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

Nusrath Tabassum[1], Farhin Faiza Neha[1], Md. Shohrab Hossain[1], and Husnu S. Narman[2]
[1]Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Bangladesh
[2]Department of Computer Sciences and Electrical Engineering, Marshall University, Huntington, WV, USA
Email: nusrathtabassum13@gmail.com, farhinfaiza@gmail.com, mshohrabhossain@cse.buet.ac.bd, narman@marshall.edu

***Abstract*—Phishing attacks are generally launched through emails or websites to acquire unauthorized access to the user's sensitive information. In recent times, many users face monetary losses due to phishing attacks. The motivation of our study is to present a prudent framework for detecting phishing websites to save users from being affected. Previous works used several supervised machine learning algorithms for classification to acquire higher accuracy for detection of phishing sites. In this paper, we have proposed a hybrid technique comprising of SVM, Decision tree, Random Forest, XGBoost by combining the idea of bagging and boosting. We have used the features of both phishing and legitimate website to mitigate the risk of phishing websites. We have evaluated classification algorithms using a number of feature subsets selected by various feature selection techniques to ascertain the most effective and efficient subset of features. Our hybrid technique achieved an accuracy of 98.28%, outperforming the state-of-the-art techniques.**

***Index Terms*—Phishing Attack; Feature Selection; Hybrid Classifier; Machine Learning; Browser Extension.**

I. Introduction

Internet has become one of the most popular, metamorphic and fast-growing technologies. The number of Internet users has increased from 413 million in 2000 to 4.54 billion in 2020 globally. Using this transformative technology, cyber-criminals often try to spread malware, illicit information, images and so on. Phishing scams and malware are the two general types of cybercrime. Phishing is a spiteful form of online identity theft that aims at gaining authorized access to user's individual information. It impersonates an honest firm's website. A common phishing tactic is to send spam emails or direct you to a fake website appeared to be legitimate and well-known individual or institution to persuade individuals to disclose personal information, such as password, credit card number, social security number, bank account number, financial data and so on often for malicious intent in an electronic communication. Phishers are those attackers who plan phishing attacks. They create phishing websites that look similar to the legitimate ones to emulate original websites for stealing user's personal and sensitive details. The information achieved by attackers are often utilized to access users confidential accounts such as twitter, facebook, email, bank etc. Many users put up with identity theft and financial losses due to the increasing number of phishing attacks [1].

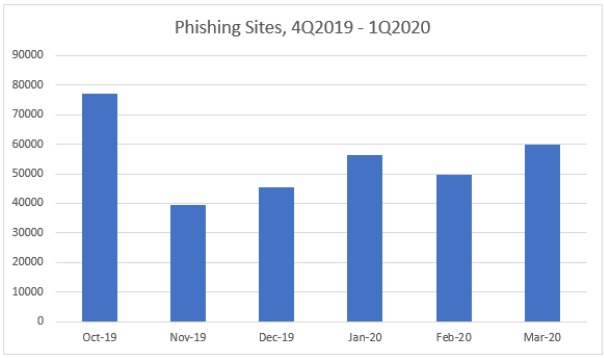

Fig. 1. Total phishing sites, 4Q2019 – 1Q2020 (according to APWG)

Due to the advancement in technology, security concerns have been increasing for various sectors like banking, education, entertainment and so on. According to Gartner, U.S. banks and credit card companies have lost 2.8 billion dollar annually due to the theft through phishing attacks [2]. According to APWG report [3], 165772 phishing sites have been detected in the first quarter of 2020 and 162155 phishing sites have been identified in last quarter of 2019 (see Fig. 1).It is a matter of great concern that attackers focus on acquiring access to corporate accounts that pertain sensitive and confidential financial information.

There have been few works on phishing website detection. Some of the works are based on Blacklist and Whitelist based technique [4]. Some are based on Content-based approach [5]. Some are based on Visual similarity-based techniques [6]. Some are Heuristics and machine learning-based techniques [7]. Abdelhamid et al. [2] examined the problem of website phishing attack using Multi-label Classifier based Associative Classification (MCAC). Nearly 94.5% accuracy was obtained using MCAC. However, their model used a dataset containing only 601 legitimate and 752 phishing websites. Only 16 features were utilized to detect phishing attack whereas there are other important features that could have been used for precise detection. In [8], the authors applied only Naive Bayes and sequential minimal optimization on two feature subsets (CFS and consistency subset) and could

achieve an accuracy of 88.17% and 94.6% for CFS subset and 83.69% and 95.39% for consistency subset respectively.

The main objective of this research work is detecting best subset of features by combining and assessing the performance of various classification algorithms for identifying phishing attack.

The main contributions of this work are as follows:

- We have derived best possible features and reduced the dimensionality of feature subset that can be used for phishing detection through the feature ranking using the combination of Random Forest algorithm, XGBoost algorithm and correlation matrix with heatmap.
- We have evaluated performance of the implemented classifiers and among them we have proposed the best hybrid classifier consisting of SVM, Decision Tree, Random Forest and XGBoost to attain higher accuracy.

Our proposed hybrid classifier will help the Internet users verify authentic websites, thereby mitigating the risk of phishing websites and ensuing secure online usage.

The rest of the paper is organized as follows. In Section II, few existing works in phishing website detection are explained. In Section III, all the features are explained briefly that we have used in our study. In Section IV, we have explained the system architecture. Feature selection techniques are explained in Section IV. In Section V, we have presented our results by combining and assessing the performance of several classifiers. Based on the evaluation, we presented our best classifier for detecting phishing attacks. Finally, we conclude the paper in Section VI.

II. Literature review

1) Blacklisting & Whitelisting based techniques: In blacklist-based approach, the requisitioned URL is contrasted with a pre-established phishing URLs. Whitelisting approach is completely opposite to the blacklist approach. In the whitelisting approach, the requested URL is compared with a preset authentic URLs. The drawback of these two approaches is that the blacklist or whitelist usually cannot cover every phishing or legitimate websites since a newly created website takes a significant time before being appended to the list.

Li et al. [4] made an assumption that a blacklist based anti-phishing toolbar is more accurate than a whitelist based one for identifying more phishing websites. authors used Anti-phishing IEPlug and Google Safe Browsing as whitelist and blacklist based anti-phishing toolbar. They found accuracy for both approach and suggested that both blacklist or whitelist can be used since they did not find any difference in toolbar.

2) Heuristics and Machine learning-based techniques: There are several techniques for machine learning like Support Vector Machine(SVM), Decision Tree, Random Forest, XGBoost, Artificial Neural Network and so on. Alswailem et al. [7] studied 36 features. Authors ignored irrelevant features and selected relevant 26 features. Random Forest classifier was chosen for classification to pursue high performance. Aminu et al. [9] worked on improving the existing methods by proposing a hybrid technique (Random Forest and XGBoost) algorithms. For ranking and selecting most relevant features, Random Forest was used. And XGBoost was utilized building the model. They collected dataset from UCI repository comprising of 11055 phishing websites. 97.2% accuracy was obtained using hybrid technique.

3) Content-based approach: Text-based contents are analysed to identify whether the website is phishing or legitimate. There are several techniques such as Deep MD5 Matching, phishDiff, TF-IDF etc. In [5], the authors proposed high-performance content-based phishing attack detection where for detecting malicious websites, a file matching algorithms is executed. Syntactical Fingerprinting algorithm compare structural components within files. This new algorithm gave low false positive rate.

4) Visual similarity-based techniques: In these techniques visual similarities between web pages are detected by extracting visual features. Chiew et al. [6] proposed a method where logo images were extracted to identify consistency between authentic and phishing websites via machine learning technique. SVM was used to classify logo and non-logo images.

III. Feature Set

We are determining whether a website is malicious or not based on its features. So we need to know clearly about those features. Basically there are four main features:

- Address bar based features (see Table I)
- Abnormal based features (see Table II)
- HTML and JavaScript based features (see Table III)
- Domain based features (see Table IV)

Address bar based feature has 12 sub-features, abnormal based features has 6 sub-features. HTML and JavaScript based feature has 5 sub-features and domain based feature has 7 sub-features. In Table I, II, III and IV, feature explanations are given [10]. In this section, we have consolidated total 30 features. After that, 23 best features are selected thorough ranking procedure.

IV. Proposed approach

Our proposed methodology is shown in Fig. 2. By investigating existing works, we collected our dataset with 30 features. After finding out a valid dataset, it is pre-processed using sampling for splitting the dataset into training and test dataset. Then, the dimensionality of feature subset is reduced and a new feature subset using vedis derifeature ranking procedure. After that, a hybrid classification algorithm is proposed by combining the concept of bagging and boosting.A chrome browser extension is also created for detecting phishing websites.

A. Data collection

We collected our dataset from UCI machine learning repository [11]. This dataset was also used by other works [11], [12]. The dataset comprises of 11055 phishing URLs with 30 features where 4898 URLs are legitimate and the remaining's

TABLE I
ADDRESS BAR BASED FEATURE

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

TABLE II
ABNORMAL BASED FEATURES

Feature Number	Feature Name	Feature Explanation
F12	Request URL	If 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.

TABLE III
HTML AND JAVASCRIPT BASED FEATURE

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.

are phishing URLs. Table I is presented to show the features and their possible values where -1 means phishing, 1 means legitimate and 0 means suspicious.

B. Sampling

We split our dataset into two parts: training and test dataset. While training dataset is used to fit an machine learning algorithm or model, test dataset comes up with unprejudiced appraisal of a final model fit on the training dataset. We used 75% for training and 25% for testing from our dataset consisting of 11055 data.

C. Feature Selection

Irrelevant features may decrease the performance of the model. For selecting the strong features, we used two techniques: feature selection by feature importance and correlation matrix with heatmap. We pointed out feature importance using XGBoost and Random Forest. Fig. 3 and 4 shows top 20 features for XGBoost and Random Forest, respectively.

In Fig. 3, X axis represents F-score and Y axis represents feature numbers. While implementing XGBoost algorithm, f6 attribute(Having subdomains) is getting more importance and is used more for making decision trees than other attributes.This is because, the more an attribute is used for making key decision, the higher it's relative importance. According to a report [3] using sub-domain registration services for launching a fake website has become a great practice. Phishers are fascinated by CO.CC domain because of it's minimum priced transactions. Again, f6 attribute is getting higher importance because subdomain services such as CO.CC domain are giving phishers an outstanding cover by providing unregulated service.

Fig. 4 shows the relative importance of different features

TABLE IV
DOMAIN BASED FEATURE

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.

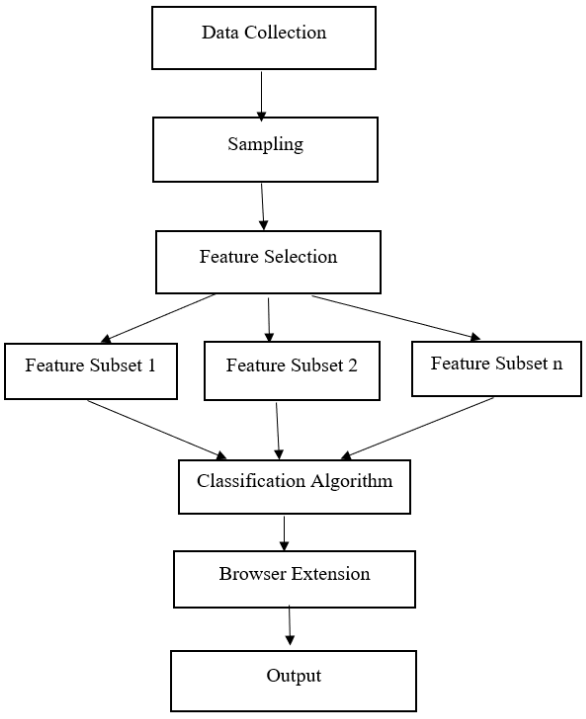

Fig. 2. Proposed system

where feature f7 (SSL final state) is the top feature. This is because if anyone enter his personal credentials without checking whether a website is authentic or not, it might be incepted by adversaries. So, before entering credentials, an user must check whether the website has encrypted connection or not. Most of the phishing websites do not use encrypted

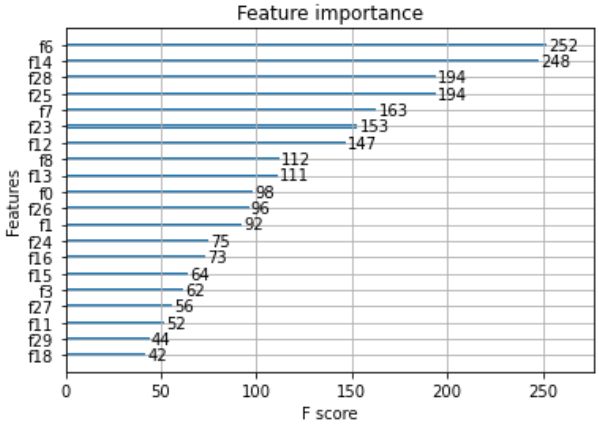

Fig. 3. Top 20 features using XGBoost

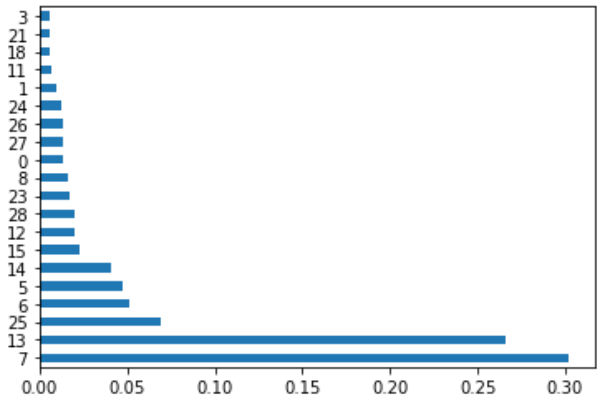

Fig. 4. Top 20 features for random forest

connection.

We also find out correlation matrix with heatmap. By using correlation matrix, we can find out highly correlated variables. Perfect negative correlation is indicated by -1, whereas +1 denotes perfect positive correlation between two variables. And 0 means no association between two variables. When the result becomes negative for some features then we have omitted those features because those features have a negative impact on the result.

Using these techniques, we created several subsets. Among them, we have proposed the best feature subset consisting of 23 features F0, F1, F3, F5, F6, F7, F8, F10, F11, F12, F13, F14, F15, F16, F20, F21, F23, F24, F25, F26, F27, F28, F29. Because other subsets do not provide better accuracy than this one. In this way, we reduced the dimensionality of feature subset. Table V represents accuracy for several feature subsets including our proposed feature subset also. In table V, first subset represents top 6 features using Random Forest feature selection technique. Second subset represents top 9 features using XGBoost feature selection technique.Third subset is chosen using correlation matrix with heatmap. We selected threshold value +0.1. In subset 4, we have added three more

TABLE V
ACCURACY FOR SEVERAL FEATURE SUBSETS USING PROPOSED HYBRID CLASSIFIER

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%

features (F0, F24, F29) contrasted with subset 3 since all positive result for corresponding features have a positive impact in correlation matrix. Here, our selected threshold value is +0.076. We have chosen F11 in subset 5 though the result becomes negative for this feature. Because F11 is ranked 17 when we use Random Forest and 18 for XGBoost feature selection technique. Last subset is our proposed one. we have added two more features (F8 and F28) in subset 6 than previous subset 5 as F8 and F28 are also important according to XGBoost and Random Forest feature selection technique.

D. Classification Algorithms

We have applied several classifiers for training, testing and evaluating the performance. Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), XGBoost and several hybrid classifiers such as RF + XGBoost, DT + XGBoost, DT + RF, DT + RF + XGBoost, SVM + DT + XGBoost,SVM + DT + RF, LR + DT + RF + XGBoost and SVM + DT + RF + XGBoost were applied.

E. Browser Extension

We have also created a browser extension. When the user enters a URL, the extension accepts the URL using the GET method and passes the same to the python code using the Java script of the extension. The python code forms an array by pulling out all the features from the URL. We then test this on the trained hybrid classifier consisting of SVM, DT, RF and XGBoost. We have tested our proposed system against some phishing urls for example paypal.de@secure-server.de/secure-environment and also against some legitimate urls for example https://www.phishing.org/ etc.

The screenshots of browser extension for detecting safe and phishing website are shown in Fig. 5 and Fig. 6, respectively.

V. PERFORMANCE EVALUATION

We have measured the effectiveness of our proposed system by the various performance metrics, such as Accuracy, Precision, Recall and F1-score which can be calculated using

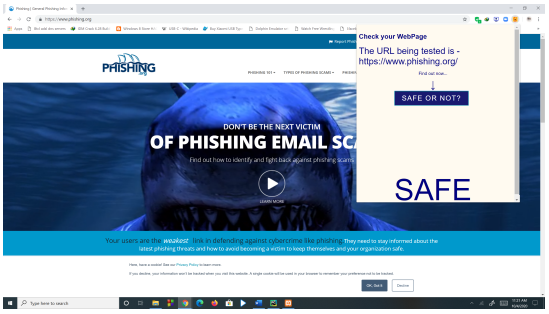

Fig. 5. Result for a safe website

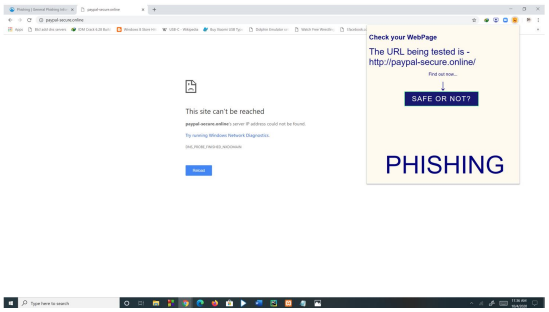

Fig. 6. Result for a phishing website

four terms: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

Table VI summarizes the results (accuracy, precision, recall, F1-score) for the phishing detection data set for all classifiers: Naive Bayes, LR, SVM, Decision Tree, Random Forest, XGBoost and combination of these classifiers for all 30 features. Our proposed hybrid classifier outperforms other classifiers by 36.69%, 5.18%, 5.11%, 1.59%, 0.63%, 0.89%, 0.34%, 1.44%, 1.23%, 0.29%, 0.37%, 0.34% and 0.14% respectively, for all 30 features. This is because we have merged the concept of bagging and boosting for classifiers which provides stability and fault-tolerance in contrast to traditional classification methods.

We have selected significant 23 features among 30 features. Various classifiers were applied on these 23 features. Such dimensionality reduction of the dataset has resulted in significant reduction in classification delay and improved the accuracy of the detection system. Table VII shows performance results such as accuracy, precision, recall, F1-score of

TABLE VI
PERFORMANCE RESULTS OF ALL CLASSIFIERS FOR 30 FEATURES

Classifier	Accuracy	Precision	Recall	F1-score
Naïve Bayes	61.87%	0.77	0.65	0.58
Logistic Regression	92.66%	0.93	0.92	0.93
SVM	92.73%	0.93	0.93	0.93
DT	96.16%	0.96	0.96	0.96
RF	97.10%	0.97	0.97	0.97
XGBoost	96.85%	0.97	0.97	0.97
RF and XGBoost	97.39%	0.97	0.97	0.97
DT and XGBoost	96.31%	0.96	0.96	0.96
DT and RF	96.52%	0.97	0.96	0.96
DT, RF and XGBoost	97.43%	0.98	0.97	0.97
SVM, DT and XGBoost	97.36%	0.97	0.97	0.97
SVM, DT and RF	97.39%	0.98	0.97	0.97
LR, DT, RF and XGBoost	97.58%	0.98	0.97	0.98
SVM, DT, RF and XGBoost	97.72%	0.98	0.98	0.98

various classifiers and their combinations for selected subset of 23 features. After assessing the performance, we found that proposed hybrid classifier performed better than others.

TABLE VII
PERFORMANCE RESULTS OF ALL CLASSIFIERS FOR PROPOSED FEATURES

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

A. Results summary

Table VIII shows comparison of our work with previous works using the same dataset [11] where our proposed method has achieved highest accuracy which is 98.28% by selecting minimum number of features and by reducing the dimensionality of feature subset. We have achieved better result due to use of the robust feature selection techniques and proposed hybrid classifier combining the concept of bagging and boosting.

VI. CONCLUSION

The number of phishing attacks has rapidly increased recently due to the rise in number of online transactions. People without much knowledge of the phishing sites face huge monetary losses due to the phishing attacks . In this paper, we have proposed a hybrid technique (SVM, DT, RF, XGBoost) for the selected features and reduced the dimensionality of feature subset to get better result. Feature importance has been computed through the use of XGBoost, Random Forest and also correlation matrix with heatmap has been generated for deriving the most important features. Results show that our system has achieved 98.28% accuracy in detecting phishing attack by analyzing the URLs of phishing website.

TABLE VIII
COMPARISON WITH PREVIOUS WORKS FOR THE SAME DATASET

	Proposed method	Accuracy	F1-score	Number of used 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

REFERENCES

[1] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3629–3654, 2017.

[2] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.

[3] "Apwg trends report," https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf.

[4] L. Li, E. Berki, M. Helenius, and S. Ovaska, "Towards a contingency approach with whitelist-and blacklist-based anti-phishing applications: what do usability tests indicate?" *Behaviour & Information Technology*, vol. 33, no. 11, pp. 1136–1147, 2014.

[5] B. Wardman, T. Stallings, G. Warner, and A. Skjellum, "High-performance content-based phishing attack detection," in *2011 eCrime Researchers Summit*. IEEE, 2011, pp. 1–9.

[6] K. L. Chiew, E. H. Chang, W. K. Tiong *et al.*, "Utilisation of website logo for phishing detection," *Computers & Security*, vol. 54, pp. 16–26, 2015.

[7] A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani, "Detecting phishing websites using machine learning," in *2nd International Conference on Computer Applications & Information Security (ICCAIS)*. Riyadh, Saudi Arabia: IEEE, 2019, pp. 1–6.

[8] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on url," in *IEEE Conference on Communications and Network Security (CNS)*. Florence, Italy: IEEE, 2015, pp. 769–770.

[9] H. Musa, B. Modi, I. A. Adamu, A. A. Aminu, H. Adamu, and Y. Ajiya, "A comparative analysis of different feature set on the performance of different algorithms in phishing website detection," *International Journal of Artificial Intelligence and Applications (IJAIA)*, vol. 10, no. 3, 2019.

[10] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Phishing websites features," *School of Computing and Engineering, University of Huddersfield*, 2015.

[11] A. A. A. Abdulrahman, A. Yahaya, and A. Maigari, "Detection of phishing websites using Random Forest and XGBoost algorithms," *International Journal of Pure and Applied Sciences*, vol. 2, no. 3, pp. 1–14, 2019.

[12] R. Das, M. Hossain, S. Islam, A. Siddiki *et al.*, "Learning a deep neural network for predicting phishing website," Ph.D. dissertation, Brac University, 2019.