IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Effective Phishing Website Identification System with Superior Accuracy

[1]Mr. Omkar M Nooli, [2]Mr. Ritin M Kankanwadi, [3]Mr. Sachin Y Shinde,
[4]Mr. Ramesh B Pogatyanatti, [5]Mrs. Arpita Patil
[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor
[1]computer Science and Engineering,
[1]SG Balekundri Institute of Technology, Belagavi, India

Abstract: The rapid development of e-commerce, e-banking, and social organizations has exposed a large number of phishing attacks to cyber security systems. To support the viability of anti-phishing efforts, we present an advanced scientific show based on machine learning. In this image, six specific calculations are tuned: estimated recurrence, K- nearest neighbours, trusted Bayes, random forest, support vector machine, and extreme gradient boosting (XGBoost). Our study used a large dataset consisting of 58,000 genuine and 30,647 phishing websites, each with 112 characteristics. Improvements in highlight selection, including dataset tuning and removal of immutable highlights, have generally improved the execution process. Extensive testing of eight unique scenarios showed that all calculations reliably achieved over 93% accuracy, with XGBoost achieving the highest score. Accuracy is 99.2%, precision is 99.1%, recall is 99.4%, and specificity is 99.1%. Additionally, the Foremost Component Examination (PCA) technique was used to significantly reduce the XGBoost calculation execution time from 1500ms to 869ms. These findings validate the suitability of our strategy for both offline and real-time dispatch, highlighting its reasonable relevance in modern cyber security operations**.**

Index Terms - Phishing Detection, Machine Learning Algorithm, Feature selection, Dataset, Model Evaluation Random Forest, Support Vector Machine, Machine Extreme Gradient, Feature Engineering Performance Metrics.

I. INTRODUCTION

In numerous jurisdictions, unfettered access to information channels and public networks is recognized as an essential right. The rise of the internet has prompted considerable adaptations in the lifestyles of individuals. With the rapid increase in e- commerce and online consumer activities, phishing has surfaced as a prominent cyber security threat. Thus, the development of robust defense mechanisms for digital assets is imperative in contemporary society [1]. By submitting sensitive information on imitation websites designed to resemble genuine ones, users unknowingly provide their credit card information, passwords, and other personal data to malicious actors. These attacks primarily affect individuals who lack extensive knowledge of cyber security risks, as well as corporations and social networking platforms. In phishing operations, fraudsters leverage the close visual similarities between authentic and counterfeit websites to mislead users. Moreover, these fraudulent sites are frequently hosted on legitimate domains that are indexed by search engines such as Google or managed by recognized hosting providers. Additionally, they often advertise these sites on legitimate websites to draw a diverse range of individuals [2]. Phishing attacks have become previously concentrated on financial deception, contemporary phishing strategies frequently involve distributing malicious software and other detrimental payloads to various devices, including computers and mobile phones [5]. Roughly half of all cyber fraud impacting internet users comprises phishing attacks. In addition to the aforementioned concerns associated with phishing, it is imperative to highlight that phishing incidents have tripled in 2022 compared to 2020, underscoring the imperative for advancements in technological solutions within this domain. Phishing emerges as the primary threat to various sectors, notably e-commerce and internet banking, particularly in developing nations where infrastructure may not be sufficiently sophisticated to preemptively identify these threats [6] it the progress made in machine learning and deep learning methodologies, the achievement of high-accuracy algorithms has become feasible [13, 14, 15, 16] it is crucial that these algorithms accurately differentiate between phishing and legitimate websites. Classification algorithms are widely employed to categorize data into discrete classes or categories, leveraging their inherent differences or similarities [18.19] as it is apparent, classification algorithms are considered appropriate for tackling phishing challenges within the spectrum of machine learning methodologies. When addressing the issue of phishing website detection, crucial features are derived from a range of sources including URLs, DNS records, or datasets sourced from external repositories such as Phish tank or WHOIS .Four unique settings are used to detect phishing in online banking. Classify web pages using the Support Vector Machine (SVM) algorithm. .

1.1 Motivation

The primary goal of phishing attacks is data theft, accounting for 85% of incidents. The top 3 mostattacked industries using social engineering from Q3 2022 to Q3 2023 (inclusive) were governmentagencies (44%), military-industrial enterprises (19%), and organizations in the field of science and education (14%). Phishing sites are also places wherepeople are tricked and tricked into paying more. Therefore, it is important to protect resources from these attacks today. In addition to damaging infrastructure, phishing attacks also cause high costs for governments and individuals. Using a phishing website when shopping online can result in people being charged up to ten times the price the primary goal of phishing attacks is data theft, accounting for 85% of incidents. The top 3 mostattacked industries using social engineering from Q3 2022 to Q3 2023 (inclusive) were government agencies (44%), military-industrial enterprises (19%), and organizations in the field of science and education (14%). Phishing sites are also places where people are tricked and tricked into paying more. Therefore, it is important to protect resources from these attacks today. In addition to damaging infrastructure, phishing attacks also cause high costs for governments and individuals. Using a phishing website when shopping online can result in people being charged up to ten times the price. In addition tothe costs mentioned above, getting to the root of these problems is time-consuming for both people and governments. To solve this problem, we have developed a fast algorithm that can be used to identify phishing websites with a low error rate. Other comments are as follows. Chapter 2 presents a review of studies in the literature. Section 4 reviews the details of implementation, evaluation, and experimental results in depth, and Section 5 describes conclusions and future work.

2. RELATED WORK

Phishing website detection (PWD) algorithmsleverage supervised machine learning techniques, also known as classifiers, to map inputs to outputs through specific functions. Throughout the learning process, classifiers analyze various features (inputs) to predict the output (response), identifying legitimacy or potential for phishing.[1,2] distinctive website attributes to determine their In a study introduced by researchers[5,6], an approach for detecting phishing websites integrated avariety of hyperlink-specific features. Machine learning algorithms were trained on 12 categories of these features. Notably, this method operated exclusively on the client side, eliminating the needfor third-party involvement. Moreover, it was designed to be language-independent, capable of identifying websites written in any language study utilized Random Forest (RF), Support Vector Machine (SVM), and other algorithms, with RF achieving the highest accuracy rate of 98.52%.

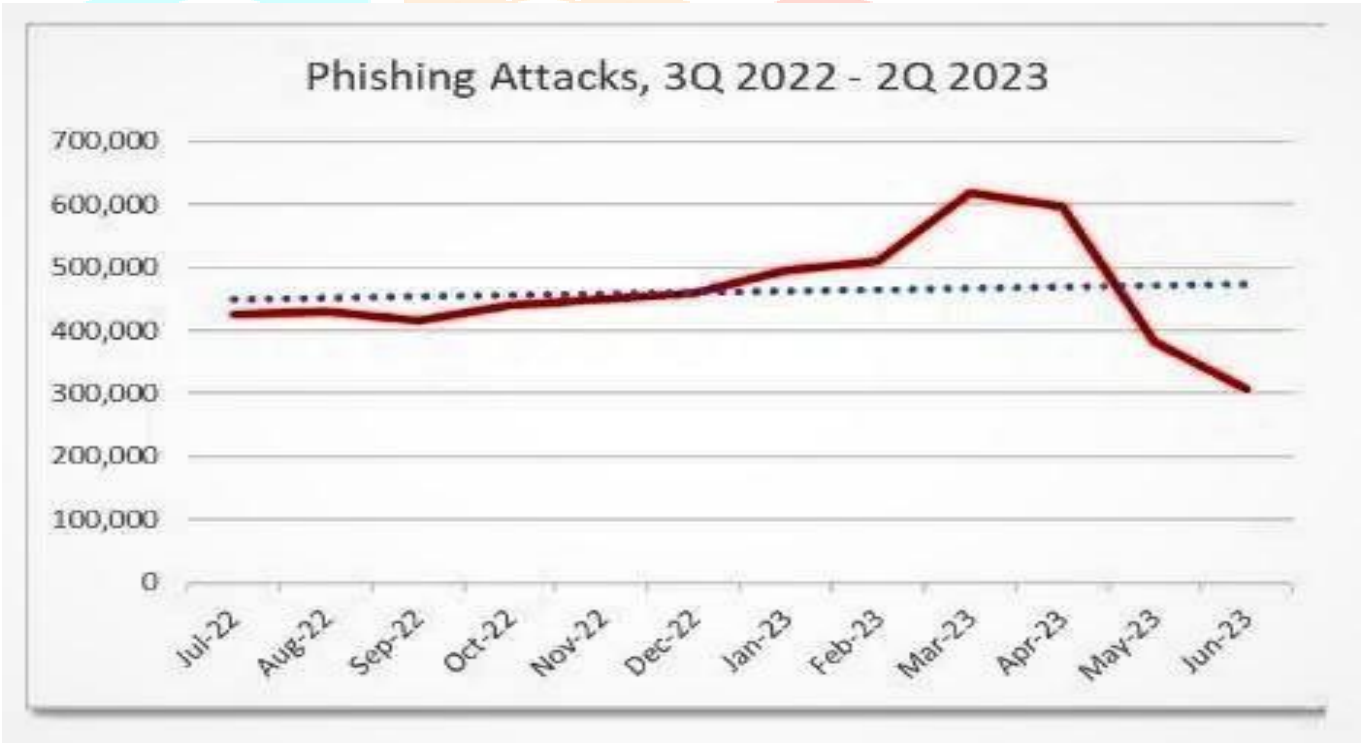

Fig 1: Phishing incidents from Q3 2022 to Q2 2023

Substantial dataset comprising over 73,500 instances extracted from URLs, covering both phishing and non-phishing URLs. The study highlighted the language independence of theirapproach, which relied on NLP techniques. Decision tree, Ad boost, and RF were among the algorithms employed, with K-star used on a subset of the datasetdue to computational constraints. Additionally, the study investigated hybrid features, combining NLP features with word vectors, resulting in significant enhancements in system performance. Findings indicated that NLP features outperformed other features in most machine learning algorithms, with hybrid features demonstrating considerable performance improvements, leading to an accuracy rate of approximately 98% achieved by RF. language processing (NLP). This system was trained The researchers sought to assess the effectiveness of various machine learning (ML) techniques andalgorithms in detecting phishing attacks, detailing their respective merits and limitations. They also developed an advanced phishing classifier system intended to exceed the capabilities of standard phishing detection methods. The investigation utilized numerical representations to compareconventional machine learning techniques, such as Random Forest (RF), K-Nearest Neighbors (KNN), Linear SVM classifier, and One-class SVM classifier. It incorporated wrapper-based feature selection that included URL data. Data for the study was aggregated from multiple sources through data mining approaches. Notably, Random Forest achieved the highest level of accuracy, registering anoverall efficacy of 96.87% [34].

The study examined phishing attacks through the application of URL and Domain Identity features, Abnormal Features, HTML and JavaScript Features, and Domain Features, collectively referred to as semantic features, to bolster phishing website detection. This methodology facilitated a more systematic and efficient classification process via the integration of semantic features. Employing 16machine learning algorithms and analyzing 10 semantic features, the research effectively detected phishing webpages within two distinct datasets. Comparative results indicated that the GradientBoosting Classifier and Random Forest Classifier were the most precise, achieving an accuracy rate of approximately 97%. Conversely, Gaussian Naive Bayes and Stochastic Gradient Descent demonstratedlower accuracy, with rates of 84% and 81%, respectively.[36].

The survey detailed in conducted a comprehensive examination of phishing attack methodologies anddefense mechanisms to overcome existing gaps in phishing detection. The research was organized into five distinct phases. The first phase analyzed the motivations, life cycles, and historical evolution of phishing, also addressing the various propagation methods used to spread phishing attacks. The subsequent phase developed a comprehensive taxonomy of phishing attack techniques relevant to both desktop and mobile platforms. In the fourth stage, the effectiveness of various phishing protection mechanisms was evaluated. The final section of the paper highlighted numerous performance challenges that developers encounter when combating this widespread threat. Moreover, the implications of phishing within new domains such as mobile platforms and online social networks were explored [41].

3. METHODLOGY

In the current digital landscape, the widespread integration of virtual platforms has spurred a surge in phishing attacks, particularly as banking systems and e-commerce entities transition to online frameworks. These malicious endeavors exact significant financialtolls on individual internet users and industries alike [1]. Consequently, there exists an urgent need for robust solutions capable of promptly and accurately identifying phishing attacks. This study presents a practical methodology for discerning phishing websites through the utilization of machine learning techniques. The methodology comprises severalsequential steps, including dataset exploration aided by visualization tools, feature selection, applicationof Imblearn's SMOTEENN method for dataset balancing, model development, testing, andsubsequent comparison of results, as delineated in Fig. 2.

The dataset forms the cornerstone for model training, as depicted in Fig. 2, with the model's output variable categorized into Training and Testing sets. An array of machine learning algorithms, encompassing Random Forest (RF), Support Vector Machine (SVM), and XGBoost, are employed for model training. Feature selection entails the identification and elimination of constant features to optimize model efficiency. The findings suggest that eliminating redundant data sources lacking pertinent information can streamline runtime and enhance overall accuracy. Additionally, dataset balance is achieved using the SMOTEENN method. The model is trained using 80% of the dataset, while the remaining 20% is reserved for testing to evaluate its efficacy. Furthermore, implementations across eight distinct scenarios are conducted, with further elucidation on the public dataset utilized in this studypending elaboration.

3.1. Data Series

Data acquisition stands as the foundational phasewithin the domain of data science endeavors, bearing processes. Our study involves a meticulous examination of a publicly accessible dataset to ascertain its efficacy in prognosticating forthcoming phishing attacks. The dataset under scrutiny, denotedas the Phishing Websites Dataset, was released in 2020 and encompasses a substantial corpus of 88,647 instances. These instances are characterized by 111 distinct features, alongside an annotated output column. A discernible class imbalance is evident, with the majority comprising 58,000 instances categorized as non-phishing (0), juxtaposed against 30,647 instances representing phishing activities (1). To rectify this imbalance, we employ a methodological approach aimed at achieving class equilibrium, expounded upon in subsequent sections. Furthermore, it is pertinent to note the absence of categorical variables and null entries within the dataset. Remarkably, all features within the dataset assume numerical data types, comprising 111 features of the int64 type and a singular feature of float64 type denoted as "time response."

. 3.2 Data Preparation

In our approach, we use normalization and model evaluation as pre-processing methods that are important fi=or design of machine learning (ML) driving applications. Feature scaling is important because machine learning algorithms will encounter problems when processing numerical input features at different scales. Given the raw quality of our data, with no significant or categorical differences, normalizationis a priori necessary. Establish a correlation throughthe modeling process by normalizing the data to followa Gaussian distribution characterized by mean μ andstandard deviation μ.

Standardization is a crucial preprocessing step in machine learning applications, aiming to scale numerical features to have a mean of zero and astandard deviation of one. This process ensures that input features are on a similar scale, which is advantageous for many machine learning algorithms.The standardization equation is represented as $y=x-\mu\sigma$

Where x denotes the original value to be standardized, μ represents the mean of the dataset, σ signifies the standard deviation of the dataset, and y is the resulting standardized value calculating the arithmetic mean of a dataset by adding all its values together and then dividing the sum by the total count of values

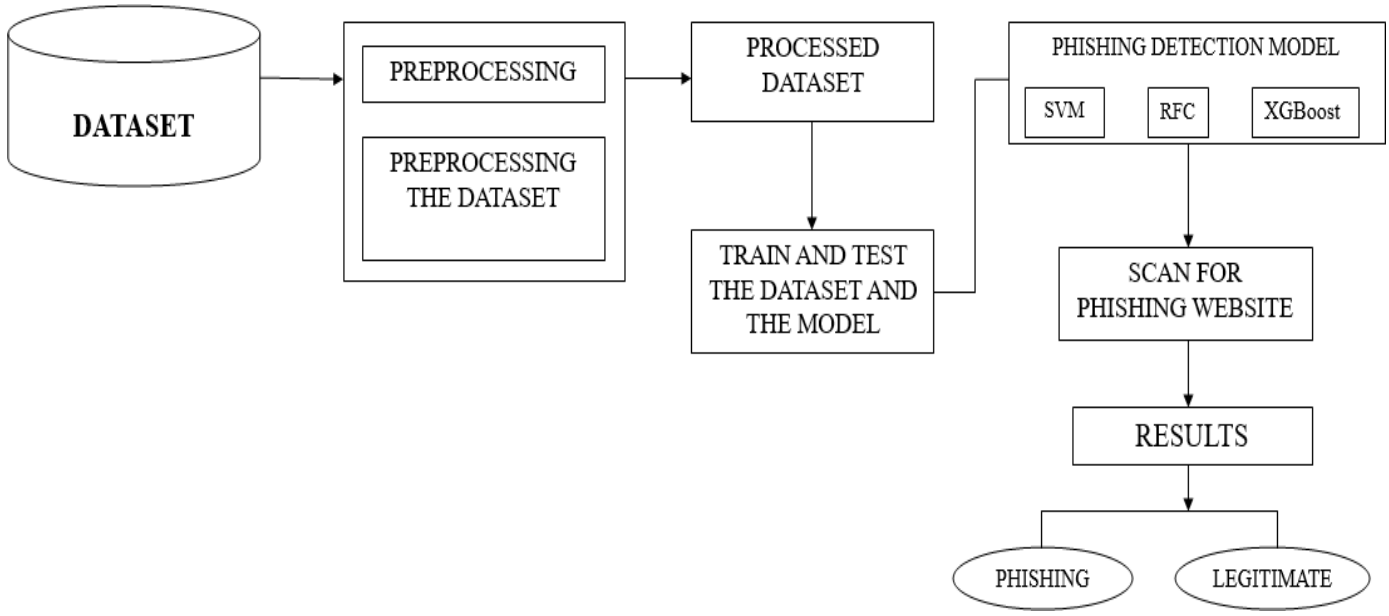

Fig. 2. Design the Future of Phishing Detection Visual Model Approach usual Model Approach

3.3 Feature Pruning

This means choosing the right one from different methods to create a good model. Feature selection can also reduce model complexity and computationalrequirements by reducing the number of variables and speeding up training. Therefore, we consider thefeature selection process in our scenario. It is often necessary to standardize data before analysis.

3.1.1 **Achieving class balance in the dataset**

In classification problems, class inequality occurs when some classes are more numerous than others, so the classification model is usually replaced by large classes and small classes are ignored. In some applications, the ratio of the small class to the large class can be very large, such as 1 to 100, 1000 to 10,000, or more. This issue occurs in many applications such as fraud/access detection, phishing detection, and diagnostics/monitoring. In the previous phase, these solutions included different types of recycling. Directional under sampling (reselecting samples for removal), oversampling to create new samples based on the data, and combinations of the above methods. Our study.selected smitten from the above method, which is acombination of oversampling and under sampling Tomsk's links and managed neighbors are two cleaning methods added to the pipeline after usingsmote oversampling to obtain clean surfaces.Smotetomek and smitten are two classes that have no parallels in the work in terms of integration ofbase models and different models.

Table 1: List of Data Parameters [part -1]

url	nb_star	tld_in_path
length_url	nb_colon	abnormal_subdomain
ip	nb_comma	tld_in_subdomain
length_hostnam	nb_semicolumn	prefix_suffix
nb_dots	nb_dollar	random_domain
nb_hyphens	nb_space	shortening_service
nb_at	nb_www	path_extension
nb_qm	nb_com	nb_redirection
nb_and	nb_dslash	nb_external_redirection
nb_or	http_in_path	nb_subdomains
nb_eq	https_token	length_words_raw
nb_underscore	ratio_digits_url	char_repeat
nb_tilde	ratio_digits_hostt	shortest_words_raw

3.4 Algorithmic Detail

In this study, we conducted a performance evaluation of 3 different classification methods, including Random Forest (RF), Support Vector Machine (SVM), and XGBoost, to determine whether an accurate machine learning-based method exists to detect phishing websites.

3.4.1 **Random Forest**

Random Forest (RF) is a powerful ensemble learningtechnique used for both classification and regression tasks. It builds multiple decision trees and mergestheir predictions to improve overall accuracy andcontrol over-fitting. Each tree in the forest is constructed using a random subset of features andtraining samples, ensuring diversity among the trees.This diversity makes Random Forest robust againstnoise and capable of handling large datasets withhigh dimensionality, making it a popular choice across various applications in machine learning.

3.4.2 **Support Vector Machine**

Support Vector Machine (SVM) is a powerful, versatile machine learning algorithm used for classification and regression tasks. It is particularly suitable for binary distribution problems. SVM is based on the idea of finding the hyper plane that best divides the data set into two groups.

3.4.3 ***XGBoost***

XGBoost (eXtreme Gradient Boosting) is a highly efficient and flexible gradient boosting framework that has gained popularity due to its performance in numerous machine learning competitions. It builds on the principles of gradient boosting machines, focusing on computational speed and modelperformance.

Table 2: List of Data Parameters [part-2]

longest_words_raw	ratio_nullHyperlinks	safe_anchor
longest_word_host	login_form	right_clic
statistical_report	ratio_extErrors	onmouseover
nb_hyperlinks	ratio_intErrors	domain_in_title
longest_word_path	ratio_intRedirection	empty_title
brand_in_path	ratio_extRedirection	domain_with_copyright
suspecious_tld	external_favicon	web_traffic
avg_word_path	links_in_tags	page_rank
brand_in_subdomain	nb_extCSS	domain_age
avg_words_raw	ratio_intMedia	dns_record
phish_hints	submit_email	status
avg_word_host	iframe	domain_registration_length
domain_in_brand	ratio_extMedia	whois_registered_domain
ratio_intHyperlinks	sfh	google_index
ratio_extHyperlinks	popup_window	

4. EXPERINMENT AND RESULT

In this study, we aim to develop and implement a high-performance handicapped device based on machine learning. Our reference takes into account publicly available data collected by Goran Vrbančič..[27]. It has 88,647 solids and 112 elements, making it very stable. Our tests were run on a Lenovo device equipped with a 2.5 GHz Intel Corei7 processor and 16 GB DDR4 RAM. The Python programming language was used along with several previouslydeveloped libraries to test the PWD process. Aftermeasuring the SMOTEENN dataset, the number of non-phishing events and phishing events are shown in Table 4.

4.1. Data Illustration

The general literature contains more and better information. While aggregate data contains only numbers, quality data contains other data. Appropriate and inappropriate information is classified according to the way it is collected, organized, stored and disclosed. Similarly, there are two types of data: continuous data and discrete data. It is usually represented using an integer or floating point number and can be any value in the range. As we mentioned above, in our study, all of the examples are numbers, integers and floating point numbers. Extracting useful information from numbers is more difficult than other types, but it can be really useful.

Table 3: Examining the Distribution of Phishing and Legitimate Instances across Imbalanced and Balanced Datasets

Dataset	Legitimate Samples	Phishing Samples	Total Instances
Imbalanced	58,000	30,647	88,647
Balanced	46,728	48,849	95,577
Total	104,728	79,496	184,224

4.2. Explored Scenarios

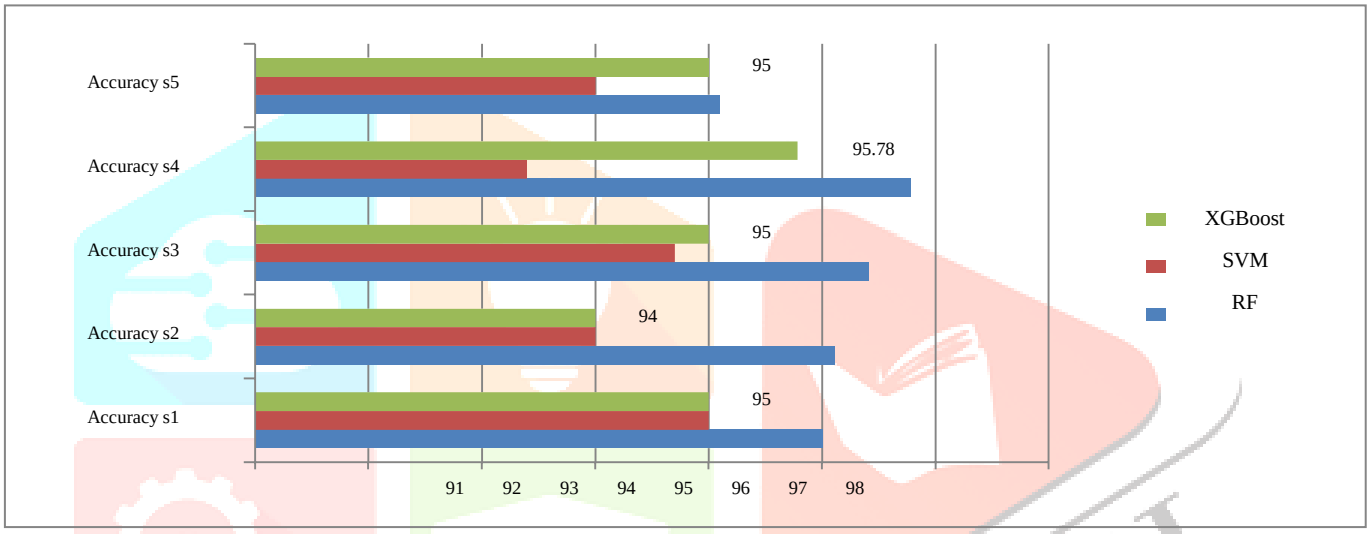

Fig 4: Accuracy (%) of all algorithms for the Explored scenario

5. CONCLUSION AND FUTURE WORK

As everything moves online, such as financial institutions and e-commerce, phishing attacks will also increase. As a result, businesses and Internet users suffered greatly. Therefore, it is necessary to develop strategic solutions that provide high accuracy and fast response time in detecting phishing attacks [63,64]. Come up. The result was a huge financial loss for businesses and Internet users. Therefore, there is a need to develop solutions that will provide accurate and rapid responses when investigating phishing attacks. Phishing network - website. These features are provided through hyperlinks contained in the source code of the website and other information collected from various sources. We developed classification algorithms such as RF, SVM, and many regular characters were removed in this study to reduce running time and improve model accuracy. Additionally, SMOTEENN is used to balance the data set to maximize it. When the data set was evaluated, 95,577 incidents were detected, 46,728 of which were legitimate and 48,849 of which were phishing. It contains 5 examples with different scenarios for all three algorithms. As a result, this model is very useful with an accuracy rate of over 93%, which is unique in the study. In our experiments, the XGBoost classifier outperformed other classifiers in phishing website classification with 96% accuracy, 96.4% recall, 96.2% F1 score, 97.1% accuracy, and 96.78% overall accuracy. Additionally, the dataset was duplicated and all fixed features were removed. The results show that neither principal component analysis (PCA) nor statistical analysis (LDA) could improve the overall accuracy of the model. The proposed method provides a unique phishing website detection strategy with high accuracy and low error rate. Strategies: (i) increasing the number of events in the next data, especially the phishing trick (ii) using various deep learning methods to detect differences and achieve higher accuracy, (iii) allowing the model to be used in real-life time while reducing the running time.

REFERENCES

[1] Badotra S, Sundas A. A systematic review on the security of E-commerce systems. Int J Appl Sci Eng 2021;18(2):1–9.

[2] Ansari MF, Sharma PK, Dash B. Prevention of phishing attacks using AI-based cybersecurity awareness training. Prevention 2022.

[3] Bhardwaj A, Al-Turjman F, Sapra V, Kumar M, Stephan.

[4] T. Privacy-aware detection framework to mitigate new- age phishing attacks. Comput Electr Eng 2021;96: 107546. Apwg.org/trends reports 2022

[5] Alshehri M, Abugabah A, Algarni A, Almotairi S. Character-level word encoding deep learning model for combating cyber threats in phishing URL detection. Comput Electr Eng 2022;100:107868

[6] Ghorbani M, Bahaghighat M, Xin Q, O¨ zen F. ConvLSTMConv network: a deep learning approach for sentiment analysis in cloud computing. J Cloud Comput 2020;9(1):1–2

[7] Hajikarimi A, Bahaghighat M. Optimum outlier detection in internet of things industries usingautoencoder. in: frontiers in nature-inspired industrial optimization. Springer; 2022. p. 77–92

[8] Khorasani F, Zanjireh MM, Bahaghighat M, Xin Q. A tradeoff between accuracy and speed for K-means seed determination. Comput Syst Sci Eng 2022;40(3):1085– 98.

[9] Rostami M, Bahaghighat M, Zanjireh MM. Bitcoin daily close price prediction using optimized grid searchmethod. Acta Univ Sapientiae Inform 2021;13(2):265–87

[10] Bahaghighat M, Abedini F, S'hoyan M, Molnar AJ. Vision inspection of bottle caps in drink factories using convolutional neural networks. In: Proceedings of the IEEE 15th international conference on intelligent computer communication and processing (ICCP). IEEE; 2019. p. 381–5

[11]Bahaghighat M, Abedini F, Xin Q, Zanjireh MM, Mirjalili S. Using machine learning and computer vision to estimate the angular velocity of wind turbines in smartgrids remotely. Energy Rep 2021;7:8561–76

[12] Shamseen A, Zanjireh MM, Bahaghighat M, Xin Q. Developing a parallel classifier for mining in big data sets. IIUM Eng J 2021;22(2):119–34

[13] phishtank.org 2022

[14] Minocha S, Singh B. A novel phishing detection system using binary modified equilibrium optimizer for feature selection. Comput Electr Eng 2022;98:107689

[15] Vrbanˇciˇc G, Fister I, Podgorelec V. Datasets for phishing websites detection. DataBrief 2020;33:106438