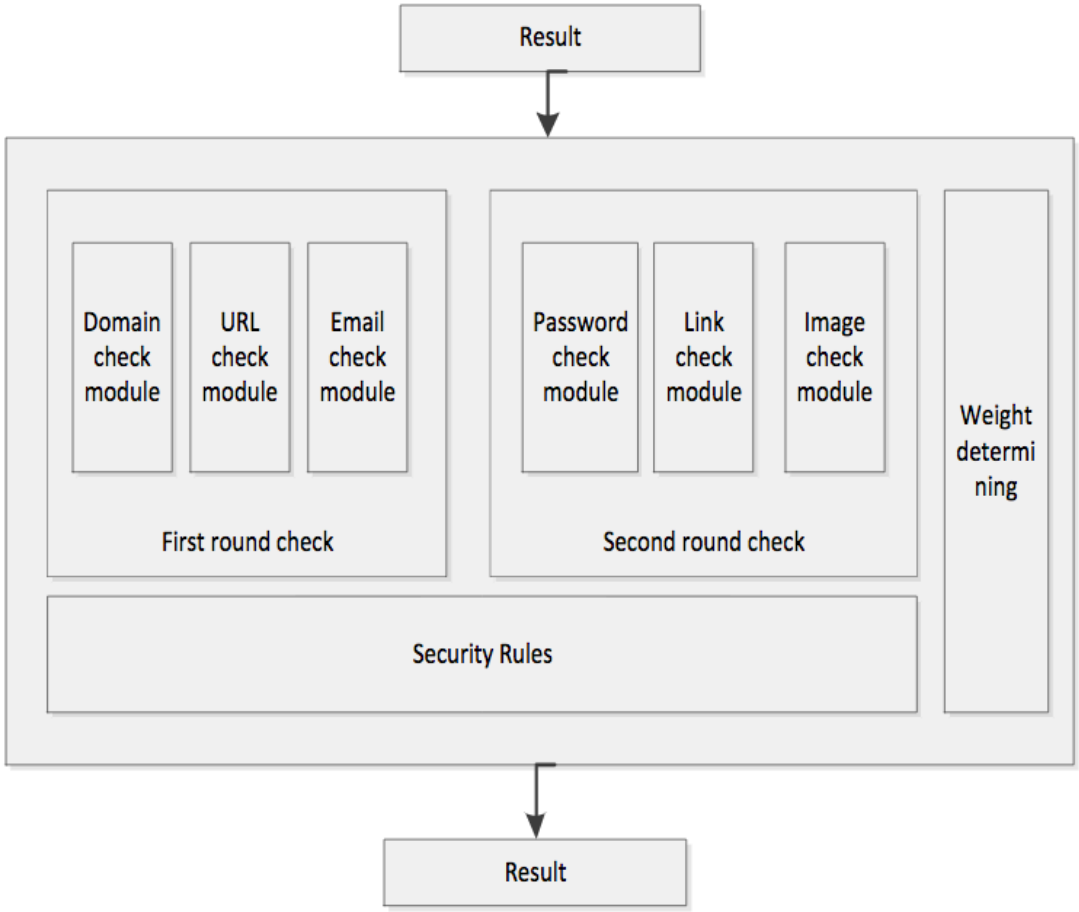
```
1   Set each data point in S as a singleton cluster;
2   For K from 1 to N in S, Do
3       Merge two closest clusters Ci and Cj into a new cluster Cm;
4       Calculate distance from Cm to all other clusters;
5       Update the distance matrix;
6       Calculate the validity index;
7       Compare and keep the best K clusters until now;
8   End For
9   Return the best K and the corresponding clusters.
```

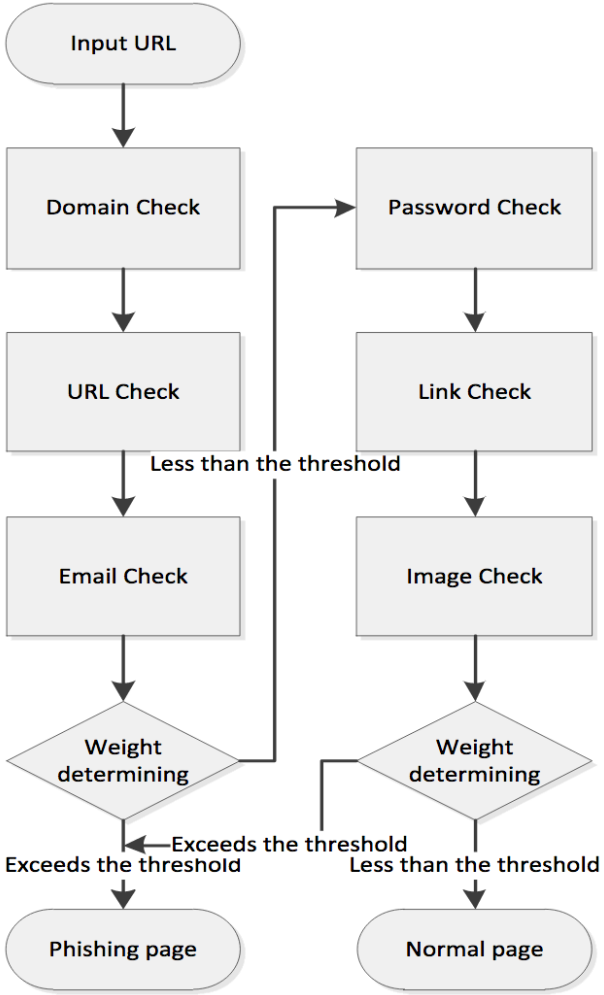
Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Intelligent detection and categorization model [11]:
    - Cluster training module
    - A cluster is a collection of phishing websites that share some common traits between them and are "dissimilar" to the phishing websites belonging to other cluster
    - Hierarchical clustering algorithm
    - Use cosine similarity to measure the distance between two data points

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Heuristics anti-phishing detection-based approaches [12]





Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Heuristics anti-phishing detection-based approaches [12]
  - URL check model:
    - It includes three steps:
    1. Check whether the URL contains any dubious username
    2. Check that the host name/domain name in the URL is not hidden.
    3. Check the page that will being navigated to is requested from a standard port.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Heuristics anti-phishing detection-based approaches [12]
  - Email check model:
    - It checks:
      - whether the clicked links link to Email addresses,
      - whether the current Email domain name is empty, and
      - whether an Email domain name is from a known website, such as yahoo.com and 163.com.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Heuristics anti-phishing detection-based approaches [12]
  - Password check model:
    - Check whether current page contains certain fields, such as 'password' or 'pass' or 'pwd'
    - if the page contains these fields and the fields hasn't been encrypted, the detection system needs to give a warning.

Phishing URL – Detection Techniques

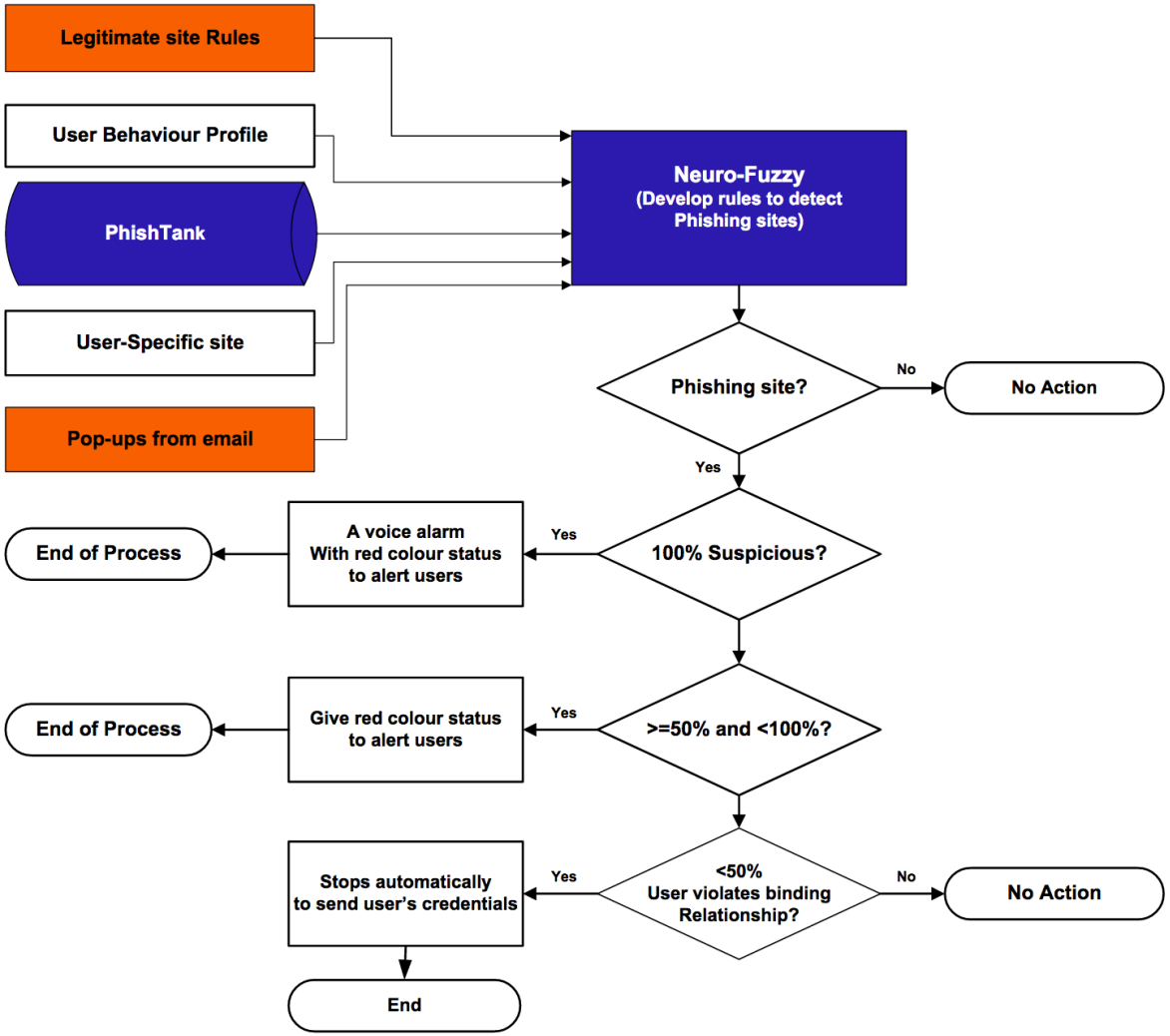
- Website Phishing Detection Techniques [9] :
  - Heuristics anti-phishing detection-based approaches [12]
  - Link check model:
    - Check whether current pages contain dubious links.
    - Dubious links: links which trigger warnings when they were was passing domain checks and URL checks
  - Image check model:
    - It compares images in current page with images from pages accessed before,
    - computes their hash values.
    - If an image in current page has the same hash value from one image accessed before, the system will give a warning.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Neuro-Fuzzy algorithm [13]
  - A combination of a fuzzy logic and a neural network with ability of reasoning and learning.
  - It has the abilities of data learning from neural network view point, and forms linguistic rules from fuzzy inference of
  - The advantage: universal approximations with ability to use Fuzzy IF...THEN rules.
    - Neural Network performs well when dealing with raw data,
    - Fuzzy Logic deals with reasoning on a higher level, using numerical and linguistic information from do-main expert.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Neuro-Fuzzy algorithm [13]



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Neuro-Fuzzy algorithm [13]
  - Five inputs or tables where features are extracted:
    1. Legitimate site rules: a summary of law covering phishing crime,
    2. User-behavior profile: a list of people's behavior when interacting with phishing and legitimate websites,
    3. PhishTank: a free community website operated by Open Domain Names where suspected websites are verified and voted as phish by the community experts,
    4. User-specific site: It contains binding requirements between a user and online transaction service providers,
    5. Pop-Ups from Email: regular phrases that are used by phishers as appear on screen.

Phishing URL – Detection Techniques

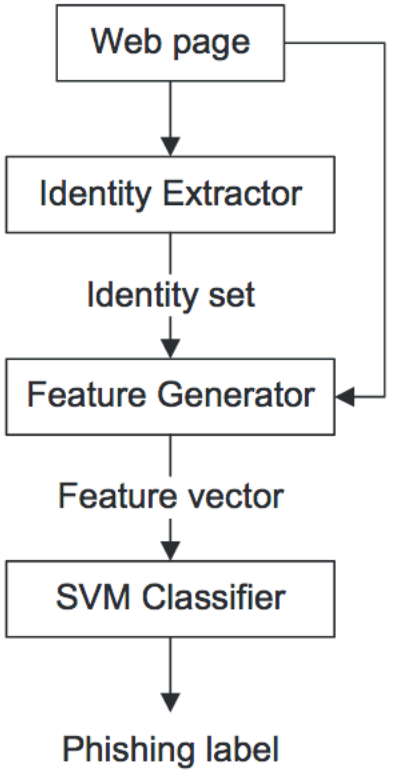
- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:

Input: a web page *P*

Output: label (-1: legitimate, 0: neutral, +1: phish)

Algorithm:

1. Construct a DOM tree of *P*
2. If *P* has a text input
   a. Extract the webpage identity//Corresponding to Identity Extractor//
   b. Generate features//Corresponding to Feature Generator//
   c. Classify *P* and return the result (-1 or +1)//Corresponding to SVM Classifier//
3. Else return neutral label (0)



Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - The webpage DOM tree, its HTTP transaction, and its identity set are inputted into the Feature Generation step to generate features in classifier format (e.g., SVM).
    - Term identify set extraction algorithm:

Input: the DOM tree of a webpage P

Output: term identity set I_t

Algorithm:

1. Extract texts from identity relevant DOM objects and properties
2. Text preprocessing
   a. Lowercase and tokenize text sentences into terms
   b. Omit short terms and stop words
3. Count *tf-idf* of each term
4. Return the top five highest *tf-idf* terms as the term identity set I_t.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - Feature 1: Suspicious page address
      - Feature 2: ID page address
      - Feature 3: Nil anchors
      - Feature 4: ID foreign anchors
      - Feature 5: Foreign anchors
      - Feature 6: ID foreign requests
      - Feature 7: Foreign requests
      - Feature 8: Cookie domain
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - Feature 1: Suspicious page address
        - » Low cost phishing web-sites usually do not have a domain name, the IP address cannot be resolved, e.g. http://61.164.117.9/ http://www.paypal.com/managament/cgi/.
        - » Whether a URL containsthe ‘‘@’’ character. This character in an URL means that the sequence of characters before character ‘‘@’’ is treated as the ‘‘userinfo’’ of the URL. E.g., , for the URLhttp:// ebay.com@www.-phish.com/, the ‘‘ebay.com’’ string is the user info part of the URL.
        - » F1is used to denote Feature 1 for the suspicious page address. If the URL of a webpage is in an irresolvable IP address form or there is a user info part, then F1= 1, otherwise F1=0.
      - Feature 2: ID page address
      - Feature 3: Nil anchors
      - Feature 4: ID foreign anchors
      - Feature 5: Foreign anchors
      - Feature 6: ID foreign requests
      - Feature 7: Foreign requests
      - Feature 8: Cookie domain
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - Feature 1: Suspicious page address
      - Feature 2: ID page address
        - » Legitimate websites: there is a relationship between members of term identity set to the base domain of its page URL. E.g., for the PayPal website http://www.paypal.com/, the base domain of the URL is PayPal.com. The term identity set of PayPalpage is {PayPal, merchant, online, ebay, account}, and one of them is a part of the base domain URL. They also have anchor links that point to its own domain.
        - » Phishing websites: they claim identity that is different from its base domain URL, either from its hyperlink structure or from its page content. Let us see a PayPal phishing page, whose URL http:// www.seccuritpya.com/ www.paypal.com/www.paypal.com/cgi-bin/index.htm. We then know that its basedomain is seccuritpya.com. This page has the majority of its anchor links pointing to PayPal.com base domain; hence the URL identityis ‘‘PayPal.com’’. The term identity set of this page is {PayPal, merchant, online, ebay, account}, all of them are not part of the base domain URL
        - » Feature F2: the ID page address. If the URL identity is different with the base domain of the page address, then F2= 1, else proceed with the base domain URL to term identity set comparison. If one of the term identity set is a substring of the base domain thenF2= 0; otherwise,F2=1
      - Feature 3: Nil anchors
      - Feature 4: ID foreign anchors
      - Feature 5: Foreign anchors
      - Feature 6: ID foreign requests
      - Feature 7: Foreign requests
      - Feature 8: Cookie domain

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - Feature 1: Suspicious page address
      - Feature 2: ID page address
      - Feature 3: Nil anchors
        - » A nil anchor is an anchor that points to nowhere, e.g. <ahref=‘‘javascript::void(0)’’>, <a href=‘‘#’’>, etc.
        - » The more nil anchors a page has, the more suspicious it becomes
        - » F3: to denote the feature for nil anchors. The feature value is set by f3 = #A-nil / #A-total. #A-nil is the number of nil anchors, #A-total is the total number o anchors.
      - Feature 4: ID foreign anchors
      - Feature 5: Foreign anchors
      - Feature 6: ID foreign requests
      - Feature 7: Foreign requests
      - Feature 8: Cookie domain
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 4: ID foreign anchors
        - » An anchor that points to a foreign domain
        - » If the URL identity of a webpage is a foreign domain, then we compare the domain of the anchor with the URL identity; otherwise, we compare it with the term identity set.
        - » F4: The feature value is set by f4 = #A-ID-Foreign / #A-Foreign. = #A-ID-Foreign is the number of ID foreign anchors, #A-Foreign is the total number of Foreign anchors.
      - Feature 5: Foreign anchors
      - Feature 6: ID foreign requests
      - Feature 7: Foreign requests
      - Feature 8: Cookie domain
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in
      - Feature 11: Number of dots in
      - Feature 12: Search engine

Input: the anchor list of a webpage P, the term identity set I_t and the URL identity I_u

Output: a_{id} value

Algorithm:

For each anchor a in the anchor list, do
If the domain of a is in foreign domain
If I_u is in a foreign domain
Compare the domain of a with I_u. Increment a_{id} if match.
Else //I_u is in its own domain//
Compare the domain of a with I_t. Increment a_{id} if it is matched with at least one of I_t set.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 5: Foreign anchors
        - » For any websites, it is normal to link to the foreign domains, too many foreign anchors would decrease the credibility of the website.
        - » F5: by f5 = #A-ForeignAnchor / #A-TotalAnchor. #A-ForeignAnchor is the number of foreign anchors, #A-TotalAnchor is the total number of anchors.
      - Feature 6: ID foreign requests
      - Feature 7: Foreign requests
      - Feature 8: Cookie domain
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 6: ID foreign requests
        - » To imitate the real website, phishing pages might request images, Javascript, CSS files and other objects from the real web-site.
        - » For each foreign request, we compare the domain with the URL identity if the URL identity is in a foreign domain; otherwise, if the URL identity is in the local domain, then we compare the domain of the request URL to the term identity set.
        - » The feature value is set by f6 = #A-ID-Foreign / #A-Foreign. = #A-ID-Foreign is the number of ID foreign requests, #A-Foreign is the total number of Foreign requests.
      - Feature 7: Foreign requests
      - Feature 8: Cookie domain
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 7: Foreign requests
        - » Similar to the foreign anchors, requests to the foreign domains are also a normal behavior. When there are too many foreign re-quests, the website could be less credible.
        - » F7: by f7 = #A-ForeignRequests / #A-TotalRequests. #A-ForeignRequests is the number of foreign requests, #A-TotalRequestsis the total number of requests.
      - Feature 8: Cookie domain
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 8: Cookie domain
        - » HTTP cookies are text data sent by a web server to a web client used for maintaining information about client users
        - » Cookies have a domain attribute, its value normally is the server domain which set the cookies.
        - » F8to denote the feature for Cookie domain. If a webpage has a domain cookie which is in a foreign domain then it may be suspicious, therefor eF8= 1.
        - » If a page has its own domain cookies or has no cookies then F8= 0.
      - Feature 9: SSL certificate
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 9: SSL certificate
        - » Secure login pages of legitimate websites, e.g. electronic commerce websites and internet banking, have an SSL certificate, while most phishing pages do not.
        - » Each certificate has website address information and this should match the page address which has the certificate.
        - » F9: feature.If the URL of the certificate is different with the page address URL then F9= +1.
        - » If no SSL certificate presents in a webpage then F9=0;otherwise, F9= 1.
      - Feature 10: Number of dots in page address
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 10: Number of dots in page address
        - » The URL address of a phishing page may have many dots to confuse users into thinking they are opening a legitimate website.
        - » F10isused to denote this feature and its value is simply the number of dots in the URL address.
      - Feature 11: Number of dots in all URLs
      - Feature 12: Search engine

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 11: Number of dots in all URLs
        - » This feature is similar to F10, but is applied for all URLs in a web-page including anchor URLs and request URLs. The hyperlink maybe in a relative form, therefore we first convert it into the absolute URL and then we count the number of dots.
      - Feature 12: Search engine

$$F_{11} = \frac{\sum_u d_u}{|U|},$$

where u is an URL in a webpage P, d_u is the number of dots in the URL u, and $|U|$ is the number of URLs.

Phishing URL – Detection Techniques

- Website Phishing Detection Techniques [9] :
  - Based on machine learning classifiers [14]:
    - Feature Generation 12 Features <f1, f2, …, f12>
      - ….
      - Feature 12: Search engine
        - » We perform a search on a search engine(we use Google) using the base domain and the term identity set of a page as the keyword.
        - » For example, for a PayPal phishing pageathttp://www.payyipal.com/whose term identity set is {PayPal,ebay, merchant, online, account}, then the keyword is: “PayPal.com PayPal ebay merchant online account”.
        - » Phishing pages usually have a low page ranking thus it would be harder to find them from search engines, while for legitimate web-site it is the opposite.
        - » If the domain of a webpage exists in the top 30 search results then we consider it as a legitimate page (F12= 0);otherwise, it is a phish (F12= 1).
        - » Zero search result from the search engine means that the page could not be found; hence it is also a phish. Note that sponsored links in the Google search results is ignored and is not a part of top 30 search results

Phishing URLs - Automation

- Algorithmic-based (Feature Extraction + Classification)
- Feature Engineering-based (Feature Extraction + Feature Selection + Classification)
- Examples of some classification schemes:
  - Naïve-based
  - Support Vector Machines
  - Random Forest
  - Neural Networks

Phishing URLs - Automation

- Anti-Phishing Tools:
- Help in decreasing the amount of contextual information
  - Google Safe Browsing
    - Notable feature: Use a blacklist of phishing URLs
    - Limitation: not recognize phishing sites not present in the blacklist
  - NetCraft Tool Bar
    - Notable feature: Risk (e.g., age) rating system used to identify phishing websites
    - Limitation: It involves using a database of sites. The database may not recognize new phishing sites
  - SpoofStick
    - Notable feature: Provide basic textual domain information to help users decide
    - Limitation: Not effective against spoofed sites opened in multiple frames
  - SpoofGuard (heuristic-based solution)
    - Notable feature:
    - Limitation:
  - SiteAdvisor
    - Notable feature: protects against spyware and ad-ware attacks.
    - Limitation: if a new phishing site does not have a rating in their database it might not be caught by this tool.

- Source: A Framework for Detection and Measurement of Phishing Attacks, G. S., Provos, N., Chew, M., Rubin, A.D.

References

1. Everything is in the Name−A URL based Approach for Phishing Detection, H. Tupsamudre, A. K. Singh, and S. Lodha, Cyber Security Cryptography and Machine Learning (pp.231-248).
2. Phishing Website Detection Using Machine Learning Algorithms: A Comparative Analysis of Phishing Websites Detection using XGBOOST Algorithm with other Machine Learning Algorithms, H. Musa, 2019
3. Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets, S. Marchal, K. Saari, N. Singh, and N. Asokan
4. Detecting Malicious URLs in E-mail – An Implementation, D. Ranganayakulu and C. Chellappan
5. A Framework for Detection and Measurement of Phishing Attacks, G. S., Provos, N., Chew, M., Rubin, A.D.
6. A Survey of URL-based Phishing Detection, E. S. Aung, C. T. Zan, and H. YAMANA
7. Detection of Phishing Webpages based on Visual Similarity, L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, X. Deng

References

8. Malicious URL Detection : A Survey. E. Sandi Aung, H. YAMANA
9. A Survey of Website Phishing Detection Techniques, Z. Shukla, K. Zala, R. Kotak
10. Detection of phishing webpages using weights computed through genetic algorithm, S. Kaur, A. Kaur
11. An intelligent anti-phishing strategy model for phishing website detection, W. Zhuang, Q. Jiang, T. Xiong
12. Financial websites oriented heuristic anti-phishing research, Y. Liu, M. Zhang
13. P. Barraclough, M. Hossain, M. Tahir, G. Sexton, N. Aslam, Intelligent phishing detection and protection scheme for online transactions
14. M. He, S. Horng, P. Fan, M. Khan, R. Run, J. Lai, R. Chen and A. Sutanto, "An efficient phishing webpage detector

Some Other Online Resources

1. Phishing - URL Analysis:
   - https://mlhale.github.io/nebraska-gencyber-modules/phishing/url-analysis/
2. Phishing URL Detection with ML, Ebubekir Buber,
   - https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5
3. How to Use URL Pattern Analysis for Phishing Detection & Mitigation, L. Havens,
   - https://info.phishlabs.com/blog/how-to-use-url-pattern-analysis-for-phishing-detection-mitigation
4. Anatomy of a Phishing Email, A. Gendre
   - https://www.vadesecure.com/en/anatomy-of-a-phishing-email/
5. URL Obfuscation—Still a Phisher's Phriend https://www.f5.com/labs/articles/threat-intelligence/url-obfuscationstill-a-phishers-phriend