IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13420

PHISHING ALERT USING MACHINE LEARNING

Mr. V. Ravikanth[1], Madimi Deekshitha[2], Palla Gnaneswar[3], Mallepogu Hari[4], Anumala Dinesh[5]

CSE, RGMCET, of JNTU Ananthapur, Nandyal, AP, India.[1-5]

Abstract: Phishing websites represent a significant threat to cyber security as they threaten the confidentiality, integrity and availability of both corporate and consumer data. These malicious sites often serve as an entry point for various cyber attacks. Despite extensive efforts by researchers over the years, effective detection of phishing sites remains a challenge. While some advanced solutions show promise, they often require extensive manual engineering of features and struggle to keep up with emerging phishing tactics.

Addressing this challenge requires strategies capable of automatically identifying phishing sites and quickly handling new, previously unseen attacks. One promising approach involves leveraging the wealth of data available on websites hosting these malicious activities. Machine learning is proving to be a powerful tool in this endeavor, offering a more automated and efficient approach compared to traditional methods.

In our research, we conducted a comprehensive literature review and proposed a new method for detecting phishing websites. This method involves extracting features from web pages and using machine learning algorithms for classification. Using a data set specifically designed for this purpose, we aim to develop a robust and adaptive system capable of accurately identifying phishing attempts, including zero-day attacks.

Through this work, we aim to improve cybersecurity measures by providing a reliable method for identifying phishing attempts, including new and previously unseen attacks. By leveraging the wealth of data available on phishing hosting websites, our approach aims to improve detection accuracy and reduce the risk of data breaches. Ultimately, our goal is to strengthen defenses against phishing attacks and protect sensitive information from unauthorized access.

Keywords: Phishing, Malicious, Cyber Security, Threat, Automation, Security

I. INTRODUCTION

The Internet has grown rapidly over the years and has billions of users worldwide. Unfortunately, this growth has also led to an increase in cybercrime, including phishing scams and malware attacks. Phishing is a sneaky way cybercriminals try to steal your personal information by pretending to be a trustworthy company or institution. They often send fake emails or redirect you to fake websites to get you to divulge sensitive information like passwords or credit card numbers. These attacks are a big problem, causing significant financial losses and threatening people's privacy. According to reports, phishing sites are on the rise, with thousands being exposed every quarter.

Detecting these phishing sites is essential to protect users from falling victim to these scams. Although there are some methods, such as using blacklists or analyzing the content of websites, they have their limitations. Researchers have been working on various techniques, including the use of machine learning, to improve detection accuracy.

One approach involves analyzing website features to determine whether they are phishing or legitimate. Some researchers have developed algorithms that can accurately classify web pages based on these features. By combining different classification algorithms, they aim to create a more robust detection system.

In our research, we focus on finding the best combination of features and classification algorithms for effective detection of phishing websites. Using advanced techniques such as Random Forest and XGBoost, we strive to achieve high accuracy in identifying phishing attacks.

Our goal is to help internet users stay safe online by providing them with a reliable way to detect phishing sites. By understanding how these detection methods work, users can better protect themselves from becoming a victim of cybercrime..

ISSN (O) 2278-1021, ISSN (P) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13420

II. RELATED WORK

1. Techniques based on blacklisting and whitelisting:
In the blacklist approach, we compare the requested URL against a list of known phishing URLs. Conversely, whitelisting involves comparing the requested URL against a list of trusted, authentic URLs. However, these methods have limitations as they may not cover all phishing or legitimate websites as it takes time to update the lists with newly created sites. Li et al. [4] compared the accuracy of blacklist-based and whitelist-based antiphishing tools and found that both can be effective. They used tools like Anti-phishing IEPlug and Google Safe Browsing.

2. Heuristics and techniques based on machine learning:
Machine learning techniques such as Support Vector Machine (SVM), Decision Tree, Random Forest, XGBoost and Artificial Neural Network are commonly used. by Alswail et al. [7] studied 36 features, selected 26 relevant ones and used Random Forest for classification. Amin et al. [9] proposed a hybrid technique combining Random Forest and XGBoost algorithms. They collected data from the UCI repository and achieved 97.2% accuracy.

3. Content-based approach:
Content-based approaches analyze text on a website to determine whether it is phishing or legitimate. Techniques such as Deep MD5 Matching, phishDiff and TF-IDF are used. In [5], a high-performance content-based phishing attack detection method using file comparison algorithms and syntactic fingerprinting to compare structural components was proposed. This approach yielded a low false positive rate.

4. Techniques based on visual similarity:
These techniques identify visual similarities between web pages by extracting visual elements. Chiew et al. [6] proposed a method that extracts logo images to compare authenticity between legitimate and phishing websites using machine learning.

5. In this study, Yong and his team developed a new method for detecting phishing websites by focusing on URL analysis, which proved to be an effective way to identify phishing attempts. Our approach involves the use of a neural network based on capsules, divided into different parts. One part removes the shallow characteristics of URLs, while the others create detailed representations of URLs and use shallow features to judge their legitimacy. The final result is calculated by combining the outputs of all parts. Through extensive testing on real-world data, we found that our system performs comparable to other advanced detection methods while being time-efficient.

6. For phishing detection, Vahid Shahrivari and colleagues used machine learning techniques, including logistic regression, KNN, Adaboost, SVM, ANN, and random forest. They found that the random forest algorithm provides good accuracy. Dr. G. Ravi Kumar used various machine learning methods, improved the results using NLP tools and achieved high accuracy using Support Vector Machine and pre-processed NLP data.

7. Amani Alswailem experimented with different machine learning models for phishing detection and found that random forest is the most accurate. Hossein et al. developed the "Fresh-Phish" framework for building machine learning datasets for phishing websites, achieving high accuracy using machine-learning classifiers.

8. Hussain et al. Creating a machine learning database for phishing websites has created a "Fresh Fish" framework to achieve high accuracy using machine learning classification.

9. X. Zhang proposed a phishing detection model based on embedding semantic words, semantic features and multidimensional statistical features to achieve successful operation. M. Aydin presented a versatile and straightforward framework for extracting features from phishing websites, using data from Phish Tank and Google, and using the feature selection method with WEKA.

Overall, this study demonstrates a different approach to phishing detection using machine learning and innovative methodologies to effectively combat online threats

III. EXISTING SYSTEM

The anti-phishing strategy aims to protect Internet users from becoming victims of phishing attacks. This strategy includes educating users and conducting technical support. In this article, we review the latest developments in technical defense against phishing attacks. Identifying phishing websites plays an important role in preventing attempts to steal user information. With the development of machine learning techniques, several machine-based approaches have been developed to improve the detection and prediction accuracy of phishing websites. The main goal of this paper is to find an effective method to prevent phishing attacks in real time.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⋟ Peer-reviewed & Refereed journal ⋟ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13420

In addition to technical protection, educating Internet users about phishing threats is essential to strengthening cyber security. By raising awareness of common phishing tactics and educating users to recognize and avoid suspicious emails, links and websites, we can empower people to protect themselves from falling victim to phishing attacks. Additionally, fostering a cybersecurity culture within organizations and promoting the use of multi-factor authentication and secure search practices can strengthen defenses against phishing attempts. Through a combination of technical measures and user education, we can work to create a safer online environment for all Internet users. Additionally, user awareness plays an important role in cyber security. Educating Internet users about common phishing tactics and recognizing and avoiding suspicious online behavior is an important step in improving online security. Additionally, promoting cybersecurity practices within organizations and supporting the use of multi-factor authentication can further strengthen defenses against phishing attempts. We strive to create a safer online environment for everyone by combining technical measures with user education. "

IV. PROPOSED SYSTEM

We've outlined our method in Figure 2. We began by studying previous research and gathering a dataset with 30 features. Once we had a good dataset, we split it into training and testing sets using sampling. Then, we reduced the number of features and created a new subset using a ranking method. Next, we developed a hybrid classification algorithm by combining bagging and boosting techniques. Additionally, we built a Chrome browser extension to help detect phishing Website.

Data Collection:
We gathered our dataset from the UCI machine learning repository. This dataset has been used in other studies as well. It contains 11,055 URLs, with 4,898 being legitimate and the rest being phishing URLs. The dataset includes 30 different features. In Table I, we show these features along with their possible values. In the table, a value of -1 indicates phishing, 1 indicates legitimate, and 0 indicates suspicious.

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

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ≍ Peer-reviewed & Refereed journal ≍ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13420

Sampling:

We divide our database into two parts: training set and test set. The training set is used to train a machine learning algorithm or model, while the test set is an unbiased evaluation of the final model trained on the training set. We used 75% of the dataset containing 11,055 data points for training and the remaining 25% for testing..

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.

Choose a feature:

It is important to choose the right features for our model, because irrelevant ones can lead to less accuracy. We use two methods to select robust features: correlation matrix with feature importance and heat map. We use XGBoost and Random Forest to determine which features are most important. Figures 3 and 4 show the top 20 features for each method.

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.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13420

In Figure 3, the X-axis represents the F-score and the Y-axis represents the feature score. We found that the attribute "Own subdomains" (f6) is very important for XGBoost. Because models are often used to make important decisions.

According to reports, it has become common to use subdomain registration services to create fake websites. Phishers are attracted to domains like CO.CC because they are cheap and easy to use, providing them for their malicious activities. That's why the "Own Subdomain" feature has become so important in our model.

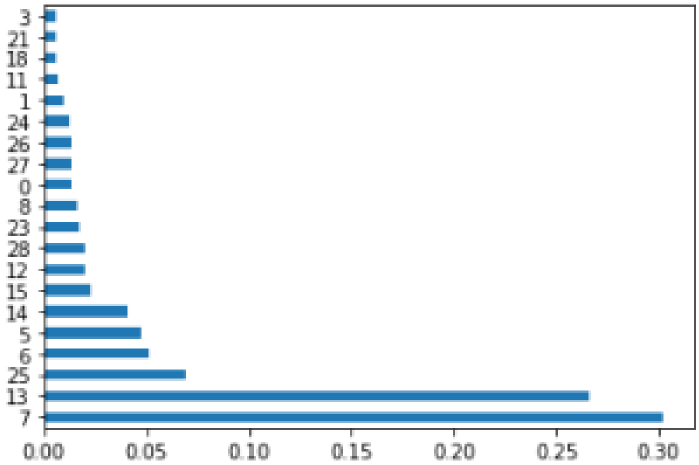


Fig. 4. Top 20 features for random forest

Figure 4 shows the importance of different features, "SSL Summary Status" (f7) being the most important. Because if someone enters their personal information on a website without verifying its authenticity, it can be intercepted by hackers. Therefore, users should check whether the website has an encrypted connection before entering sensitive information. Most phishing websites do not use encrypted connections, which makes it easier for attackers to steal data.

We also analyze the relationship between variables using heat maps. A correlation of -1 indicates perfect negative correlation, +1 indicates perfect positive correlation, and 0 indicates no correlation. Negatively correlated features were removed because they had a negative effect on the results.

Using this technique, we create several feature sets. The best subset of 23 features was selected based on higher accuracy compared to other subsets. We have reduced the dimensions of feature components to improve efficiency.

Table V shows the accuracies of several feature subsets including the proposed subset. Our department outperformed others in accuracy.

In addition, we have added more features to the proposed section based on the importance based on the feature selection method. This further improves the accuracy of our model. Through this technique, we refined the feature set to create a more effective fishing detection system.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⵛ Peer-reviewed & Refereed journal ⵛ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13420

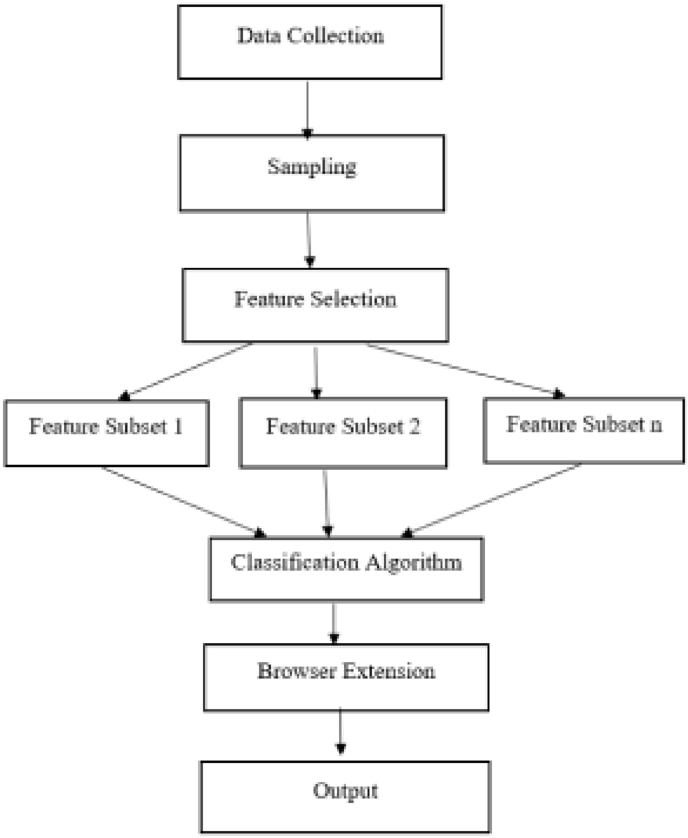


Fig. 2. Proposed system

V. ALGORITHM

We use different classifiers to train, test and evaluate the performance of our systems. These include naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), XGBoost and RF + XGBoost, DT + XGBoost and hybrid classifiers.

We have also developed browser extensions. If the user enters a URL, the extension passes the URL to our Python code using JavaScript. Python code extracts features from URLs and tests them using our hybrid classifier, which includes SVM, DT, RF, and XGBoost. We test our system against phishing and legitimate URLs such as "paypal.de@secure-server.de/secure-environment" and "https://www.phishing.org/".

We have developed a browser extension that runs when the user enters a URL. This extension takes a URL using the GET method and sends it to our Python code using JavaScript inside the extension. The Python code then extracts all the attributes from the URL and creates an array. We use this array to test our system and a hybrid classifier combining SVM, DT, RF, and XGBoost algorithms.

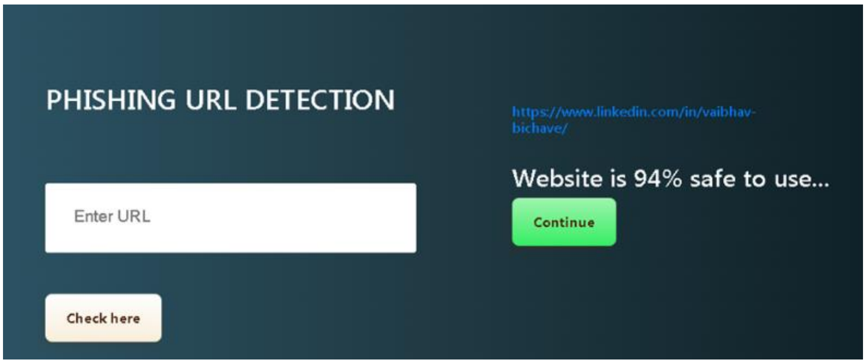
To evaluate our system, we tested it with several phishing URLs such as "paypal.de@secure-server.de/secure-environment" and legitimate URLs such as "https://www.phishing.org/". We also took screenshots of the browser extension, demonstrating its ability to detect legitimate and phishing websites. These screenshots are shown in Figures 5 and 6 respectively.

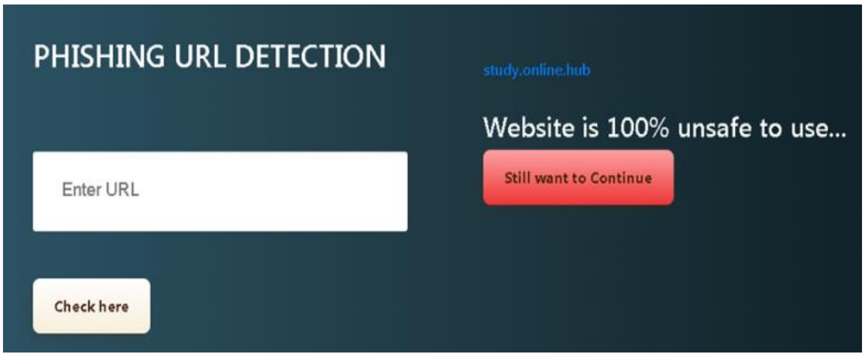
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⪤ Peer-reviewed & Refereed journal ⪤ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13420





Accuracy measures the overall accuracy of our system's predictions, while Precision represents the proportion of correctly identified positive cases out of all identified positive cases. On the other hand, only the proportion of true positive cases is identified out of all true positive cases. F1-score is the harmonic mean of Precision and recall, which provides a balance between the two dimensions.

In Table VI, we present the performance results (accuracy, precision, recall, F1-score) for different classifiers, including Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, XGBoost, and combinations of these classifiers. 30 features. Our proposed hybrid classifier outperforms others in all parameters. This improvement can be attributed to the fact that the combination of packing and reinforcement methods improves the stability and fault tolerance of our system compared to conventional methods.

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%

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⪤ Peer-reviewed & Refereed journal ⪤ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13420

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

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

We further refined our system by choosing the 23 most important features from the original 30. Applying a different classifier to this reduced feature set results in an increase in latency and classification accuracy as shown in Table VII.

Lowering these measurement levels increases the effectiveness and efficiency of our catch detection system. Table VII summarizes performance results such as precision, accuracy, recall and F1-score for this optimized set of features.

Accuracy = T P + T N T P + F P + T N + F N (1)

Precision = T P T P + F P (2)

Recall = T P T P + F N (3)

F1 − score = 2 ∗ Precision ∗ Recall Precision +Recall (4)

In Table VIII, our method outperforms the previous approach using the same dataset, reaching a maximum accuracy of 98.28%. This success is due to our strategy of selecting multiple features and reducing the dimensionality of the feature set. We attribute our superior results to the use of robust feature selection methods and our new hybrid classifier that combines bagging and boost elements.

VI. CONCLUSION

In conclusion, the increase in online transactions has led to a significant increase in phishing attacks, leading to significant financial losses for those unaware of these fraudulent sites. To solve this problem, our paper presents a hybrid method using SVM, decision tree (DT), random forest (RF), and XGBoost algorithm. By selecting key features and reducing feature dimensions, we aim to improve detection accuracy.

We use XGBoost and Random Forest algorithms to estimate feature importance, and generate a correlation matrix heatmap for feature detection. Our system showed promising results, achieving an incredible 98.28% accuracy in detecting phishing attacks by analyzing URLs associated with these malicious websites.

Overall, it offers a secure solution to combat phishing threats using advanced techniques to improve detection capabilities and reduce financial risks for online users**.**

REFERENCES

[1]. Gupta, B. B. et al.: This paper provides an overview of the current status and future challenges in combating phishing attacks. Published in 2017 in the journal Neural Computing and Applications, it provides information on strategies to combat phishing threats.
[2]. Abdelhamid, N. et al. APWG Trend Report: This document provides insight into trends in fishing attacks. It serves as a valuable resource for understanding the evolving phishing threat landscape and is accessible on the APWG website in 2020.
[3]. Lee, L. et al. A 2014 study published in the Journal of Behavioral and Information Technology examines the results of usability tests for anti-phishing software.
[4]. Wardman, B. et al.: Paper discusses high-performance content-based detection of phishing attacks. Presented at the 2011 eCrime Researchers Summit explores methods for detecting phishing attacks based on content analysis.
[5]. Chiew, K. L. et al.: This study examines the use of website logos for phishing detection. The 2015 edition of Computers & Security tested the effectiveness of using website logos as a feature to detect phishing websites.
[6]. Alswailem, A. et al.: The paper discusses the detection of phishing websites using machine learning techniques. A machine learning approach for phishing detection, presented at the 2nd International Conference on Computer Applications and Information Security 2019.
[7]. Aydin, M. et al.: This paper explores feature extraction and classification of phishing websites based on URL features. Presented at the 2015 IEEE Conference on Communications and Network Security, explores how to extract features from URLs to detect phishing.
[8]. Musa, H. et al. Published in 2019 in the International Journal of Artificial Intelligence and Applications, it evaluates the performance of different algorithms using different feature sets.
[9]. Muhammad, R. M. et al.: This report presents the characteristics of phishing websites. It provides insight into the characteristics of phishing websites developed by the School of Computing and Engineering at the University of Huddersfield in 2015.
[10]. Abdulrahman, A. A. A. et al. Published in 2019 in the International Journal of Pure and Applied Sciences, it evaluates the effectiveness of machine learning algorithms for phishing detection.
[11]. Das, R. et al.: This dissertation explores the use of deep neural networks to predict phishing websites. Presented at the University of Brac in 2019, it explores the use of deep learning techniques for website prediction.
[12]. Wahid Shahrivari and others. Various machine learning techniques such as logistic regression, KNN, Adaboost, SVM, ANN and random forest are used for fishing. Their research shows that the random forest algorithm offers better accuracy.
[13]. Dr G. Ravi Kumar uses various machine learning techniques and natural language processing (NLP) tools to detect phishing attacks. They achieved high accuracy by applying support vector machines to pre-processed data using NLP techniques.
[14]. X. Zhang proposed a phishing detection model based on mining semantic features, word input, semantic features and multidimensional statistical features on Chinese websites. The model uses AdaBoost, Bagging, Random Forest, and SMO algorithms to detect successful fishing using databases from DirectIndustry and China Anti Fishing Alliance online guides.
[15]. M. Aydin presented a versatile and straightforward framework for extracting features, data from Phish Tank, and real URLs from Google.C.