A Hybrid Machine Learning based Phishing Website Detection Technique through Dimensionality Reduction

Nusrath Tabassum,
Farhin Faiza Neha, Md.
Shohrab Hossain

Department of CSE, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Husnu S. Narman

Department of Computer Sciences and Electrical Engineering, Marshall University, Huntington, WV, USA

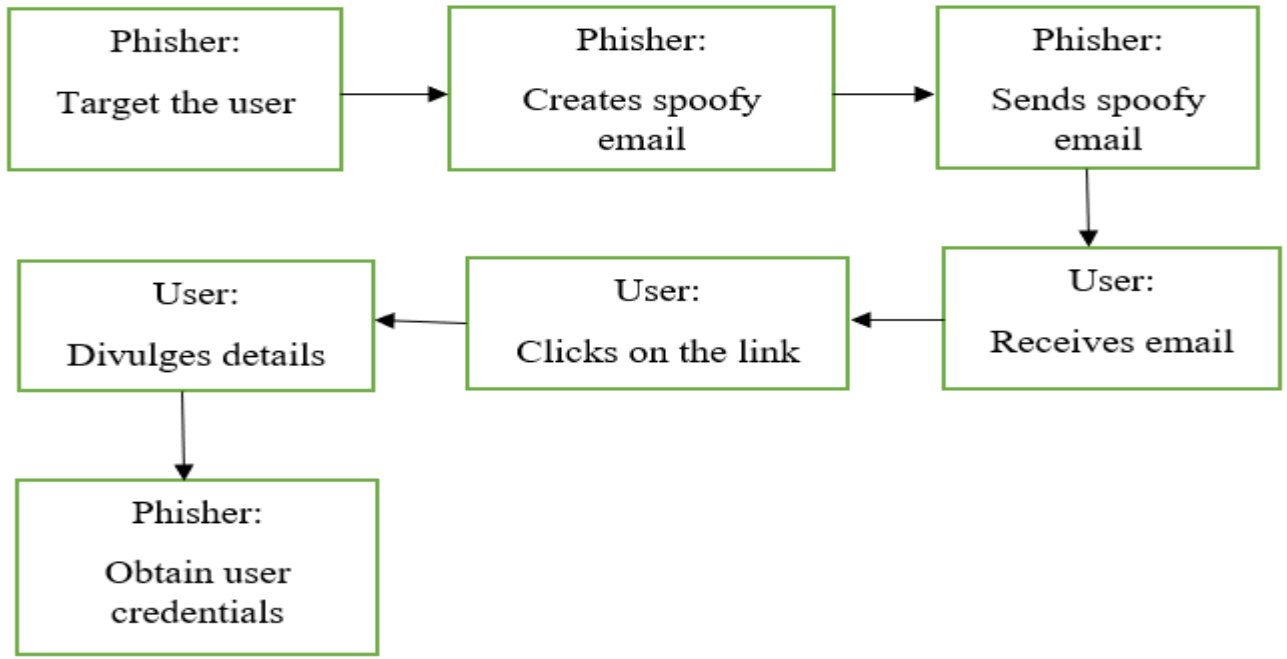
IEEE BlackSeaCom

May 24 – May 28

What is Phishing?

Phishing is a spiteful form of online identity theft that impersonates an honest firm’s website and aims at gaining authorized access to user’s individual information.

Phishing Life Cycle



Phishing Motives

- Financial gain
- Identity hiding
- Fame and notoriety

Approaches in designing technical anti-phishing solutions

- Blacklisting & Whitelisting based techniques
- Heuristic based techniques
- Content based techniques
- Visual similarity based techniques

Motivation

- Phishing attack results in identity theft and monetary losses
- It is important to detect phishing websites so that those malicious websites can be blocked by the firewall

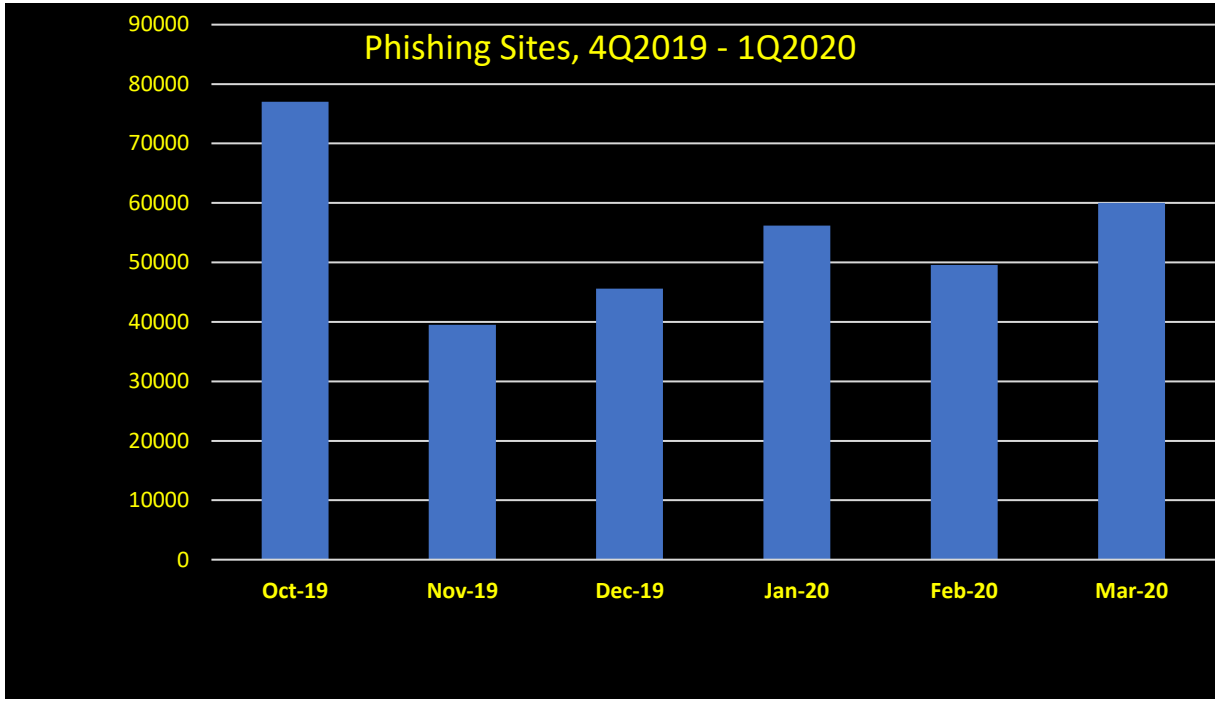


Fig. 1: Total phishing sites, 4Q2019 – 1Q2020 (according to APWG Phishing Activity Trends Report)

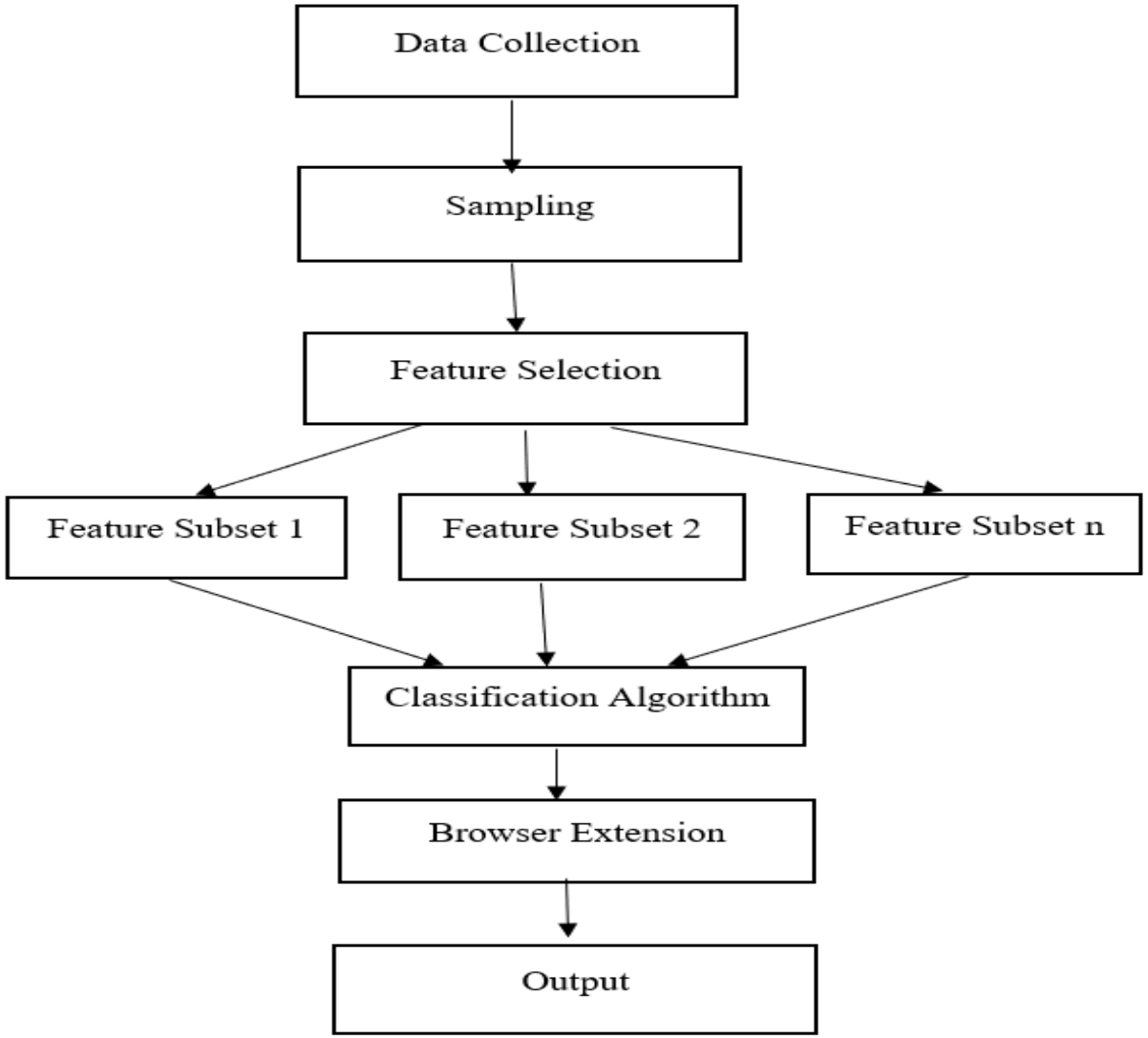
Objective

- To detect best subset of features so that phishing website detection can be made faster
- To identify the best performing classification algorithms

Contribution

- We have reduced the dimensionality of feature subset through the feature ranking.
- We have evaluated performance of the various classifiers and proposed the best hybrid classifier consisting of SVM, Decision Tree, Random Forest and XGBoost.

Proposed Methodology



Proposed Methodology(Contd.)

- Preparing Dataset:
  - The dataset was obtained from the UCI - Machine Learning Repository.
  - Dataset URL: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites

- Sampling:
  - 75% for training and 25% for testing.
  - Ran at least five times and select the average one.

Feature Categories for Phishing Detection

No.	Feature	Category
1	Using the IP Address	Address Bar based Features
2	URL-Length	
3	Shortining-Service	
4	having-At-Symbol	
5	double-slash-redirecting	
6	Prefix-Suffix	
7	having-Sub-Domain	
8	SSLfinal-State	
9	Domain-registration-length	
10	Favicon	
11	port	
12	HTTPS-token	
13	Request-URL	Abnormal Based Features
14	URL-of-Anchor	
15	Links-in-tags	
16	SFH	
17	Submitting-to-email	
18	Abnormal-URL	
19	Redirect	HTML and JavaScript based Features
20	on-mouseover	
21	RightClick	
22	popUpWidnow	
23	Iframe	
24	age-of-domain	Domain based Features
25	DNSRecord	
26	web-traffic	
27	Page-Rank	
28	Google-Index	
29	Links-pointing-to-page	
30	Statistical-report	

Feature Number	Feature Name	Feature Explanation
F0	Using IP Address	Phishing: IP address exists in domain part Legitimate: IP address does not exist in domain part
F1	URL Length	Phishing: URL length >75 Suspicious: URL length >=54 and <=75 Legitimate: URL length <54
F2	Using URL Shortening Service	Phishing: Use of Tiny URL Legitimate: Otherwise
F3	URL having the @ symbol	Phishing: URL having @ symbol Legitimate: Otherwise
F4	URL has redirect symbol	Phishing:The position of the last occurrence of "//" in the URL >7 Legitimate: Otherwise
F5	Prefix or suffix	Phishing: Domain name part includes (-) symbol Legitimate: Otherwise
F6	Having subdomains	Phishing: After omitting www. and .ccTLD if dots in domain part > 2 Suspicious: Remaining dots in domain part = 2 Legitimate: Remaining dots in domain part = 1
F7	SSL final state	Phishing: Use https and Issuer Is not trusted and age of certificate <= 1 year. Suspicious: Use https and Issuer Is not trusted. Legitimate: Use https and Issuer Is trusted and age of certificate >= 1 year
F8	Domain registration length	Phishing: Domain expires on <= 1 year Legitimate: Otherwise
F9	Having Favicon	Phishing: Favicon loaded from external domain Legitimate: Otherwise
F10	Having non standard port	Phishers take advantage if a URL has some open ports.
F11	HTTPS token	Phishing: Use HTTP token in domain part of the URL Legitimate: Otherwise

Address bar based features

Feature Number	Feature Name	Feature Explanation
F12	Request URL	The webpage address and most of the objects within the webpage have same domain then we consider it legitimate based on the percentage.
F13	Anchor URL	If the <a>tags and the website have different domain names then we count it suspicious or phishing based on the percentage.
F14	Links in tags	If the <Meta>, <Script>, <Link>and the website have different domain names then we consider it suspicious or spoofy based on the percentage.
F15	Server from handler	If SFH is blank or empty, it is considered as phishing. If SFH refers to a different domain, then it is suspicious.
F16	Submitting to email	If "mail()" or "mailto" PHP function is used,it is considered as phishing.
F17	Abnormal URL	If the host name is not included in the URL, it is classified as phishing.

Abnormal based features

Feature Number	Feature Name	Feature Explanation
F18	Redirect	If a website page is redirected less than or equal one, it is considered as legitimate. If a website page is redirected at least four times, it is marked as phishing.Otherwise it is suspicious.
F19	Status bar customization	If onMouseOver changes status bar, it is marked as phishing.
F20	Disabling right click	If the right click is disabled, it is considered as phishing.
F21	Having pop up window	If the pop-up window asks users to submit their personal details then we can count it spoofy.
F22	Iframe redirect	If iframe is used, it is referred as phishing.

HTML & JavaScript based features

Feature Number	Feature Name	Feature Explanation
F23	Age of domain	If the age of domain is greater than or equal 6 months, it is classified as legitimate.
F24	DNS record	If the DNS record for the domain is not found, it is marked as phishing website.
F25	Web traffic	A higher ranked website has less chance of being spoofy. If the domain has no traffic or is not recognized by Alexa database, it is considered as phishing.
F26	Page rank	If the page rank is less than 0.2, it is marked as phishing.
F27	Google indexed	If the website is in Google's index, it is classified as legitimate.
F28	Links pointing to page	If number of links pointing to the website is zero, it is considered as phishing. Because phishing websites have short life span.
F29	Statistical report	If the host of the website belongs to any top phishing domains, it is classified as phishing.

Domain based features

Feature Selection

- To rank the features, we have used:
  - ✓Random Forest
  - ✓XGBoost
  - ✓Correlation matrix with heatmap

Feature Selection (Contd.)

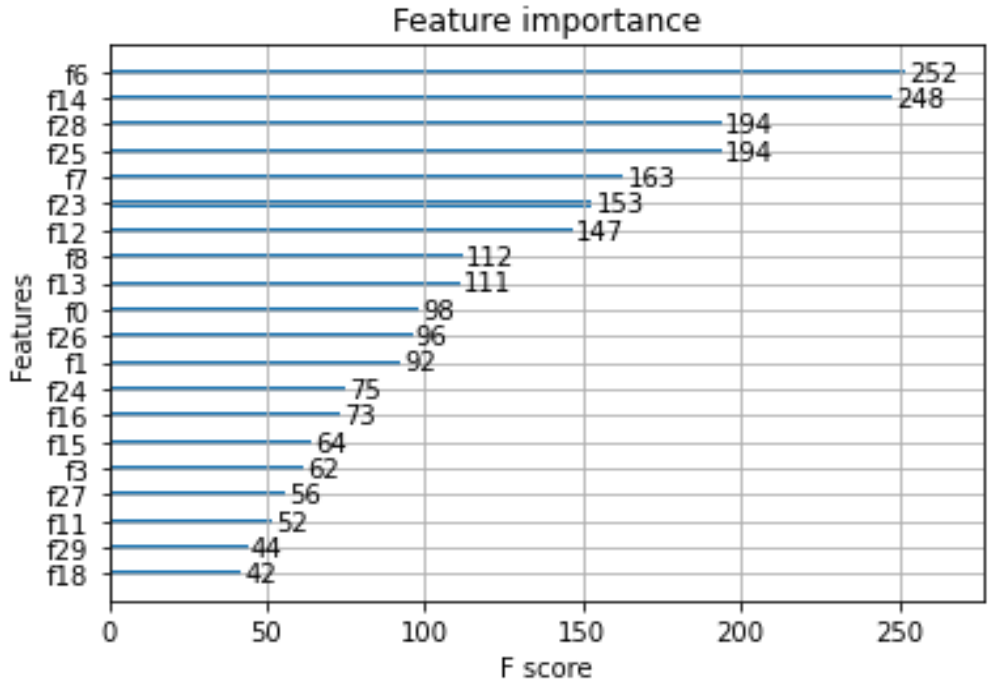


Fig. 2: Using XGBoost

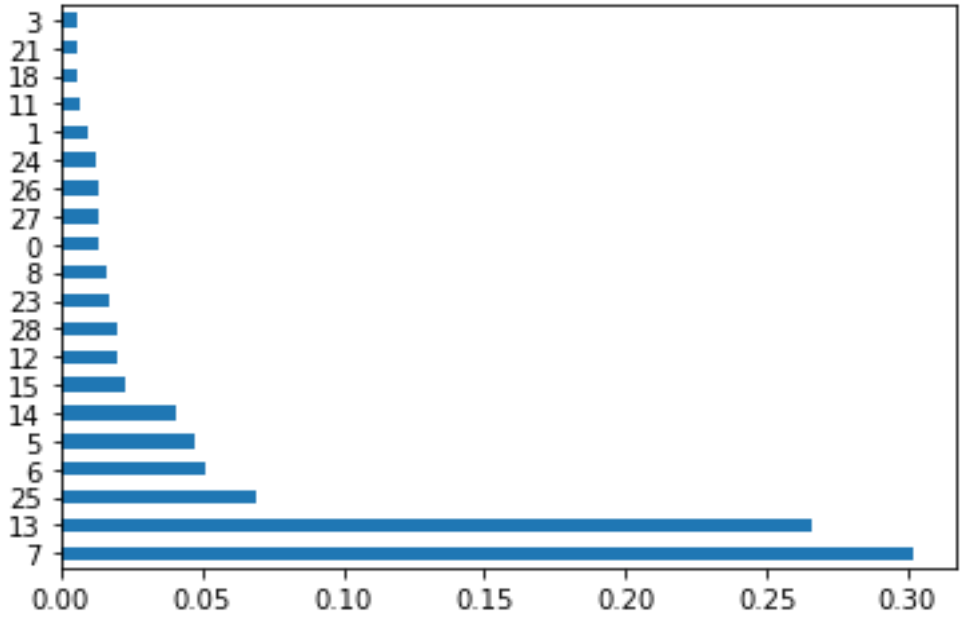


Fig. 3: Using Random Forest

Feature Selection (Contd.)

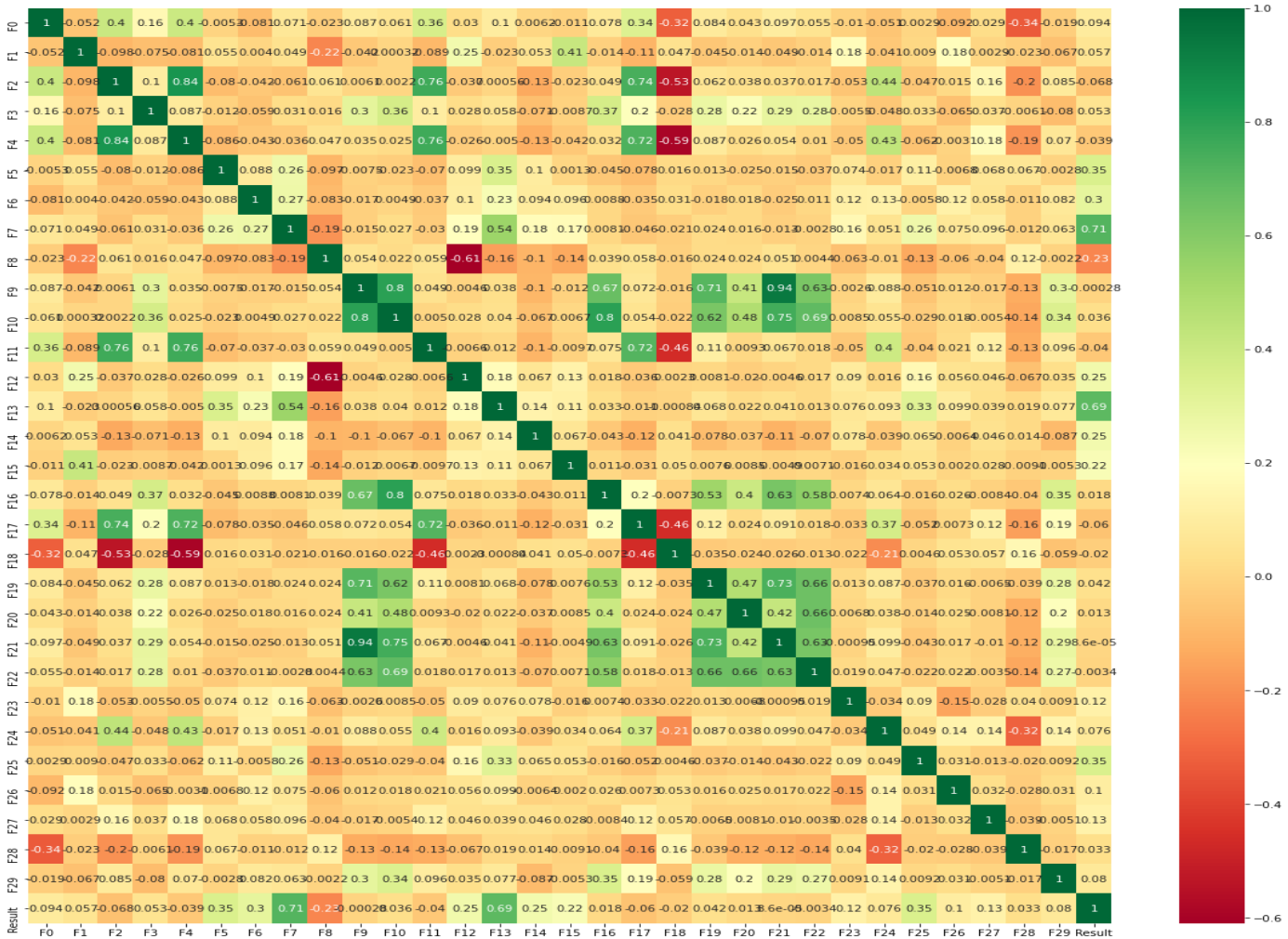


Fig. 4: Correlation matrix with heatmap

Feature Selection (Contd.)

SL.	Feature Subsets	Accuracy
1	F5, F6, F7, F13, F14, F25	93.60%
2	F6, F7, F8, F12, F13, F14, F23, F25, F28	94.21%
3	F5, F6, F7, F12, F13, F14, F15, F23, F25, F26, F27	94.46%
4	F0, F5, F6, F7, F12, F13, F14, F15, F23, F24, F25, F26, F27, F29	96.24%
5	F0, F1, F3, F5, F6, F7, F10, F11, F12, F13, F14, F15, F16, F20, F21, F23, F24, F25, F26, F27, F29	95.93%
6	F0, F1, F3, F5, F6, F7, F8, F10, F11, F12, F13, F14, F15, F16, F20, F21, F23, F24, F25, F26, F27, F28, F29	98.28%

Accuracy for several feature subsets using proposed hybrid classifier

Selected Features

- Finally our proposed features are:

  - F0, F1, F3, F5, F6, F7, F8, F10, F11, F12, F13,F14, F15, F16, F20, F21, F23, F24, F25, F26, F27, F28, F29.

Classification Algorithm

- Naïve Bayes
- Logistic Regression
- Support Vector Machine
- Decision Tree
- Random Forest
- XGBoost
- Several Hybrid Classifiers

Performance Evaluation

- Precision
  - Precision= TP/(TP+FP)
- Recall
  - Recall= TP/(TP+FN)
- F1-score
  - F1-score= (2*Precision*Recall)/(Precision+Recall)
- Accuracy
  - Accuracy= (TP+TN)/(TP+TN+FP+FN)

Performance Evaluation of all classifiers

Classifier	Accuracy	Precision	Recall	F1-score
Naïve Bayes	0.6187	0.77	0.65	0.58
Logistic Regression	0.9266	0.93	0.92	0.93
SVM	0.9273	0.93	0.93	0.93
DT	0.9616	0.96	0.96	0.96
RF	0.9710	0.97	0.97	0.97
XGBoost	0.9685	0.97	0.97	0.97
RF and XGBoost	0.9739	0.97	0.97	0.97
DT and XGBoost	0.9631	0.96	0.96	0.96
DT and RF	0.9652	0.97	0.96	0.96
DT, RF and XGBoost	0.9743	0.98	0.97	0.97
SVM, DT and XGBoost	0.9736	0.97	0.97	0.97
SVM, DT and RF	0.9739	0.98	0.97	0.97
LR, DT, RF and XGBoost	0.9758	0.98	0.97	0.98
SVM, DT, RF and XGBoost	0.9772	0.98	0.98	0.98

For 30 features
(Without feature selection)

Classifier	Accuracy	Precision	Recall	F1-score
Naïve Bayes	62.05%	0.77	0.65	0.58
Logistic Regression	92.58%	0.92	0.92	0.92
SVM	92.85%	0.93	0.92	0.92
DT	96.56%	0.97	0.97	0.97
RF	97.19%	0.97	0.97	0.97
XGBoost	97.47%	0.97	0.97	0.97
RF and XGBoost	97.38%	0.97	0.97	0.97
DT and XGBoost	96.83%	0.97	0.97	0.97
DT and RF	97.01%	0.97	0.97	0.97
DT, RF and XGBoost	97.47%	0.98	0.97	0.97
SVM, DT and XGBoost	97.64%	0.97	0.98	0.97
SVM, DT and RF	97.06%	0.97	0.97	0.97
LR, DT, RF and XGBoost	97.83%	0.98	0.97	0.98
SVM, DT, RF and XGBoost	98.28%	0.98	0.98	0.98

For 23 features
(With feature selection)

Comparison

	Proposed Method	Accuracy	F1-Score	Number of Features
Abdulrahman et al. [11]	Hybrid classifier (RF and XGBoost)	97.26%	0.9721	24
Das et al. [12]	LSTM	96.55%	0.969	30
Our proposed method	Hybrid classifier (SVM, DT, RF & XGBoost)	98.28%	0.98	23

Comparison with previous works for the same dataset

Conclusion

- Our proposed hybrid classifier will help the Internet users verify authentic websites.
- So our system will mitigate the risk of phishing websites.


Thank
you