PHISHING DETECTION USING MACHINE LEARNING TECHNIQUES

T.N. Ranganadham, Assistant Professor in Department of CSE, Annamacharya Institute Of Technology And Sciences (Autonomous), Rajampet, Andhra Pradesh, India-516126

S. Harshitha[2], K. Jahnavi[3], K. Jyothirmai[4], T. Guna Harshitha[5] Department of Computer Science And Engineering, Annamacharya Institute Of Technology And Sciences (Autonomous), Rajampet, Andhra Pradesh, India-516126

Abstract

The trend of shifting practically all real-world operations to the cyberworld has been developing in recent years due to the increased use of mobile devices. Due to the Internet's anonymous nature, even though it simplifies our daily lives, it also leads to numerous security breaches. The majority of attacks can be avoided by using firewall and antivirus software. However, skilled attackers aim to exploit computer users' vulnerabilities by sending them phoney websites. To steal sensitive information like user names, passwords, bank account numbers, credit card details, and more, these pages mimic well-known social media, e-commerce, banking, and other websites. A number of solutions, including a blacklist, rule-based detection, anomaly-based detection, etc., but they are not giving the accurate results. This paper will discuss the machine learning algorithms such as SVM, gradient boosting, Random Forest, XG boost and naive bayes by applying all these algorithms on the dataset and the best algorithm having the best precision and accuracy is selected for the phishing website detection. This work can provide more effective defenses for phishing attacks of the future.

I. INTRODUCTION

We do the majority of our work on digital platforms in our daily lives. Our personal and professional lives are made easier by using computers and the internet in numerous ways. It enables us to finish our transactions and operations rapidly in a variety of industries, including banking, aviation, health, education, communication, research, and other technical fields as well as the arts and public services. Since the advent of mobile and wireless technologies, users that require access to a local network can now quickly and simply connect to the Internet from any location. Despite the fact that this circumstance is really convenient, it has exposed major information security flaws. Users in cyberspace must now take precautions to protect themselves from such threats.

Cybercriminals, pirates, non-malicious (white-capped) attackers, and hacktivists are only a few examples of those that carry out attacks [1]. The objective is to get access to the computer, the data it holds, or to collect personal data in various methods. Morris Worm, an internet worm, was used to launch the attacks, which have continued to this day. Fraud, forgery, coercion, shakedowns, hacking, service blocking, virus software, illicit digital materials, and social engineering are the major targets of these attacks [2]. Attackers try to obtain a lot of data and/or money by targeting a broad spectrum of users. Kaspersky's statistics indicates that, depending on the size of the attack, the average cost of an attack in 2021 will be close to $5 billion. The average cost of an assault in 2019 is between $ 108K and $ 1.4 billion, according to Kaspersky's research, depending on the scale of the attack. A further $ 124 billion is spent annually on goods and services related to global security [3].

Among these assaults, "phishing attacks" are the most pervasive and important. The use of email or other social networking contact channels by cybercriminals in this kind of assault is particularly prevalent. The main aim is to steal the sensitive information like user personal information or the bussiness information. Every attacker choosing the phishing because it is simple and there is no need of virus software for it. Attackers trick people into becoming victims by making it appear as though the post came from a reputable organisation, such a bank, an online store, or another comparable entity. They attempt to access their sensitive data in this way [4]. After exploiting this information, attackers get access to their victims' accounts. As a result, it results in financial loss as well as intangible losses.

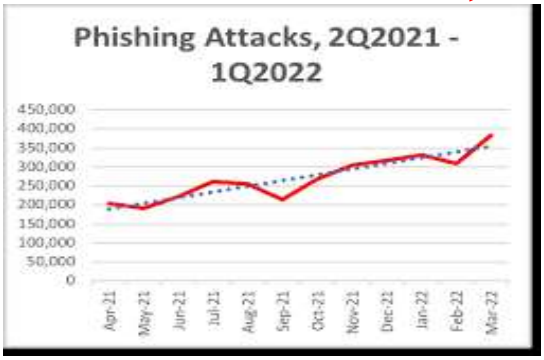


Figure 1. Phishing Attacks Statistics [6]

Since the previous decade, phishing assaults have become more sophisticated in how they reach their target users. This technique was used in the 1990s as an algorithm-based technique, in the early 2000s as an email-based technique, later as Domain Spoofing, and most recently via HTTPs. The cost and impact of the attacks on the users have been substantial due to the extent of the mass attacks in recent years. In 2019, phishing assaults that resulted in a data breach cost an average of $ 3.86 million, and business email compromise (BEC) incidents cost an estimated $ 12 billion. Furthermore, it is well-known that 15% of those who are attacked are at least two targets [5]. In light of this conclusion, it can be claimed that phishing assaults will still be conducted in the next years.

The Anti Phishing Working Group's (APWG) periodic reports serve as an important manual for researchers in this area. The largest monthly total in APWG's reporting history, 384,291 attacks were recorded in March 2022, according to the report. APWG recorded 1,025,968 phishing attempts in total during the first quarter of 2022. For the first time ever, the quarterly total for phishing incidents has exceeded one million, and this quarter was the worst APWG has ever recorded in terms of phishing incidents. In the final three months of 2021, there were 888,585 attacks, which set the previous record. Since early 2020, when APWG was tracking between 68,000 and 94,000 attacks per month, the number of phishing attacks had more than tripled. However, it can be argued that not only will phishing attacks continue, but there will also be more different attack types this year than there was the year before.

This increase suggests that attackers are employing phishing attacks more frequently. since they are simple to design. As seen in Figure 2, phishing attacks are dependent on the attacker building a false website. A phisher creates bogus websites first, along with a phishing kit. The prepared email is then used to direct the victim to the fake website. The victim visits the bogus website by clicking on the URL because they think the email and URL are secure. Following this, the victim's login information is obtained by the phishing kit and sent to the phisher. Finally, using the victim's login information, the scammer creates a bogus earnings account on a reputable website. The aesthetics on these websites are frequently strikingly similar or even identical. The intended recipient is forwarded to this phoney website in an email that appears to have come from a reliable source. The victim uses an email account that she/he trusts to access the website at the appropriate URL and enters the data that the attacker is trying to get. The essential information is obtained by the attacker, who then applies it to the actual system.

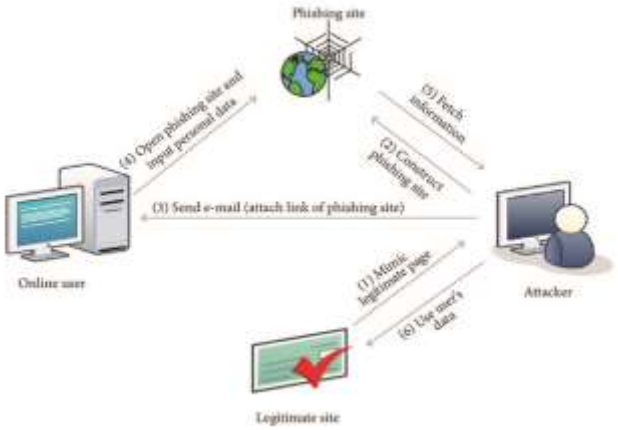


Figure 2. Phishing Attack Life Cycle

The attacker obtains data and/or money in this manner. For the victim to believe, trustworthy email contents are manufactured in a variety of ways. Prior to this, people would send each other emails with unlikely offers, urgent texts, links, or attachments from unorthodox senders. Reputable organisations nowadays are sought, as are connections to them.

The true URL is served by modifying in a way that is nearly identical to the original by attackers who prefer to communicate with victims using a secure communication protocol. The victim can defend himself from the attack at this point if he is aware that the website is a fake. Due to the fact that most of these messages warn users and attempt to induce panic for submitting the victim's private information on the forwarded page, it is quite difficult for the victim to identify the assault on their own.

To trick users, there are many different kinds of phishing assaults. To counter phishing attempts, a variety of tools and approaches are now available. Detecting online phishing involves a variety of methods, including classification [6]. The following list includes typical phishing attack types and classification methods:

A. Different types of phishing attacks

To test the strength of internet security, attackers employ a variety of techniques. In order to take advantage of security system flaws, they constantly look for them. The following list contains a variety of phishing attacks that are unique from one another:

1) Phishing using algorithms

America Online (AOL), a website built using an algorithm, stopped the first phishing attempt. To match America Online accounts' credit card numbers, the cheater used an algorithm.

2) Disingenuous phishing

To trick online visitors, the deceiver employs a variety of techniques. For account verification, fishers email users. Links and buttons are used, and clicks are requested. The user's personal information is stolen and stored on a website that is accessible only through links that are hidden.

3) URL spoofing

The use of a hidden link is used in another type of phishing assault called Universal Resource Locator (URL) phishing. The hackers' website is accessible via the link. The user's information is stored on the hackers' website when they are forwarded to it after they click the link.

4) Hosts Report Poisoning

The Windows operating system platform uses it to corrupt the host file. When a person uses a search engine to find a website, they may find that the site is blocked from them. The user's data is captured and stolen if it can redirect to the bogus site.

5) Content-Injection Phishing

Users are the target of hackers that pose a false website as the real thing. To deceive the user or present the company inaccurately is the motivation. Content spoofing is another name for it. In order to trick the user and gather data for their server, the attackers employ this tactic.

6) Clone Phishing

Clone refers to the process of creating persons who are identical to the original. Frequently, it occurs in genetic engineering. Another type of phishing attack is clone phishing, in which the sender's or recipient's email is compromised by a rival party. first and send a message to the party, the original, the original, the original, the original, the original They ask for the original document to be sent in an updated form.

7) Whaling

Higher-ranking company executives are the intended target of this kind of phishing. The executives are the intended recipients of the email, which contains information on crucial topics. A customer complaint could be the email's message.

8) Sword-fishing

Email scams like spear phishing are one type that target particular people and businesses. To elicit a response from the victim, the attacker sends emails. The email is written in a way where the sender pretends to know a lot of details about the victims, including their name, place of employment, email address, and other details.

To safeguard the end user against phishing assaults, various decision support or detection systems have been created. These systems employ a variety of methodologies, including blacklists, rule-based systems, similarity-based systems, machine learning-based systems, etc.

B. Methods for Detecting Phishing

To classify legal and phishing URLs, respectively, in the List Based method, there are two lists, dubbed the Whitelist and Blacklist. Only if the URL is on the whitelist may a website be accessed in [7]. Utilized in [8] is the blacklist. A

phishing URL's structure is examined using a heuristic-based technique. Previously phishing-related URLs are grouped together in a pattern. Depending on how well a URL follows this pattern, it is categorized. To effectively classify websites, it is important to consider the techniques employed to process the URL's properties [9].

Visually comparable based technique compares how comparable the website pages look on the surface. By examining websites from the server-side, as in [10], websites are categorised as phishing or not. Then, using image processing methods, these two data are compared. Fake web sites are made to look very similar to the real thing; hence, using image processing techniques, it is easier to spot little discrepancies that users cannot readily detect.

Analyzing the content of the pages using a content-based approach. Utilizing search engines, DNS servers, and page content, this strategy extracts features for use on web pages. Authors suggested a detection technique in [11] by assigning weights to the terms that are extracted from URLs and HTML contents. In order to make the URL appear legitimate, attackers may use brand names in the text. Weights are set based on where they appear in distinct URL locations. To find the domain name with the highest frequency among the top 30 results, the most likely terms are picked and sent to Yahoo search. In order to determine whether a website is a phishing scam or not, the domain name owners are compared. By comparing authentic and fraudulent online pages, they were able to identify web pages in [12] by using a logo image.

It is possible to handle ambiguous variables using a fuzzy rule-based technique. After that, human specialists are integrated to classify the variables and their relationships. Using a particular set of measurements and predetermined rules, it is used to categorize web sites according to the amount of phishing that is present in the pages [13]. Less features contribute to improved accuracy for fuzzy logic systems, according to the experimental findings in the research. Relevant features will lessen the efficacy of the classifier and vice versa if a fuzzy logic method is influenced by them.

A machine learning technique uses supervised learning algorithms to develop machine learning models that can determine whether a particular URL is phishing or not. To determine the effectiveness of each model, various algorithms are trained on a dataset and then tested. The model's performance is directly impacted by any changes to the training set of data. With this method, phishing can be detected using effective strategies that operate well. Since machine learning-based phishing detection is a prominent area of study, there are numerous articles that cover it.

The research in this context were thoroughly assessed after a thorough study of the literature. Since they have a defence against zero-day threats, machine learning-based solutions are currently favoured. As a result, the objective of this research is to create a phishing detection system based on a machine learning algorithm for examining the URL address of the target web page. The system is intended to detect phishing attempts quickly, without the need for outside services, and without having to wait for the updating of blacklists. It does this by utilising the system's already-existing, improvable techniques.

II. LITERATURE SURVEY

A few literature reviews on detection of phishing by the various methods like list, rule, visual similarity, and machine learning algorithms are discussed in this chapter, along with a review of a few research work on already existing projects and their results are summarized.

A. Detection systems for list-based phishing

When classifying phishing and non-phishing websites, these systems use two lists. Blacklist and whitelist are the names given to this. While the blacklist includes sites that have been flagged as phishing, the whitelist only includes trustworthy and reliable websites.

Using the whitelist, researchers in [7] were able to recognise phishing websites. Only URLs that are on the whitelist are permitted access to websites during the trial. Using a blacklist is an additional strategy. A few studies employing blacklists like [8] have been done in the literature in addition to using tools like Google Safe Browsing API and PhishNet. The URL is checked against the list in blacklist-based systems, and access is denied if it is not on the list. These systems' major flaw is that a tiny URL change makes it impossible for a match to occur in the list. Additionally, such security measures cannot detect the most recent attacks, known as zero-day attacks.

B. Detection Systems for Rule-Based Phishing

Relational rule mining is used in these systems to acquire the features. According to the guidelines, aspects that are more typical of phishing URLs will be highlighted [14]. This type of system is used in studies with the intention of classifying data more effectively. There are established rules in these systems. When taught with these rules, the system performs with a greater accuracy rate.

The Term Frequency - Inverse Document Frequency (TF-IDF) and criteria were utilised in this situation, similar to the CANTINA [15] study, to identify phishing attacks. Additionally, models were built utilising a few features and criteria in investigations of a similar nature.

C. Phishing Detection Systems for Visual Similarity-Based

These methods compare web sites' aesthetic similarities as their foundation. By looking at them from the server-side, websites are divided into phishing and non-phishing ones. With the aid of image processing methods, these two data are compared. The designs of fake websites frequently resemble those of the real ones. However, there are slight cosmetic variations amongst them. Image processing techniques make it simpler to spot these discrepancies, which users find difficult to perceive. The website's phishing potential is determined based on the similarities found. There are studies that identify distinctions based on fundamental commonalities in the literature, such as the study [16].

D. Detection Systems for Machine Learning Based Phishing

The classification of the specified features using some artificial intelligence approaches is the basis for the detection of the phishing website in machine learning based phishing detection systems. Features are made by compiling elements from various categories, including URL, domain name, website features, and website content, among others. It has a great popularity for user security because of the dynamic structure, particularly for the detection of abnormality in the websites.

Several works on this kind of detecting method can be found in the literature. The machine learning approach was also used in the aforementioned CANTINA project [15]. They found a 90% accuracy rate using heuristic and Tf-Idf techniques. PhishWHO, a phishing defence system created by researchers, uses three methods to ascertain whether a website is trustworthy in [17]. They discovered an accuracy percentage of 96.10% using their 3-Tier Identity Matching System. Phishing websites are categorised using URL data such length, quantity of special characters, directory, domain name, and file name in [18]. Incoming email's title and priority order are covered in [19].

In [20], characteristics relating to transport layer security and URL-based features are combined (Length, slash number, point number and location). Using the rules produced by the apriori algorithm, they discovered a 93% accuracy rate. In [21], a nonlinear regression method is employed to identify phishing websites. Harmony search and Support Vector Machine (SVM) techniques were used to run the system. 20 features and 11055 websites were used. The wrapper was replaced with the decision tree technique for choosing features. Using nonlinear regression based on HS led, they were able to determine the accuracy rate to be 92.80%. Using 209 word-vector characteristics and 17 NLP-based features, a phishing detection system was proposed in a different study [22]. The accuracy rate for the Random Forest method in the hybrid approach was 89.9%, which was the best result when the algorithms Random Forest, SMO, and Nave Bayes were examined.

The number of NLP vectors was raised in the system suggested in [23], and the accuracy values of three different machine learning methods were compared. The Random Forest method in the hybrid approach produced the greatest results, with an accuracy rate of 97.2%, when compared to the SMO and Nave Bayes algorithms. A phishing detection system was developed by researchers in [24] using adaptive self-configuring neural networks for classification. 17 different aspects, including third-party services, are used in the study. As a result, it was indicated that significantly more time was required for implementation in real-world settings.[25] Employed a machine learning algorithm utilising 19 features from the URL and Source code, which are independent of any third party, to differentiate between authentic websites and phishing websites. The findings demonstrated that this method calculated an accuracy rate of 99.09%.

The Monte Carlo algorithm and the risk reduction concept are used in [26] to present a neural network-based classification solution for the identification of phishing websites. [34] concentrated on the impact of training functions on neural networks to improve the effectiveness of the suggestions. E-mail headers, URLs in the message, HTML content, and main text are the four categories listed in [27]. Using 50 features from these categories, the classification was created using machine learning.

This study shows that by combining the use of several features with an analysis of prior research, a greater accuracy rate can be attained. In contrast to other research, a new study was made using features that were picked out and coded from a larger number of features. The model training timeframes and accuracy rates of several methods were contrasted using the machine learning technique.

III. PROPOSED SYSTEM

Using several supervised-learning techniques, the proposed method will focus on increasing the precision of faked website detection. Data was obtained via Kaggle. 11056 occurrences and 32 characteristics make up the dataset. The dataset is then partitioned according to entropy. The fine-tuned dataset shows accuracy. Afterward, the divided dataset is used to observe accuracy. By using correlation and a working model, the optimal attributes for each leaf node are determined. The model's accuracy was seen to be hypertuned depending on the best attributes for each division. ML-based solutions have an advantage over blacklists because they can reduce the impact of zero-hour faked assaults, just like heuristic checks can. It's interesting to note that ML approaches can build their own categorization models by examining vast amounts of data. Due to ML algorithms' ability to discover their own models, manually creating heuristic tests is no longer necessary.

The analysis's module description is represented by the framework in figure 3.

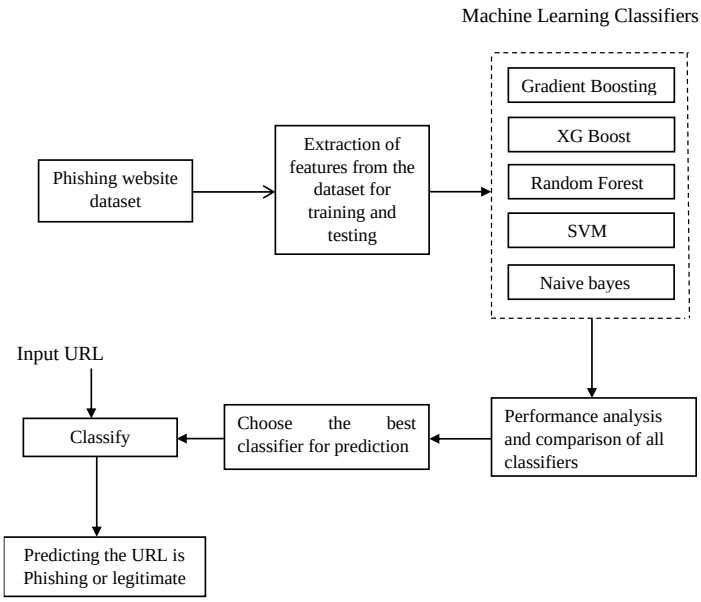


Figure 3. Proposed Methodology.

A. Dataset

In this model, we've combined data sets that we've created with phishing datasets that we've obtained from several online resources, including Kaggle. We test our model using 30% of the Kaggle phishing dataset, and we train our model using the remaining 70%. Data from phishing and genuine websites are included in the dataset, which has 32 columns and 11056 rows.

B. Data preprocessing

The steps involved in data preprocessing include cleaning, instance selection, feature extraction, normalization, transformation, etc. The training dataset as a whole is the end outcome of data preprocessing. How data is pre-processed could have an impact on how the final results are understood. Filling in the gaps in the data, reducing noise, identifying and eliminating outliers, and resolving incompatibilities are all steps in the data cleaning process. The addition of certain databases or data sets may be accomplished through a technique called data integration. When collecting and normalizing data to measure a certain set of data, data transformation is taking place. Data reduction allows for the creation of a very compact dataset overview that nevertheless contributes to the analysis's ability to yield a consistent result.

C. Train-test split

In order for the training dataset to be utilised to detect phishing websites on the testing dataset, the dataset is divided into two subsets: testing set and training set. In order for the training model to adequately train and learn the data, 30% of the data is examined for the testing set.

D. Machine Learning Algorithms

1) Random Forest

When numerous decision trees are combined to form a "forest," it is known as Random Forest, a very versatile and promising supervised machine learning technique. Both classification and regression issues are addressed by its application. It endorsed the concept of ensemble learning, which is a method that combines a number of classifiers to solve a complex problem and improve the performance of the model. For a more accurate prediction, Random Forest brings together various decision trees. The Random Forest model is based on the notion that a combination of models performs significantly better than a single model alone. Each tree provides a vote when Random Forest is used to classify data.

According to the number of votes, the forest choose the category. However, when using Random Forest to model regression, the forest considers the output from every tree. Even though some of the different decision trees might make mistakes, most of them are accurate, therefore it is acceptable to accept the common outcome under the proper guidance. Comparing it to other strategies, the training period is shorter. The outcome is accurately predicted by it. For large datasets, it still functions well. When an oversized data point is absent, it still maintains accuracy. To create

sample datasets for each model, Bootstrap employs row sampling.

In order to see and integrate them, aggregation condenses these sample datasets into a single, concise statistic. Variance refers to a mistake caused by minute fluctuations in the training dataset. In place of the signal, which is the outcomes that are expected, high variance trains noisy or irrelevant data from the dataset. Overfitting is a problem that exists. In testing, a model that is overfit will not be able to tell the difference between the signal and noise, even though it will perform better in training. The bootstrap approach is a technology of high difference in bagging. All things considered, Random Forest is flawless, efficient, and astonishingly rapid to construct.

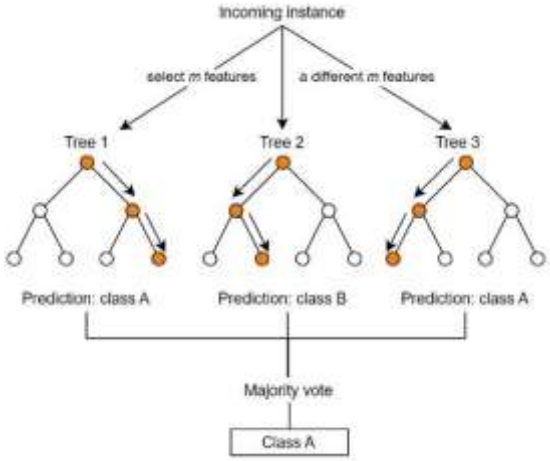


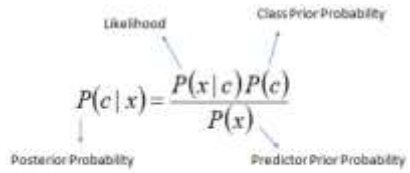
Figure 4. shows the Random Forest Simplified

2) XG Boost

Extreme Gradient Boosting is what the acronym XG Boost means. Gradient-boosted decision trees are used in this application because of their quickness and effectiveness. Using sophisticated approaches, boosting is an ensemble learning strategy that corrects flaws in previously presented models. Up until we discover that no further improvement is possible, models are added one after the other. It employs a gradient descent method to reduce the loss as it adds new models. Utilizing this approach will result in more effective memory and computation resources. This design's goal was to create the greatest model training environment possible using the available sources. Working with XG Boost is primarily motivated by execution speed and model performance. Models for both classification and regression can be supported by this strategy.

3) Naive Bayes

A method of classifying data that relies on the independence of predictions and is based on the Bayes Theorem. For predicting the dataset's class, naive Bayes is employed. By doing so, a multi-class forecast can be made. When compared to other algorithms like logistic regression, Naive Bayes is significantly more effective if the assumption of independence is true. In addition, the categorization needs fewer training data. Spam filtering and document classification are two examples of practical uses for the Naive Bayes classifier. But all that is acknowledged is that it is a poor estimator. This method is simple and effective. The posterior probability of the provided predictor's class (target) is denoted by P(c/x) (attribute).



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

The prior probability of the class is P(c).
P(x/c), is the probability of the predictor given the class.
P(x) is the predictor's prior probability.

4) Support Vector Machine

An example of a supervised machine learning algorithm is the support vector machine, which offers data analysis for regression and classification. Typically, SVM is used for classification. Every feature's value is the same as the coordinate's value. The optimum hyperplane that separates the two classes is then found. The training data are used to create as large a gap as feasible between the points in space that represent the support vector machine's representation as points compared into categories. Because it uses only a portion of the training points in the decision function and is effective and efficient in high-dimensional areas, it is also renowned for having good memory performance. Through the use of five-fold cross-validation, the technique indirectly offers probability estimates

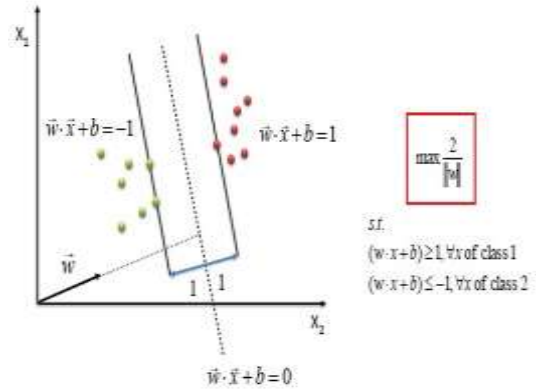


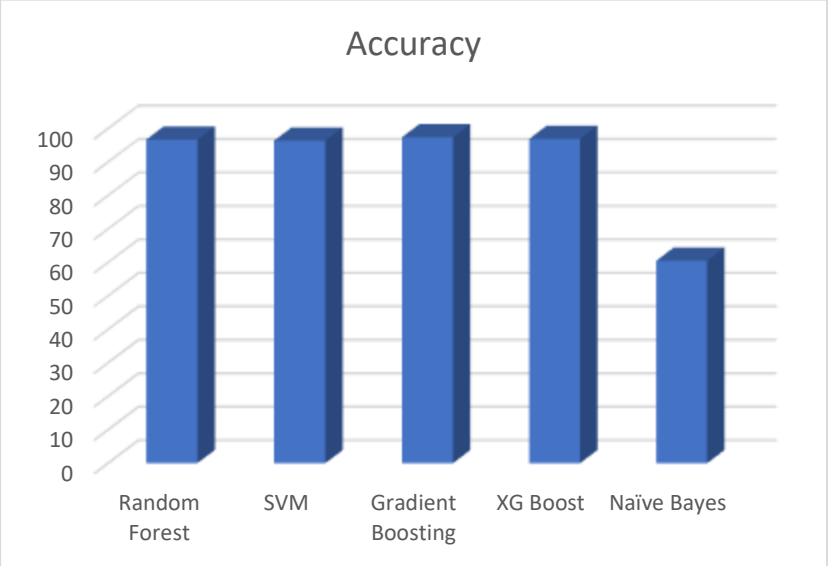
Figure 3. Two support vectors with hyperplane

5) Gradient Boosting

A group of machine learning techniques known as gradient boosting classifiers combines weak learning models to produce a powerful predicting model. Gradient boosting frequently makes use of decision trees. As a result of their success in categorising complicated datasets, gradient boosting models are gaining popularity. A method called gradient boosting can quickly overfit a training dataset. It benefits from regularisation techniques that punish different aspects of the algorithm, and by lowering overfitting, it improves algorithm performance.

IV. RESULTS

High-quality findings from the experiments demonstrate how well the chosen strategy handled the supplied unbalanced dataset. Furthermore, in every experiment carried out, there was no over fitting for the data trained and assessed. Gradient boosting outperformed the other classifiers, according to the data, as it had an accuracy of 97.6% overall. The other classifiers did, however, also attain fairly high accuracies. Additionally, the naive bayes classifier had the lowest performance of all the others, with an accuracy rate of approximately (60.5%). The accuracy outcome, which counts how many of the tested items were correct decisions.



V. CONCLUSION

In this study, numerous machine learning models were tested and trained on a variety of feature sets. This method appears to offer significant advantages for phishing attack detection and prevention and simply relies on the URLs in the headers of web requests. The outcomes shown in this paper also demonstrate the robustness of machine learning approaches in the security area.

The work will continue to go deeper into utilising the online content characteristics and the real-time learning capabilities in the future. This would aid in developing powerful, simultaneously-learnable security appliances in the future.

REFERENCES

[1] State of Cybersecurity Implications for 2016. An ISACA Conference Available: HTTPs resources/state-of cybersecurityimplications-for-2016. [Accessed: 09-Mar-2020].
[2] Republic of Turkey, "National Cyber Security Strategy, 2016," Ministry of Transport Martime Affairs and Communications.
[3] R. Loftus, "What cybersecurity trends should you look out for in 2020?"Daily English Global blogkasperskycom.Available:https://www.kaspersky.com/blog/secure-futures-magazine/2020cybersecurity-predictions/32068/. [Accessed: 09-Mar-2020].
[4] E. Buber, Ö. Demir and O. K. Sahingoz, "Feature selections for the machine learning based detection of phishing websites," 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, 2017, pp. 1-5.
[5] "Retruster," Retruster. [Online]. Available: https://retruster.com/blog/2019-phishing-and-email-fraud-statistics.html. [Accessed: 09-Mar-2020].
[6] "Phishing Activity Trends Reports, 1st-2nd-3rd Half" APWG. [Online]. Available: https://apwg.org/trendsreports/. [Accessed: 09-Mar-2020].
[7] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," Proceedings of the 4th ACM workshop on Digital identity management - DIM 08, pp. 51–60, 2008.
[8] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840–843, 2008.
[9] 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.
[10] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," Special interest tracks and posters of the 14th international conference on World Wide Web-WWW 05, pp. 1060-1061, 2005.
[11] C. L. Tan, K. L. Chiew et al., "Phishing website detection using url assisted brand name weighting system," 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), IEEE, pp. 054-059, 2014.
[12] 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.

[13] K. M. kumar, K. Alekhya, "Detecting phishing websites using fuzzy logic," International Journal of Advanced Research in Computer Engineering Technology (IJARCET), vol. 5, no. 10, 2016.
[14] M. Khonji, Y. Iraqi, and A. Jones, "Phishing Detection: A Literature Survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.
[15] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina,a content based approach to detecting phishing web sites" Proceedings of the 16th international conference on World Wide Web - WWW 07, pp. 639-648, 2007.
[16] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," Special interest tracks and posters of the 14th international conference on World Wide Web - WWW 05, pp. 1060-1061, 2005.
[17] C. L. Tan, K. L. Chiew, K. Wong, and S. N. Sze, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder," Decision Support Systems, vol. 88, pp. 18–27, 2016.
[18] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," 2011 Proceedings IEEE INFOCOM, pp. 191-195, 2011.
[19] R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," Journal of Network and Computer Applications, vol. 36, no. 1, pp. 324–335, 2013.
[20] S. C. Jeeva and E. B. Rajsingh, "Intelligent phishing url detection using association rule mining," Human-centric Computing and Information Sciences, vol. 6, no. 1, Oct. 2016.
[21] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," Soft Computing, vol. 23, no. 12, pp. 4315–4327, 2018.
[22] E. Buber, B. Diri, and O. K. Sahingoz, "Detecting phishing attacks from URL by using NLP techniques," 2017 International Conference on Computer Science and Engineering (UBMK), pp. 337-342, 2017.
[23] E. Buber, B. Diri, and O. K. Sahingoz, "NLP Based Phishing Attack Detection from URLs," Advances in Intelligent Systems and Computing Intelligent Systems Design and Applications, pp. 608–618, 2018.
[24] R. M. Mohammad, F. Thabtah, and L. Mccluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, no. 2, pp. 443–458, 2013.
[25] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," Telecommunication Systems, vol. 68, no. 4, pp. 687–700, 2017.
[26] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han & J. Wang, "The application of a novel neural network in the detection of phishing websites," Journal of Ambient Intelligence and Humanized Computing, pp 1-15, 2018.
[27] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," Decision Support Systems, vol. 107, pp. 88–102, 2018.
[28] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," Neural Computing and Applications, vol. 31, no. 8, pp. 3851–3873, Jun. 2018.
[29] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 300-301, 2018.
[30] R. S. Rao, T. Vaishnavi, and A. R. Pais, "CatchPhish: detection of phishing websites by inspecting URLs," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 2, pp. 813–825, Oct. 2019.
[31] PhishTank-Friends of PhishTank," PhishTank. [Online]. Available: https://www.phishtank.com/friends.php. [Accessed: 09-Mar-2020].
[32] Amazon, Alexa Statistic, [Online]. Available: http://s3.amazonaws.com/alexa-static/top-1m.csv.zip. [Accessed: 09Mar-2020].
[33] G. Karatas, O. Demir and O. K. Sahingoz, "Deep Learning in Intrusion Detection Systems," 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), ANKARA, Turkey, 2018, pp. 113-116, doi: 10.1109/IBIGDELFT.2018.8625278.
[34] G. Karatas and O. K. Sahingoz, "Neural network based intrusion detection systems with different training functions," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-6, doi: 10.1109/ISDFS.2018.8355327.

