A Robust Model For Phishing Websites Detection Using Novel Generic Gradient Boost Classifier

Chirayu Agarwal
School of Computing
SRM Institute of Science and Technology,
Kattankulathur- 603203
ca3075@srmist.edu.in

Kavya A
School of Computing
SRM Institute of Science and Technology,
Kattankulathur- 603203
kz4129@srmist.edu.in

Kavisankar Leelasankar
Department of Computing Technologies
School of Computing
SRM Institute of Science and Technology,
Kattankulathur- 603203
kavisanl@srmist.edu.in

***Abstract*— The massive increase in online communication, social media interaction, and social network services, also increases the number of consequences occurring by it. Phishing attacks are getting increased in recent days because of the continuous people's engagement in the internet community also creating hints for hackers and spammers. The phishing attack enters the one's system premises through emails, social media, and other forms of electronic communication means. Phishing is a kind of resilient and robust act by the hacker to drag the maximum information about the user. The proposed system considers the detection of phishing attacks as a serious task and performed a comprehensive analysis of various algorithms to detect malicious activity on the websites in a rapid way. The proposed system utilizes the benefit of machine learning algorithms using the Novel Generic Gradient Boost Classifier algorithm (NGGB), which detects whether the website is Phishing or Legitimate. This study considers the phishing website dataset collected from Kaggle. After comparative analyses of various existing systems with the proposed system, it is observed that the proposed system detects phishing websites more accurately.**

Keywords— Phishing websites, Malware detection, Machine learning, Cyber-attacks, Email attacks.

INTRODUCTION

A massive amount of internet users are getting affected by malicious attacks every day. People often spend more time on the internet for activities such as bank transactions, entertainment, and business, academic and professional growth. Websites like Facebook, Twitter, and youtube are frequently accessed by users for many purposes. The frequently accessed websites have a record of the footprints of the users. The attractive notification, and subsequent links, redirect the user's attention towards malicious and unauthorized websites easily. [1] The hackers and spammers today, yield an attractive trap for users to enter into the malicious websites even through a single click. The more often, a user visits the malicious websites, there is a higher chance of vulnerable applications enter into the user system premises. Phishing websites create multiple traps for users, frequently accessing the social network websites [2].

A. Types of phishing

The basic element, where phishing starts is through continuous messages, communication via emails, social media, and other electronic means. A phisher may be described as a public resource such as social networking websites, entertainment providing links, websites, etc. these sources grab the user information such as job, name, email id as well as user interest and activities log. [3] This information is further processed by the hacker in the backend to create fake communication through emails, contact numbers, etc. The art of phishing websites is it looks similar to the original information that a company or trusted organization could provide. The impact of phishing started by creating trust between the user and the email sender. Most of the business communication emails and job confirmation emails nowadays are sent with phishing probes [4].

Email phishing is created via trustable websites or organization links generated by the hacker or spammer. The goal of the hacker is to get the user's trust and create a communication with them. Through trusted information and one or more communication follow-ups, the hacker grabs the user information, sometimes credential data without confirmed knowledge of the user. Trusted communications enable the user to provide all the data despite getting a certain benefit. These kinds of email-based phishing attacks are more frequently happening current days. On the other hand, email phishing attacks are sent via links. By clicking the certain link shared in the mail, the user becomes the victim of losing control of the current network and being trapped by the hacker. Within a certain frame of communication window, the user credentials are extracted rapidly by the hacker[5].

Spear Phishing attack includes, are critical ones, where the user loses money for the fake benefit offered by the spammer. By collecting the contact number of users, colleagues family members, and other friends, the scammer sends a fake link on benefits.

Whaling attacks were similar to phishing attacks in which the techniques often occur in senior employees when commonly connected with the public domain. The attackers craft the information with a high level of computing Technology. These attackers do not apply malicious tricks and

fake links. Instead, they leverage personalized information based on the research and victim. [2] Whaling attacks or commonly based on tax returns, the discovery of sensitive data, research revealing, etc.

Another type of phishing attack is called missing in walls fraudulent SMS messages that are involved in phone conversations. In this attack, the hacker scammed the credit card details from the company or bank and victimized the user's account. The complete details on transaction and credential information and with bankers are hacked by the scammers. In some cases, these kinds of phishing attacks can be involved by making the call and by using the phone keypad [7].

The Presented paper is formulated as a detailed literature background discussion in section II, followed by system design modeling, problem statement discussion, and tool selection in section III, further design methodology system architecture, detailed description of the present system is discussed in Section IV, concluded with future enhancement in terms of performance of the proposed methodology is discussed in the last section.

BACKGROUND STUDY

C. -Y. Li, et al., (2021) the author presented Android-based phishing attack detection in terms of privacy-preserving models. The system model incorporating the web keypads' role in phishing attack manipulation is discussed. The leakage of user credential data using Android mobile phones is discussed here. The phishing attack detection to be installed in Android devices with the privacy-preserving model is evaluated. Interfacing of third-party e attack detection application for Android app and the accuracy of 98% [7].

C. Esposito, et al., (2022) the author presented a system in which the trustworthiness of the social network users was analyzed. Analysis of legitimate users in the social networks is analyzed here for the prediction of phishing attack spammers involved in the community. The presence of phishing and legitimate users are classified with social media interactions and text analysis. The percentage system Utilizes Fuzzy logic for determining the Class [8].

S. He et al., (2021) the author presented a malicious URL detection that sometimes balances the data set. The presented system considers 600000 URLs for the analysis of malicious content present in the data set. The percentage system utilizes an XG boost algorithm for the classification of legitimate and malicious URLs present in the complete database. The presented XG booster algorithm compass with state of art approaches to existing algorithms in terms of feature-based analysis; data-based analysis, and biased classifiers [9].

Maroofi, et al., (2021) Data presented by email and disposing of methodology and analysis of legitimate users and malicious farmers entering into the network by using emails. The presented system considers configuring rules DMARC records and analyzing Accounting email logs. The system also verifies the email domains coming up from a verified resource [10].

B. Sun et al., (2021) Detection and prevention of cyber security attacks from emails and documentation. The presented here Deceptive documentation employed with targeted emails is analyzed in depth using the Privacy-preserving algorithm. The author presented a system in which the email logs documentation or completely verified in terms of the presence of legitimate and malicious users. The presented system achieves an accuracy of 97 % and is compared with state-of-the-art approaches [11].

S. Chen et al., (2021) the author presented a robust model to detect squatting attacks in Android systems. Using deep learning algorithms, phishing apps on the Android platforms are detected. The presented system detects the attack for long-term mobile users and Android devices that are prolonged connected with the internet resources. In the presented algorithm to conditions such as page confusion and logic, deception is considered for revealing the existing phishing attacks [12].

SYSTEM DESIGN

The serious impact of phishing attacks on websites and social media networks is discussed here. The drawback of wormhole websites, slow phishing, resilient communication window, and poor phishing detectors in recent implementations motivated the research towards the creation of a robust phishing website detection system. The presented system is implemented through python for the machine learning algorithms. Simulated via Anaconda3 software. Python is the high-level computing language in which the utilization of scientific computing libraries, numerical libraries, and machine learning models are helpful to create, and interprets detection mechanisms accurately. The presented system is focused on creating an efficient prediction system using a Novel generic gradient boost algorithm and also comparing the presented model with state-of-art approaches.

Data collection

The phishing data set is collected from the publicly available website in which text files and CSV files or provided with resources of input for building the detection models. a collection of many URLs around 11000+ websites are organized in the data set. This sample has thirty website parameters labeled with a class to identify phishing websites or not. The data set is formulated as + 1 and -1 for identifying the phishing label. The data set also acts as an input for project scoping and developing functional and non-functional requirements. The percentage system utilizes Python kit learning for classifying the input phishing data set into normal and phishing websites with a higher accuracy score.

METHODOLOGY

A. System Architecture

Fig. 1. Shows the system implementation architecture of the Phishing website detector. The input data set is pre-processed by removing the junk values not a number and infinite values by applying filters. The labeled data set is applied for the feature extraction process. The feature extraction process segregates the input data set into URL-based feature

extraction, HTML script-based feature extraction, JavaScript-based feature extraction, and domain-based feature extraction. The feature extracted text of the URL is further mapped concerning the labeled data set. On the other hand visualization of raw data and feature extracted data or mapped. The feature extracted data set is divided into training sets and testing sets. Initially to create a prediction model 70% of the training set is considered and 30% is considered for testing. Once the Novel generic gradient boosting classifier algorithm tests the given values with higher accuracy that is adaptable by the statistical methods then the real-time testing happens. The real-time testing data is collected randomly from the given data set. a part of the testing data is again switched to the created model for the analysis of prediction. despite classification results, the predicted results concerning the actual result are calculated and father considered for evaluation metrics.

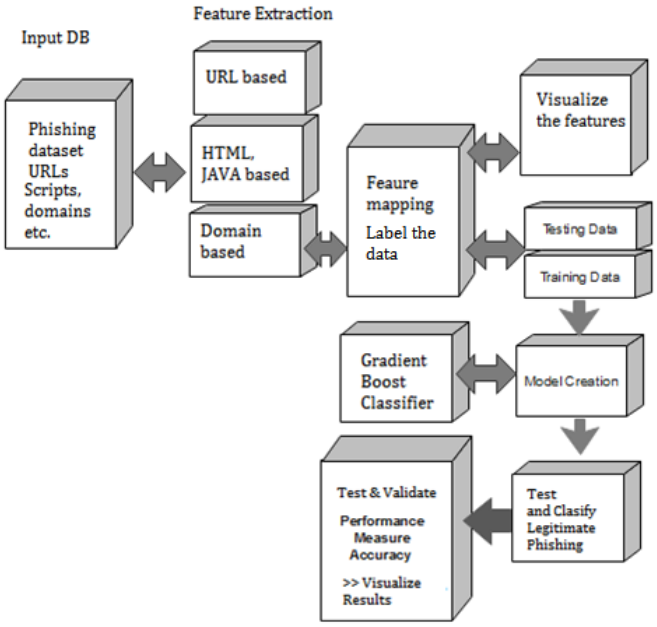

Fig. 1. The system architecture of Proposed Phishing website detector

B. Ensemble models

Rather than using a single indicator, Ensemble Learning sums up a number of indicators and preparations in the data and their consequences, often yielding a higher score than using a single model.Supporting is a unique type of Ensemble Learning method that combines a few weak students (indicators with terrible precision) into a strong student (a model with solid precision).This is accomplished by each model concentrating on the flaws of its forefathers.

C. Analysis

Many machine learning algorithms help attend the production process of phishing websites. Logistic regression acts as a binary classifier that determines the presence of legitimate users and phishing website users in the given data set. K nearest neighbor algorithm is one of the unsupervised models used for the detection of phishing websites that is discussed in the presented system. Support vector machines for robust machine learning algorithms are used to automatically segregate the pattern of training vectors in two different classes of HyperTerminal.

A naive Bayes classifier is used to classify the text models concerning the relativity between the data. The exploratory data analysis model uses a naive Bayes classifier for classifying the text based on weightage by a softer input text. A decision tree algorithm is nothing but the combination of smaller decisions that determine the training pattern and testing pattern correlations. A random forest algorithm is the probability distribution of input data in a scattered manner in which the relative correlated data alone is considered for decision making and finally evaluates the highly correlated data pattern as predicted results. A random forest algorithm is the combination of small decision-making algorithms evolved together to find the final decision.

CatBoost algorithm is one of the recently used techniques that determine the correlated factors based on category-based analysis. The input data set is categorized for the training data set and the statistical values such as mean median maximum or measured. Based on the correlated features between the training data set and testing data set the final results are evaluated. A multilayer perceptron is considered the robust neural network architecture used to make the statistical relativity between the training data and the testing data. The performance measured or evaluated using cross-entropy methodology and the layers of multilayer multilevel perceptron determine the in-depth Analysis of data patterns. Higher the relativity present with the data the higher the correlation would be.

From the analysis of various machine learning algorithms, the Novel generic gradient boosted classifier used for the analysis of phishing website detection, and from the results obtained it is considered that the gradient boost classifier outperforms comparing with other states of our approach models.

Feature importance

The given dataset is comprised of various attributes that are framed within the URL links. Some of the impacted features are, "UsingIP", which determines the current IP address in usage. 'LongURL' determines the URL length as per the recorded information. 'ShortURL' determines the content of the website, sometimes the URL resembles a malicious website, hence it is considered an important feature to process. 'Symbol@ represents the unique special characters present in the link. 'Redirecting// represents the control transfer to other third-party links. 'PrefixSuffix is considered to check if any malicious content is present. 'SubDomains', and 'HTTPS' both are helpful to check only the authorized website is getting connected. HTTPSDomainURL determines the domain URL under test. The complete HTTP link resembles. 'RequestURL' is the connectivity of the subdomain, with the main URL link connected live. 'WebsiteTraffic' determines the hint of website traffic as a variable. 'PageRank' determines the information about the website, the google rank in the form of a constant. Other important features in the website analysis are 'WebsiteForwarding', 'StatusBarCust', 'DisableRightClick',

'UsingPopupWindow', 'IframeRedirection', 'AgeofDomain', 'DNSRecording', 'PageRank', 'GoogleIndex', 'LinksPointingToPage', 'StatsReport', 'class' etc.

D. Implementation Summary using NGGB classifier

In Gradient Boosting, every indicator attempts to develop its ancestor by lessening the blunders. However, the intriguing idea behind Gradient Boosting is that instead of fitting an indicator to the information at each focus, it fits another indicator to the past indicator's residual faults.

- For each example in the preparation set, it works out the residuals for that case, or, as such, the noticed worthless the anticipated worth.
- Whenever it does this, it constructs a new Decision Tree that attempts to predict the recently determined residuals.

To begin predicting on the dataset, we must first determine the log of the chances of the objective component i.e. class label. This is usually the ratio of the number of True values (values equal to 1) to the number of False values (values equivalent to 0).It's 1 and -1 for legitimate and phishing in our scenario. Log (odds) = log (4/2) = 0.7 for six occurrences out of which four phishing websites and two trustworthy websites.This is the base estimator.Use a logistic function to turn the values into probabilities and make predictions once we know the logarithms (odds).If the logarithmic (odds) value is 0.7, the logistic function is also around 0.7, as in the previous example. The algorithm predicts 0.7 as the base estimate for each instance. The formula for converting logarithms (odds) to probabilities is:

$$P = \frac{e * Log(odd_{num})}{(1 + e * Log(odd_{num}))}$$

Eq. 1. The formula for finding probability

The prediction formula is base log odds + (learning rate * predicted residual) for each instance of the training set.The learning rate is a hyperparameter that is used to determine each tree's contribution and to sacrifice bias in order to increase variance.To avoid data overfitting, multiply that number by the predicted value.We need to convert the logarithmic (odds) forecast to a probability using the prior formula for turning the logarithmic (odds) value to a probability.

E. Evaluation metrics

Using the NGGB classifier, the system needs to be developed to find the presence of Phishing with reduced processing time. Ultimately the system performance needs to be evaluated. Using statistical measures, the True positive rate decides how much-anticipated outcomes are the same as the normal outcomes. True Negative rate decides the way that true data is anticipated correctly on misleading and expressed as True data. In case, more negative results came expressed output is negative. The false-positive rate determines the occurrence of the true value for false data. The false-negative rate determines the true value for false data. Accuracy is calculated despite predicted results vs. trained results by the formula below.

$$Accuracy = \frac{Number\ of\ Correctly\ predicted\ Samples}{Total\ population\ of\ the\ Data} X\ 100$$

Eq. 2. The formula for finding accuracy

RESULTS AND DISCUSSIONS

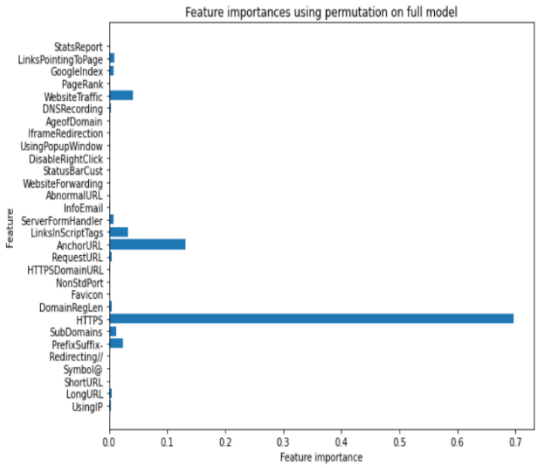

Fig. 2. Feature importance using permutation on full model

Fig. 2. Shows the feature importance using permutation on full model. From the figure, it is depicted that website traffic, Anchor URL, Links in the Script Tags, HTTPS are considered as most impacted features among all the features from the dataset.

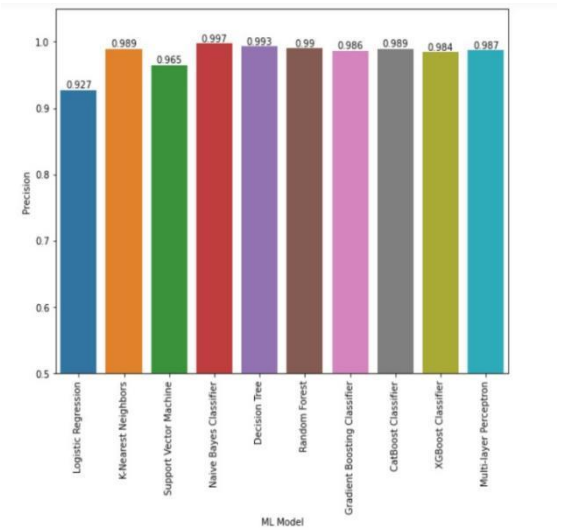

Fig. 3. Comparison of Precision values of Existing and proposed systems

Fig. 3. Shows the graphical representation of Precision value in which Novel generic gradient boosted classifier achieved 0.986 with respect to other existing machine learning classifiers.

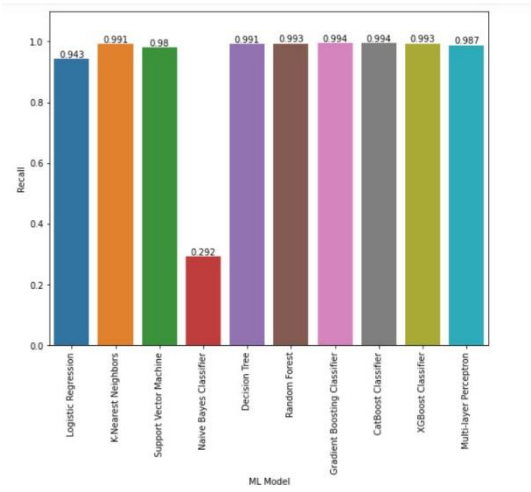

Fig. 4. Comparison of Recall values of Existing and proposed systems

Fig. 4. Shows the graphical representation of Recall value in which gradient boosted classifier achieved 0.994 compared to other existing machine learning classifiers.

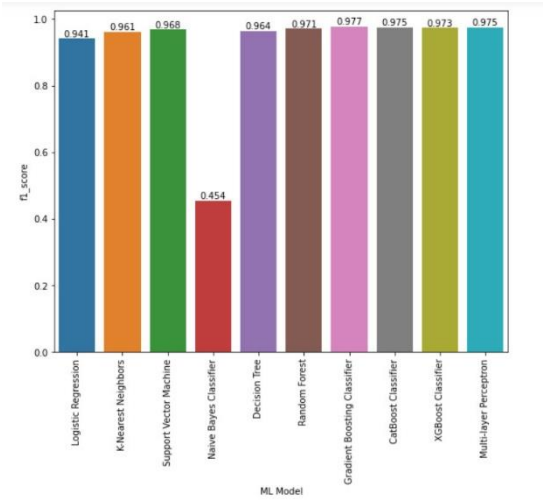

Fig. 5. Comparison of F1Score values of Existing and proposed systems

Fig. 5. Shows the graphical representation of F1Score value in which gradient boosted classifier achieved 0.977 compared to other existing machine learning classifiers.

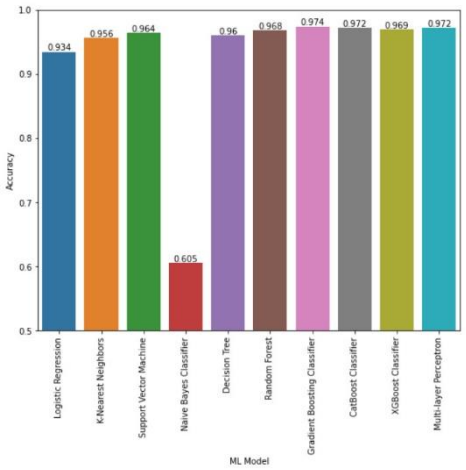

Fig. 6. Comparison of Accuracy values of Existing and proposed systems

Fig. 6. Shows the graphical representation of Accuracy value in which gradient boosted classifier achieved 0.974 compared to other existing machine learning algorithms. The proposed system uses decision trees, increased learning rate and max depth which helps to provide better accuracy.

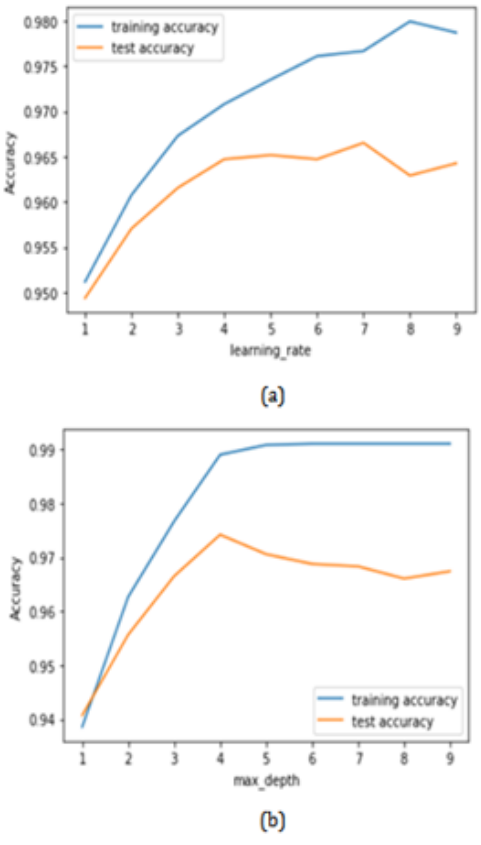

Fig. 7 Prediction Accuracy (a) Learning rate vs Accuracy (b)Maximum depth vs Accuracy

Fig. 7. Shows the obtained prediction accuracy at with respect to learning rate and, the simulation of training accuracy vs. testing accuracy with respect to maximum depth is shown.

CONCLUSION

A novel implementation of a phishing website detection system is developed here. The present system considers a sufficient data set from a specific website and considers the data set URL for analysis. Here the feature extraction involves URL based feature extraction process, a script-based feature extraction process, and a domain-based feature extraction process. The feature extracted data or split it up into training data and testing data to create a model using gradient boosted classifier (NGGB) algorithm.Logistic regression, K nearest neighbour, Support Vector Machine, Naive Bayes classifier, decision tree, random forest algorithm, CatBoost classifier, XG boost classifier, and multi-level perceptron are some of the machine learning techniques that are compared with the proposed algorithm. From the presented analysis the Novel generic gradient boost classifier achieves an accuracy of ~98%. Further, the present system is improved by developing deep learning algorithms and transfer learning approaches to improve the accuracy and in-depth analysis of phishing websites.

REFERENCES

1. C. Tankard, ``Advanced persistent threats and how to monitor and deter them," Netw. Secur., vol. 2011, no. 8, pp. 16_19, Aug. 2011.
   Gha_r and V. Prenosil, ``Advanced persistent threat attack detection: An overview," Int. J. Adv. Comput. Netw. Secur., vol. 4, no. 4, pp. 50_54, 2014.
   Advanced Persistent Threat. Accessed: Apr. 6, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Advanced_persistent_threat
2. W. Dai, M. Qiu, L. Qiu, L. Chen, and A. Wu, ``Who moved my data? Privacy protection in smartphones," IEEE Commun. Mag., vol. 55, no. 1, pp. 20_25, Jan. 2017.
3. Spear Phishing. Accessed: Apr. 6, 2021. [Online]. Available: https://www.kaspersky.com/resource-center/de_nitions/spear-phishing
4. S. Le Blond, C. Gilbert, U. Upadhyay, M. G. Rodriguez, and D. Choffnes, ``A broad view of the ecosystem of socially engineered exploit documents," in Proc. Netw. Distrib. Syst. Secur. Symp., 2017, pp. 1_15.
5. C. -Y. Li, H. -Y. Wang, W. -C. Wang and C. -Y. Huang, "Privacy Leakage and Protection of InputConnection Interface in Android," in IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3309-3323, Sept. 2021, doi: 10.1109/TNSM.2021.3077010.
6. C. Esposito, V. Moscato and G. Sperlí, "Trustworthiness Assessment of Users in Social Reviewing Systems," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 1, pp. 151-165, Jan. 2022, doi: 10.1109/TSMC.2020.3049082.
7. S. He, B. Li, H. Peng, J. Xin and E. Zhang, "An Effective Cost-Sensitive XGBoost Method for Malicious URLs Detection in Imbalanced Dataset," in IEEE Access, vol. 9, pp. 93089-93096, 2021, doi: 10.1109/ACCESS.2021.3093094.
8. Maroofi, M. Korczynski, A. Hölzel and A. Duda, "Adoption of Email Anti-Spoofing Schemes: A Large Scale Analysis," in IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3184-3196, Sept. 2021, doi: 10.1109/TNSM.2021.3065422..2019.2956035.
9. B. Sun et al., "Leveraging Machine Learning Techniques to Identify Deceptive Decoy Documents Associated With Targeted Email Attacks," in IEEE Access, vol. 9, pp. 87962-87971, 2021, doi: 10.1109/ACCESS.2021.3082000.
10. S. Chen, L. Fan, C. Chen, M. Xue, Y. Liu and L. Xu, "GUI-Squatting Attack: Automated Generation of Android Phishing Apps," in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 6, pp. 2551-2568, 1 Nov.-Dec. 2021, doi: 10.1109/TDSCS.
11. Z. Guo, J. -H. Cho, I. -R. Chen, S. Sengupta, M. Hong and T. Mitra, "Online Social Deception and Its Countermeasures: A Survey," in IEEE Access, vol. 9, pp. 1770-1806, 2021, doi: 10.1109/ACCESS.2020.3047337.
12. S. Z. Nizamani, S. R. Hassan, R. A. Shaikh, E. A. Abozinadah and R. Mehmood, "A Novel Hybrid Textual-Graphical Authentication Scheme With Better Security, Memorability, and Usability," in IEEE Access, vol. 9, pp. 51294-51312, 2021, doi: 10.1109/ACCESS.2021.3069164..