ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.971

Intelligent System for Phishing detection on web pages using Random Forest

1st Alvaro Araujo
Peruvian Univ. of Applied Sciences
*L*ima, Peru
u201713444@upc.edu.pe

2nd Gonzalo Felix-Diaz
Peruvian Univ. of Applied Sciences
*L*ima, Peru
u201519058@upc.edu.pe

3rd Juan-Pablo Mansilla
Peruvian Univ. of Applied Sciences
*L*ima, Peru
juan.mansilla@upc.pe

Abstract: During the last years there have been many technological problems, among them, one of those that stands out for its impact worldwide is Phishing, which is a Social Engineering technique that seeks as a benefit to illegally appropriate confidential and private information. In Peru, this problem is spreading considerably due to the lack of interest of both state and private entities in being able to make a contingency plan against this problem, and more and more citizens are affected and there is no clear statement or response from the authorities. In response to this major problem, this paper proposes the use of an intelligent system that uses a Random Forest algorithm, which is based on machine learning and can detect in time the URLs that may be malicious.

Keywords: Phishing, emails, text messages, Banks, website, URL, personal information, cyber attacks, cybersecurity.

Intelligent System for Phishing detection on web pages using Random Forest

1st Alvaro Araujo
Peruvian Univ. of Applied Sciences
*L*ima, Peru
u201713444@upc.edu.pe

2nd Gonzalo Felix-Diaz
Peruvian Univ. of Applied Sciences
*L*ima, Peru
u201519058@upc.edu.pe

3rd Juan-Pablo Mansilla
Peruvian Univ. of Applied Sciences
*L*ima, Peru
juan.mansilla@upc.pe

***Abstract*—During the last years there have been many technological problems, among them, one of those that stands out for its impact worldwide is Phishing, which is a Social Engineering technique that seeks as a benefit to illegally appropriate confidential and private information. In Peru, this problem is spreading considerably due to the lack of interest of both state and private entities in being able to make a contingency plan against this problem, and more and more citizens are affected and there is no clear statement or response from the authorities. In response to this major problem, this paper proposes the use of an intelligent system that uses a Random Forest algorithm, which is based on machine learning and can detect in time the URLs that may be malicious. The operation of the system is based mainly on the analysis of the characteristics of the URLs, which is complemented with a dataset with which the algorithm was trained, and thus give a comprehensive detail of analysis. According to the surveys made to the users of the system, a great acceptance of the intelligent system was obtained due to the precision that it has and in this way avoid falling in the deceptions based on phishing techniques. The results of the use of the Random Forest algorithm for this project were good, having an accuracy of 0.98 compared to 0.93 of the Decision Trees algorithm and 0.91 of Neural Networks algorithm.**

I. Introduction

Phishing is the crime of tricking people into providing sensitive information, such as passwords or credit card numbers. As with Phishing, there are multiple ways to entrap victims, but this tactic is among the most common. Victims receive emails or text messages pretending to be a trusted person or entity, such as: Banks or Government Agencies. When the victim opens the email or text message, they are presented with a frightening message intended to instill fear and impair judgment. The message urges the victim to visit the website and take immediate action or face the consequences. The distinguishing characteristics of Phishing, ranging from the bad wording, the type of logo and URL, also the type of message that appears such as: suspended account, two-factor authentication, tax refund or order confirmation where they request personal information from the user.

New and more convincing ways to trick the user are emerging. In a 2022 State Phish Report that was endorsed by 600 expert IT professionals, a survey was conducted by the State Phish Report that in 2021, 83 percent of organizations abroad (Australia, France, Germany, Japan, Spain, UK, France and USA) suffered successful phishing attacks, 54 percent ended in customer data breaches and 48 percent ended up with compromised credentials and accounts. Other consequences of this survey were 46 percent ransomware infections, 44 percent had major data loss and 27 percent had malware infections. According to in the FBI, in the year 2018, in the world there has been a loss of about 2.7 trillion dollars, due to phishing attacks. According to the Anti-Phishing Working Group (APWG 2018) organization, in 2018 there were 785,920 unique phishing web pages reported, this presented a growth of 69.5 percent compared to the 463,750 cases presented in 2014, in 2021 that there were more than 850 thousand reports and with that there was 6900 million dollars of loss, indicating that the figure has been increasing over the years. In a study called State of Cyber Risk in Latin America in times of Covid-19, a survey of 600 Peruvian companies was made where they confirmed that 49 percent of these companies noticed that there was an increase of cyber attacks between 2020 and 2021, being Phishing the most common. And in the face of this increase, only 20 percent decided to invest more in their cybersecurity budgets. One solution is to be able to detect them before they can deceive the user.

II. Related Work

In this section, techniques, solutions and studies on phishing detection will be reviewed and the different techniques will be evaluated. In [44], the authors examined that the use of the Internet as a means of communication has increased and this also means a growing threat of information piracy for both individuals and organizations. In [14], the author develops an analysis where users are harmed by Phishing attacks, the responses given by the users against the attack are verified, as a result it is evaluated whether the users calibrated their system. In [3], the authors explain the most used method by cybercriminals to steal user data, this involves the channels they use such as mail, SMS and social networks. Due to their nature, as long as cybercriminals continue with their phishing attacks many more people and organizations will suffer data breaches. In [4], the author seeks to solve the problem that users suffer from phishing attacks, this has established as a means to steal private information through fraudulent websites due to the evolution of electronic fraud techniques and the ignorance of users. In [6], the authors also seek to solve the problem of cyber phishing attacks, as these have become in recent years a major threat to governments, businesses and in-

dividuals worldwide. Also, these have evolved rapidly causing problems in their detection for existing methods. According to [28], counterfeit and theft functions ensure high recovery rates by avoiding missing detections as much as possible and removing as many legitimate sites from the dataset as possible. On the other hand, the multi-secular membership functions and evaluation functions ensure high accuracy and low false detection rates. In [7], the main contribution of the paper is based on proposing a novel approach using character-level URL encoding to prevent phishing. In [32], the main contribution of the paper is based on proposing machine learning models that use a limited number of features to classify COVID-19 related domain names as malicious or legitimate for the purpose of improving anti-cyberattack or malware alternatives. In [12], the authors conducted a systematic literature research (SLR) to identify, evaluate and synthesize the results on Deep Learning approaches for phishing detection as reported by selected scientific publications. In [48], the authors contribute with the MFPD framework, a phishing detection approach with multidimensional features based on a fast detection method by using deep learning. In [40], the main contribution of the paper is based on proposing three mutation-based attacks, which differ in the knowledge of the target classifier, and address a key technical challenge: to automatically create an adversarial sample of a known phishing website in a way that can mislead the classifiers. In [10], sources confirm that Phishing is one of the fastest growing cyber threats. This added to the easy distribution of this attack, carried out over the Internet, are factors that make the authors consider it pertinent to design and implement a solution that is up to this growth. In [41], the author provides an intelligent ensemble learning approach based on weighted soft voting for the detection of phishing websites. In [23], the author highlights that Phishing is the method by which cyber attackers or cyber criminals trick internet users to hand over their sensitive data associated with their work, credit cards or personal data itself. Also in [26], the authors argue that the number of pages created for fraudulent purposes has increased over time with the development of the e-commerce industry. Attackers or malicious agents use means such as mail or SMS with fake messages to lure their future victims and trick them into unwittingly handing over their personal data.

In [19], the author proposes an approach to Phishing detection that requires 9 lexical features for effective detection. This is given in order to help users to see the legitimacy or maliciousness of the URL. In [38], the authors propose a real-time web page phishing detection system by analyzing the URL of the page. The performance of such a system is evaluated with seven machine learning classification algorithms: Naive Bayes, Random Forest, kNN, Adaboost, K-star, SMO and Decision Tree and two feature extraction techniques: natural processing language (NLP) and word vectors or word embedding. In [1], the authors aim to use different properties of the URLs of websites and use a Machine Learning model for the classification of URLs that are phishing or not. In [18], the main contribution of the paper is based on a collaborative approach for early detection of unwanted malicious emails and its application in large enterprises. In [25], the main contribution of the article is based on proposing a blockchain-based computing verification protocol, called EntrapNet, for distributed shared computing networks, an emerging underlying network for many Internet of Things (IoT) applications. All with the purpose of reducing the possibility of receiving incorrect computing results from untrusted service providers that have offered computing resources. In [30], the authors' contribution was to present a simple but efficient deep learning model for email classification. For it considers and applies different performance measures to build a 3-class email filter capable of separating emails. In [8], the authors seek to resolve the absence of a classification algorithm that dominates in terms of performance and efficiency for the proposed systems. In [43], in this paper, the authors perform performance experimentation of eight supervised learning classification algorithms with three public datasets: UCI-2015, UCI-2016 and MDP-2018. The algorithms used are the following: AdaBoost (Adaptive Boosting), Classification and Regression Tree (CART) or classification and regression trees, Gradient Tree Boosting or gradient boosting, k-Nearest Neighbours or k nearest neighbors, Multilayer Perceptron (MLP) or multilayer perceptron, Naive-Bayes or naive Bayesian classifier, Random Forest or random forests and Support-Vector Machine (SVM). In [47], the proposed method includes 3 phases: Data collection and preprocessing module, Feature extraction module and Detection module. In [27], the main contribution of the paper is based on proposing a multi-scale semantic information of different web page modules and extracting multi-scale semantic information of URL, title, body text and invisible text (HTML tags) from both URL and HTML, and performing their fusion from different depths, which is more efficient compared to the methods based on limited text information, third party services or artificial heuristic features. In [11], the authors provide an engine that uses supervised machine learning algorithms based on a combination of features that are uniquely extracted from the URL to block phishing attacks. In [35], the authors propose a logo-based phishing detection mechanism to verify the identity between real and projected entities of a website using a hybrid technique including image-based similarity-based approaches and Machine Learning. In [5], proposed a set of hybrid functions including URL character sequence functions without the knowledge of an expert, various hyperlink information, plain text and noisy HTML data based functions within the HTML source code. In [31], a phishing detection model is proposed using machine learning techniques to partition the dataset to train a detection model and validate the results using test data to capture inherent features of email text and other features to classify them as phishing or legitimate. In [22], the authors provide a human-centric data-driven attention enhancement mechanism for phishing prevention called ADVERT1. In [51], the authors contribute with OFS-NN, a model based on optimal feature selection (OFS) and neural network (NN) method for phishing website detection. In [42], the authors propose a framework

that employs deep learning and whose main function is to detect URLs of phishing sites when they are entered in a web browser. In [46], the authors propose a solution based on the use of convolutional neural networks (CNN). The authors comment that several cases of phishing occur when Internet users do not recognize that the URL of the site they are visiting has one or more "misspellings". In [15], the authors propose a system developed with an algorithm created by themselves. The algorithm in question is called Cumulative Distribution Function gradient (CDF-g), which will be employed to identify features and the cutoff rank. In [9], the author proposes SPWalk, which is an unsupervised feature learning algorithm for Phishing detection, which are similar property nodes that reference a collection of Phishing web pages or legitimate web pages. In [29], the author presents different Machine Learning based Phishing detection techniques, in addition to that he analyzes different mechanisms and taxonomy used in each detection technique followed by an upper bound computation time. In [33], the author proposes a framework for detecting phishing websites using a stacking model. For this, various machine learning algorithms are going to be used. In [34], the author proposes a comprehensive analysis of various machine learning algorithms to evaluate their performance on multiple datasets. In [17], the most effective technique against phishing is education. Educated Internet users will know how to detect when a phone message or email has phishing intentions. In [49], the authors mention that they have come up with the solution because, although there are currently techniques to identify phishing cases on websites, they have some limitations. On the one hand, for machine learning solutions, certain features need to be extracted to train the models, which is time-consuming and requires personnel with a certain level of technical knowledge. In [21], the authors were able to identify the existence of solutions proposed by other researchers for phishing detection; however, they consider that these have certain disadvantages in each of the techniques used, and for this reason they fail to fully protect users from this type of malicious attacks.

III. Intelligent System

A. Machine Learning

There are currently several solutions proposed by different authors for detecting phishing on web pages; however, this type of attack is on the rise and continues to find victims in the digital world. Several researches have been conducted to prevent, mitigate and even correct phishing attacks. Most of the research focuses on the use of different machine learning models, deep learning models and/or combinations of models. Machine Learning is a discipline in the field of AI, which is capable of identifying patterns in massive data and making predictions, thus enabling computers to perform processes autonomously and without the need to be previously programmed. Machine learning and deep learning models are mainly statistical models designed to perform specific tasks effectively without external instructions, yet they lack accuracy in performing those specific tasks, resulting in wrong classifications. Behind Machine Learning there is a logic that in this case is mathematics where algorithms are applied. The most popular of these is "Classification", it is one of the tasks most frequently performed by so-called Intelligent Systems. Therefore, a large number of paradigms developed either by Statistics or Artificial Intelligence (Neural Networks, Decision Trees and Random Forest) are able to perform the classification tasks.

As a result of applying a classification method, two errors will be made, in the case of a binary variable that takes values 0 and 1, there will be zeros that are incorrectly classified as ones and ones that are incorrectly classified as zeros. From this count the following classification table can be constructed:

Real Value Y_i

Estimated Value $\hat{Y}$

$$Y_i = 0 \tag{1}$$

$$Y_i = 1 \tag{2}$$

$$\hat{Y}_i = 0; (P_{11}, P_{12}) \tag{3}$$

$$\hat{Y}_i = 1; (P_{21}, P_{22}) \tag{4}$$

Where P_{11} and P_{12} will correspond to correct predictions (values 0 well predicted in the first case and values 1 well predicted in the second case), while P_{21} and P_{22} will correspond to erroneous predictions (values 1 low predicted in the first case and values 0 low predicted in the second case). From these values you can define the following indices that appear:

- Hit rate
  Quotient between the correct predictions and the total predictions

$$\frac{P_{11} + P_{22}}{P_{11} + P_{12} + P_{21} + P_{22}} \tag{5}$$

- Error rate
  Ratio of incorrect predictions to total predictions

$$\frac{P_{12} + P_{21}}{P_{11} + P_{12} + P_{21} + P_{22}} \tag{6}$$

- Specificity
  Ratio between the frequency of correct zero values and the total number of zero values observed

$$\frac{P_{11}}{P_{11} + P_{21}} \tag{7}$$

- Sensitivity
  Ratio between the frequency of correct one values and the total observed one values

$$\frac{P_{22}}{P_{12} + P_{22}} \tag{8}$$

- False zero rate
  Ratio between the frequency of incorrect zero values and the total zero values observed

$$\frac{P_{21}}{P_{11} + P_{21}} \tag{9}$$

- False ones rate
  Ratio between the frequency of incorrect one values and the total observed one values

$$\frac{P_{12}}{P_{12} + P_{22}} \tag{10}$$

B. Artificial neural networks

A neural network is a method of artificial intelligence that teaches computers to process data in a way that is inspired by the way the human brain does. This is a type of machine learning process called deep learning, which uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and continually improve. In this way, artificial neural networks try to solve complicated problems, such as document summaries or face recognition, with greater precision.

What has attracted the most interest in neural networks is the possibility of learning. Given a given task to solve, and a class of functions F, learning consists of using a set of observations to find $f^* \in F$ which solves the task in some optimal way.

This implies the definition of a cost function:

$$C : F \to \mathbb{R} \tag{11}$$

such that, for the optimal solution:

$$f^*, C(f^*) \leq C(f) \forall f \in F \tag{12}$$

That is, no solution has a cost less than the cost of the optimal solution.

C. Decision tree

They are statistical algorithms or machine learning techniques that allow us to build predictive data analytics models for Big Data based on their classification according to certain characteristics or properties, or on regression through the relationship between different variables to predict the value of another. .

A tree can be "learned" by partitioning the initial set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion ends when the subset at a node all has the same value of the target variable, or when the partition no longer adds value to the predictions. This top-down induction of decision trees (ITDAD) process is an example of a greedy algorithm, and is by far the most common strategy for learning decision trees from data.

The data comes in records of the form:

$$(x, Y) = (x_1, x_2, x_3, ..., x_k, Y) \tag{13}$$

The dependent variable, Y, is the target variable that we are trying to understand, classify, or generalize. The vector x is made up of the input variables, x_1, x_2, x_3 etc., which are used for that task.

D. Random Forest

Random forest is a set of decision trees that solve classification and regression problems, one of the most important features is data management. They present estimates of variable importance, that is, neural networks. It also provides a great way to handle missing data. Missing values are replaced with the variable that occurs most frequently at a particular node. Among all available classification methods, random forests provide the highest accuracy. The Random Forest technique can also handle large data with numerous variables running into the thousands. You can automatically balance data sets when one class is less prevalent than other classes in the data. The method also handles variables quickly, making it suitable for complicated tasks.

Whether you have a regression or classification task, Random Forest is an applicable model for your needs. It can handle binary features, categorical features, and numeric features. There is very little pre-processing that needs to be done. The data does not need to be rescaled or transformed. It's faster to train than decision trees because we're only working on a subset of the features in this model, so we can easily work with hundreds of features. The prediction speed is significantly faster than the training speed because we can save the generated forests for future use. The Random Forest algorithm training applies the general aggregation technique, giving a training set:

$$X = x_1, ...x_n \tag{14}$$

with answers:

$$Y = y_1, ...y_n \tag{15}$$

With "n" times a random sample is selected and with them the trees are adjusted to this sample. With "n" training instances of "X", "Y" becomes X_n, Y_n.

$$\hat{f} = \frac{1}{N} \sum_{n=1}^{n} f_n(x') \tag{16}$$

After training, predictions for the unseen samples x' are made by averaging the predictions of the individual regression trees at x'.

TABLE I
SOLUTIONS

	Neural Networks	Decision tree	Random Forest
Standars	Score	Score	Score
Adaptability	3	3	4
Integration	3	1	3
Efficiency	4	4	4
Total	10	8	11

The previous table explains the reason why Random Forest was used, adaptability, integration and efficiency were evaluated. The three solutions are very good and developable, but Random Forest manages to be better as a solution for the present project.

E. Smart System

This proposal proposes to be able to detect Phishing in real time by identifying URL's, for which our system will have as its main task the exact detection of Phishing of a URL, which usually arrives in emails or text messages that are sent by cyber criminals. The system begins when the Administrator enters the "ENTRY" module and starts the system and creates a network, in which you can create different accounts and be able to manage them independently and / or together, once an account is created it is assigned to a user, which you can log in with your respective credentials. The user already with their credentials created from a device (PC or laptop), connected to the internet, enter the system, which is hosted on Heroku (Fig 1).

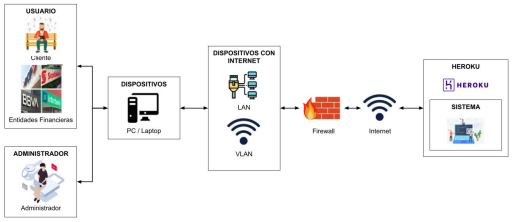


Fig. 1. Physical Architecture.

In the case of Logical Architecture, it is presented how the system is composed, Front End and Back End, the Front End is using HTML, Java Script, CSS library and Bootstrap that are the presentation layer. In the case of the logical and data layer, which is the Back end, it is made up of the Machine learning technique, Python language, and the database, which is MySql (Fig. 2). With both layers the system is developed and allows the detection of URLs with Phishing.

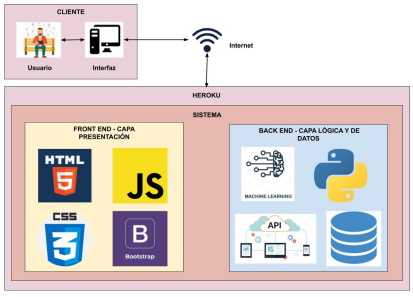


Fig. 2. Logic Architecture.

With both layers, the start, registry, login and analysis system interfaces are shown:

Fig. 3. Homepage.

In the beginning part, the following buttons are shown as "About" where the timeline, mission and vision are explained; the "Services" offered by the intelligent system; the "Work Team" where the individuals involved in the system are detailed and finally the "Login".

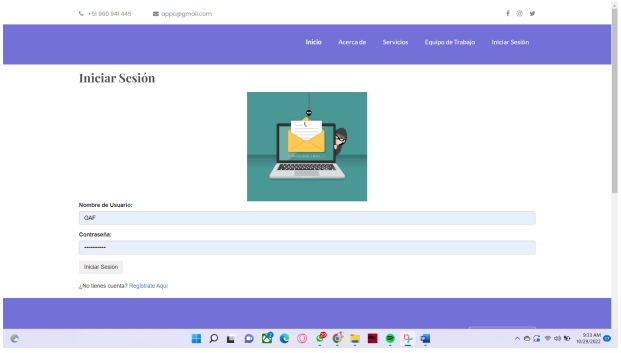


Fig. 4. Login.

In the login part we enter the previously registered data, in case they did not have credentials, the user must register.

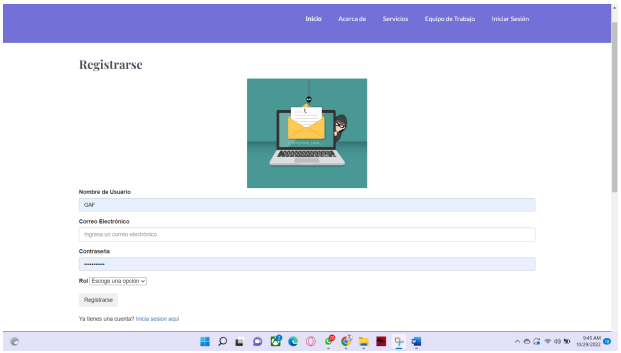


Fig. 5. Sign in.

In the registration part, the user must register with a username, a valid email, a password and choose a role.

In the part of the validator is where to place the link of suspicion of Phishing and we click on validate.

Fig. 6. URL validator.

Fig. 7. Analyzed link.

Once the link has been analyzed we can see the details of why it is legitimate or suspected of Phishing.

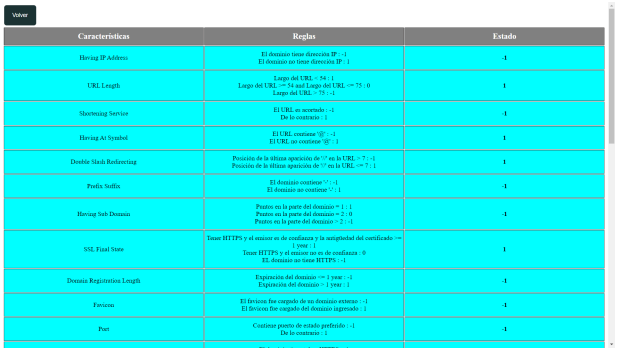


Fig. 8. Detail.

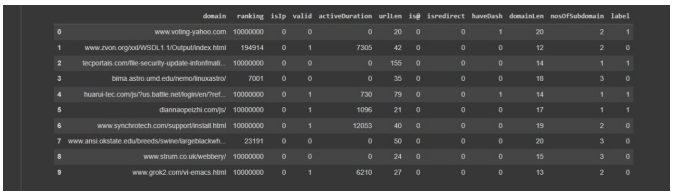
Here it is possible to appreciate that it was valid for the link to be valid or suspected of phishing.

IV. Experiments

A. Dataset

The functions that were taken were from a public Dataset, hosted in the Kaggle repository [33]. The creators of this Dataset classified the basic characteristics that should be able to identify a suspicious URL in this case, such as the domain, URL length, the position of the "//", HTTPS, number of times the website is redirected, etc. With this, our model is trained in such a way that over time it manages to predict a Phishing page more effectively.

The dataset indicates whether the URL is legitimate or suspected Phishing. In addition to this, within the dataset it already has a considerable number of pages that have already been cataloged as suspected Phishing, so when they are a URL already included in the dataset, the response time is faster (Fig. 9).



	domain	ranking	isIp	valid	activeDuration	urlLen	is@	isredirect	haveDash	domainLen	nosOfSubdomain	label
0	www.voting-yahoo.com	10000000	0	0	0	20	0	0	1	20	2	1
1	www.zvon.org/xxl/WSDL1.1/Output/index.html	194914	0	1	7305	42	0	0	0	12	2	0
2	tecportals.com/file-security-update-infonfmati...	10000000	0	0	0	155	0	0	0	14	1	1
3	bima.astro.umd.edu/nemo/linuxastro/	7001	0	0	0	35	0	0	0	18	3	0
4	huarui-tec.com/js/?us.battle.net/login/en/?ref...	10000000	0	1	730	79	0	0	1	14	1	1
5	diannaopeizhi.com/js/	10000000	0	1	1096	21	0	0	0	17	1	1
6	www.synchrotech.com/support/install.html	10000000	0	1	12053	40	0	0	0	19	2	0
7	www.ansi.okstate.edu/breeds/swine/largeblackwh...	23191	0	0	0	50	0	0	0	20	3	0
8	www.strum.co.uk/webben/	10000000	0	0	0	24	0	0	0	15	3	0
9	www.grok2.com/vi-emacs.html	10000000	0	1	6210	27	0	0	0	13	2	0

Fig. 9. Dataset URL.

B. Model Training

In this phase, there is a considerable amount of data, which is separated into a part for training the algorithm and giving all the information so that it successfully finds the necessary patterns and later so that its predictions are much more accurate.

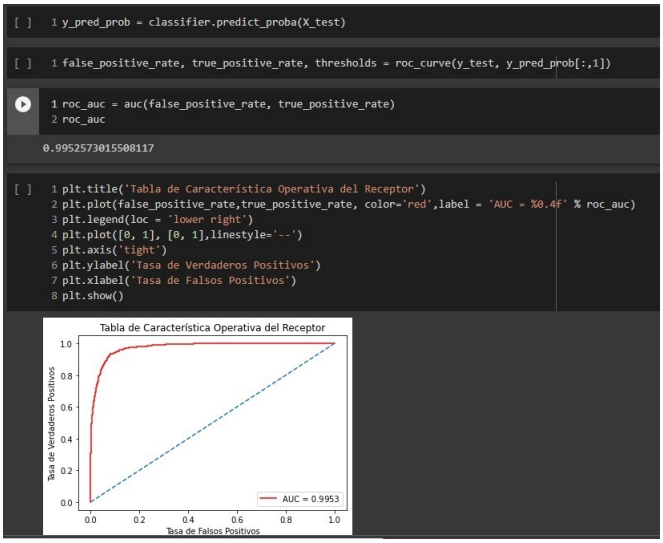


Fig. 10. Algorithm training.

C. Results obtained

Once the training with the three solutions, which are Artificial Neural Networks, Decision Trees and Random Forest, is finished, the data obtained from the three algorithms will be used for the testing phase. With which we will be able to ask the algorithms questions and evaluate the answers that each one gives, that is, we will know in the training phase which algorithm was more successful and which not so much.

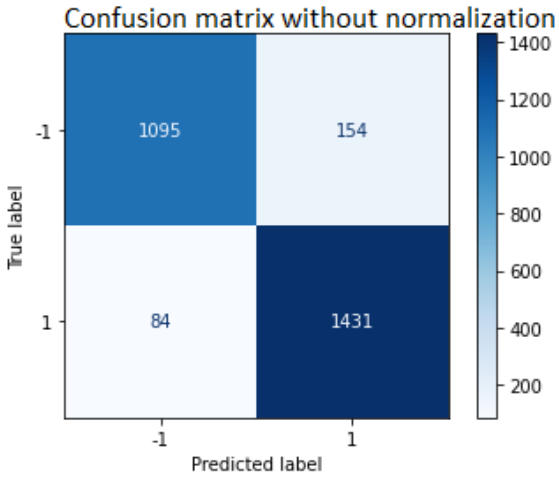


Fig. 11. Confusion matrix without norm. DT

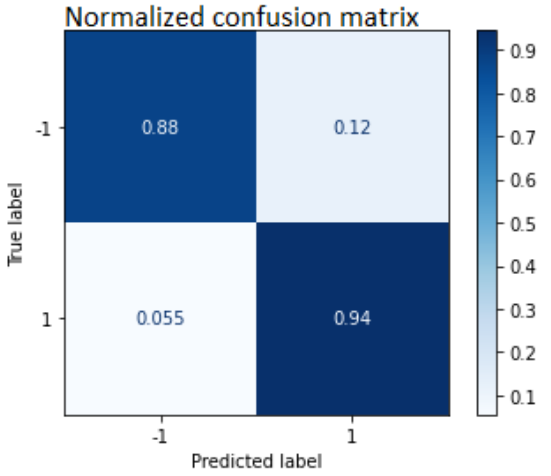


Fig. 12. Confusion matrix norm. DT

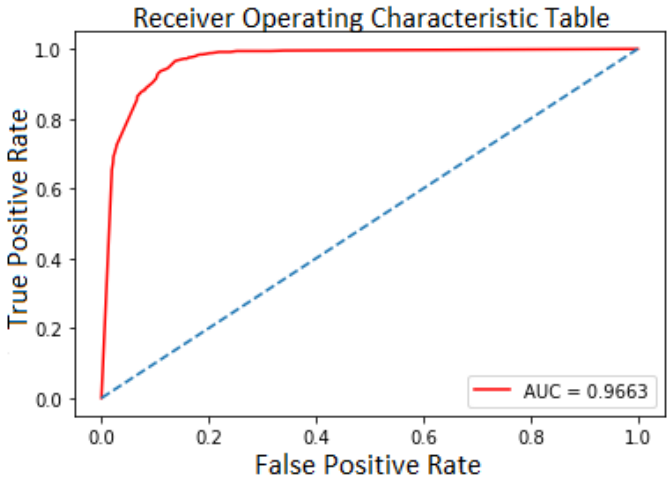


Fig. 13. False positive rate DT

TABLE II
DT

	precision	recall	f1-score	support
-1	0.93	0.88	0.90	1249
1	0.90	0.94	0.92	1515
accuracy			0.91	2764
macro avg	0.92	0.91	0.91	2764
weighted avg	0.91	0.91	0.91	2764

The previous table and graphs show us the results obtained from the training of the decision tree algorithm and we see that from each graph the necessary and relevant information is extracted to form the table and give how much its precision was, which in this case was 0.93 and 0.90. coming in third place.

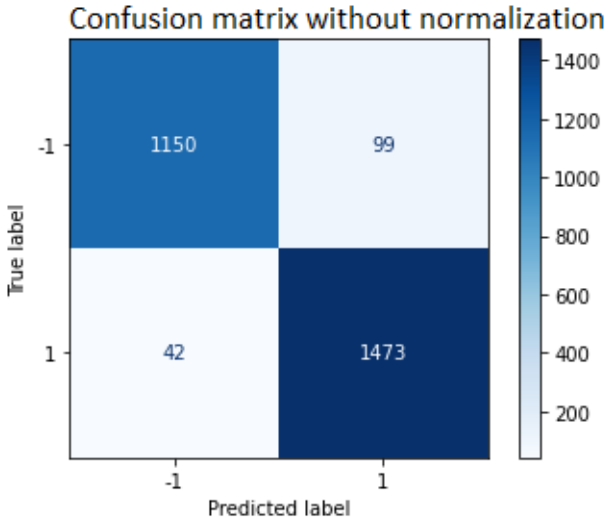


Fig. 14. Confusion matrix without norm. RF

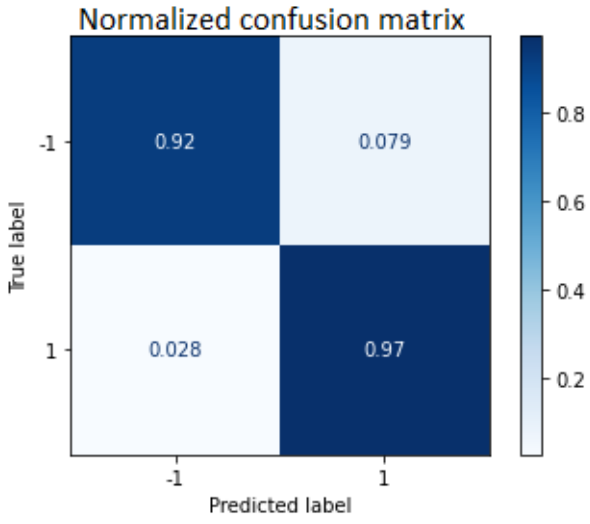


Fig. 15. Confusion matrix norm. RF

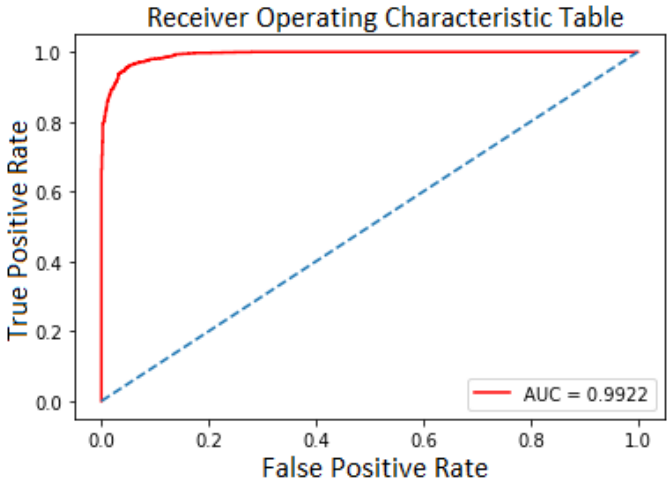


Fig. 16. False positive rate RF

TABLE III
RF

	precision	recall	f1-score	support
-1	0.99	0.92	0.94	1249
1	0.98	0.97	0.95	1515
accuracy			0.95	2764
macro avg	0.95	0.95	0.95	2764
weighted avg	0.95	0.95	0.95	2764

The previous table and graphs show us the results obtained from the training of the decision tree algorithm and we see that from each graph the necessary and relevant information is extracted to form the table and give its precision, which in this case was 0.99 and 0.98. arriving in the first place and the reason why the present project uses said algorithm.

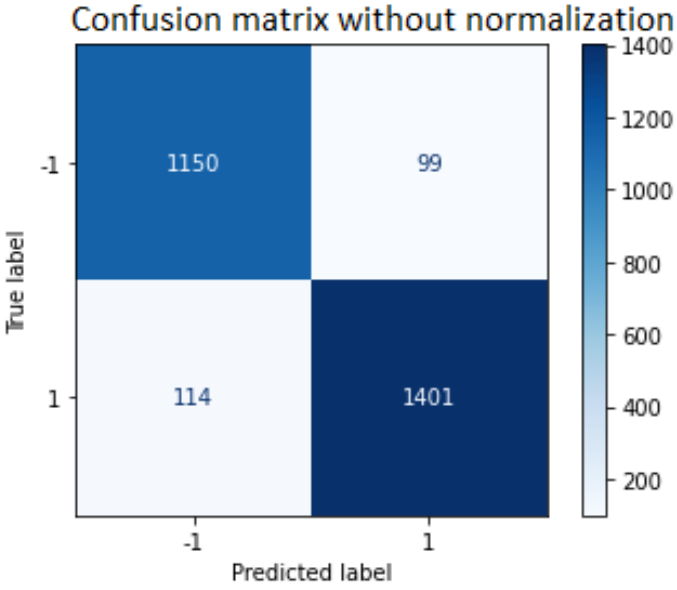


Fig. 17. Confusion matrix without norm. NN

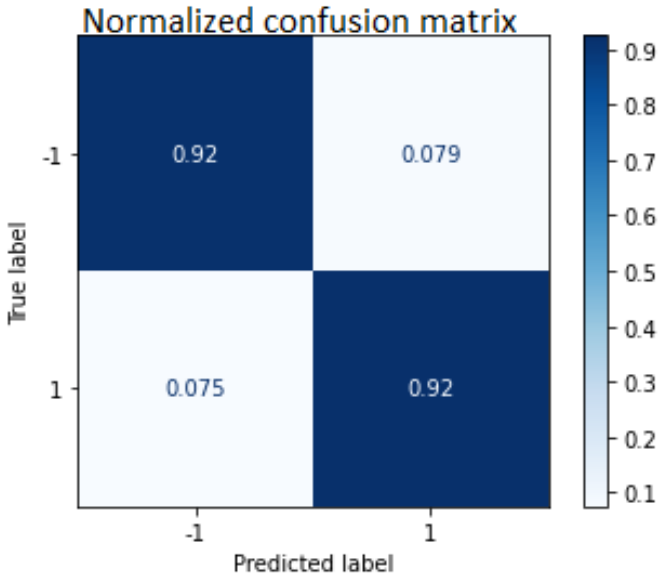


Fig. 18. Confusion matrix norm. NN

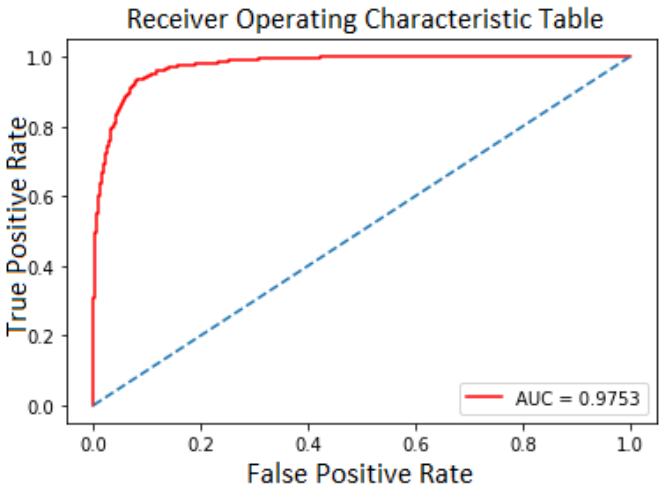


Fig. 19. False positive rate NN

The previous table and graphs show us the results obtained from the training of the decision tree algorithm and we see that from each graph the necessary and relevant information is extracted to form the table and give how much its precision was, which in this case was 0.91 and 0.93. coming in second

TABLE IV
NN

	precision	recall	f1-score	support
-1	0.91	0.92	0.92	1249
1	0.93	0.92	0.93	1515
accuracy			0.92	2764
macro avg	0.92	0.92	0.92	2764
weighted avg	0.92	0.92	0.92	2764

place.

TABLE V
RESULTS COMPARISON

DT	RF	NN
0.93	0.99	0.91

The table above shows the classification results of each of the 3 algorithms mentioned in the project. The 3 algorithms have a good performance for the work, however, the one that stands out the most is the Random Forest algorithm with its 0.99 efficiency.

V. CONCLUSIONS

Machine learning algorithms for phishing detection have helped to reduce cyberattacks, so the proposed intelligent system must grow and have greater precision at the time of phishing detection. With proper algorithm training, the decision when choosing an algorithm will be very accurate. Cyber attacks have increased and will continue to increase, but the correct way to counter them is with the help of systems and education on how to detect phishing on web pages. The intelligent system for detecting phishing on the pages had a successful deployment thanks to the selected machine learning method, which in this case is Random Forest. Random Forest was chosen for its random classification method that allows a better and faster search. The intelligent system was tested by 3 expert users in the financial sector and with knowledge of Systems Engineering and 20 basic users without experience in the field, the results indicate that the proposed system is functional and friendly, for which a user without knowledge of systems could easily navigate the application. With a survey of fifteen users without knowledge in the field, more than ninety percent indicated that the system is friendly and easy to use, they also indicated that they understand the results obtained from the analysis of the URLs. The proposed intelligent system has a continuity plan and therefore its detection in the future will be more accurate and of great help to the virtual financial community.

REFERENCES

[1] Abutair, H., Belghith, A., AlAhmadi, S. (2019). CBR-PDS: A case-based reasoning phishing detection system. Journal of Ambient Intelligence and Humanized Computing, 10(7), 2593-2606. doi:10.1007/s12652-018-0736-0

[2] Al-Sarem, M., Saeed, F., Al-Mekhlafi, Z. G., Mohammed, B. A., Al-Hadhrami, T., Alshammari, M. T., Alreshidi, A., Al-shammari, T. S. (2021). An optimized stacking ensemble model for phishing websites detection. Electronics (Switzerland), 10(11) doi:10.3390/electronics10111285

[3] Ali, W., Malebary, S. (2020). Particle swarm optimization-based feature weighting for improving intelligent phishing website detection. IEEE Access, 8, 116766-116780. doi:10.1109/ACCESS.2020.3003569

[4] Aljofey, A., Jiang, Q., Qu, Q., Huang, M., Niyigena, J. -. (2020). An effective phishing detection model based on character level convolutional neural network from URL. Electronics (Switzerland), 9(9), 1-24. doi:10.3390/electronics9091514

[5] Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. Scientific Reports, 12(1) doi:10.1038/s41598-022-10841-5

[6] Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., Alazzawi, A. K. (2020). AI meta-learners and extra-trees algorithm for the detection of phishing websites. IEEE Access, 8, 142532-142542. doi:10.1109/ACCESS.2020.3013699

[7] Alshehri, M., Abugabah, A., Algarni, A., Almotairi, S. (2022). Character-level word encoding deep learning model for combating cyber threats in phishing URL detection. Computers and Electrical Engineering, 100 doi:10.1016/j.compeleceng.2022.107868

[8] Anupam, S., Kar, A. K. (2021). Phishing website detection using support vector machines and nature-inspired optimization algorithms. Telecommunication Systems, 76(1), 17-32. doi:10.1007/s11235-020-00739-w

[9] Barraclough, P. A., Fehringer, G., Woodward, J. (2021). Intelligent cyber-phishing detection for online. Computers and Security, 104 doi:10.1016/j.cose.2020.102123

[10] Bustio-Martínez, L., Álvarez-Carmona, M. A., Herrera-Semenets, V., Feregrino-Uribe, C., Cumplido, R. (2022). A lightweight data representation for phishing URLs detection in IoT environments. Information Sciences, 603, 42-59. doi:10.1016/j.ins.2022.04.059

[11] Butnaru, A., Mylonas, A., Pitropakis, N. (2021). Towards lightweight url-based phishing detection. Future Internet, 13(6) doi:10.3390/3fi13060154

[12] Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., Shukla, S. (2022). Applications of deep learning for phishing detection: A systematic literature review. Knowledge and Information Systems, 64(6), 1457-1500. doi:10.1007/s10115-022-01672-x

[13] Chen, J. -., Ma, Y. -., Huang, K. -. (2020). Intelligent visual similarity-based phishing websites detection. Symmetry, 12(10), 1-16. doi:10.3390/sym12101681

[14] Chen, Y., Zahedi, F. M., Abbasi, A., Dobolyi, D. (2021). Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools. Information and Management, 58(1) doi:10.1016/j.im.2020.103394

[15] Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S. C., Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences, 484, 153-166. doi:10.1016/j.ins.2019.01.064

[16] Das, M., Saraswathi, S., Panda, R., Mishra, A. K., Tripathy, A. K. (2021). Exquisite analysis of popular machine Learning–Based phishing detection techniques for cyber systems. Journal of Applied Security Research, 16(4), 538-562. doi:10.1080/19361610.2020.1816440

[17] Ding, Y., Luktarhan, N., Li, K., Slamu, W. (2019). A keyword-based combination approach for detecting phishing webpages. Computers and Security, 84, 256-275. doi:10.1016/j.cose.2019.03.018

[18] Gallo, L., Maiello, A., Botta, A., Ventre, G. (2021). 2 years in the anti-phishing group of a large company. Computers and Security, 105 doi:10.1016/j.cose.2021.102259

[19] Gualberto, E. S., De Sousa, R. T., De Vieira, T. P. B., Da Costa, J. P. C. L., Duque, C. G. (2020). From feature engineering and topics models to enhanced prediction rates in phishing detection. IEEE Access, 8, 76368-76385. doi:10.1109/ACCESS.2020.2989126

[20] Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. Computer Communications, 175, 47-57. doi:10.1016/j.comcom.2021.04.023

[21] Hr, M. G., Mv, A., Gunesh Prasad, S., Vinay, S. (2020). Development of anti-phishing browser based on random forest and rule of extraction framework. Cybersecurity, 3(1) doi:10.1186/s42400-020-00059-1

[22] Huang, L., Jia, S., Balcetis, E., Zhu, Q. (2022). ADVERT: An adaptive and data-driven attention enhancement mechanism for phishing prevention. IEEE Transactions on Information Forensics and Security, 17, 2585-2597. doi:10.1109/TIFS.2022.3189530

[23] Jain, A. K., Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing, 10(5), 2015-2028. doi:10.1007/s12652-018-0798-z

[24] Khan, S. A., Khan, W., Hussain, A. (2020). Phishing attacks and websites classification using machine learning and multiple datasets (A comparative analysis) doi:10.1007/978-3-030-60796-826 Retrieved from www.scopus.com

[25] Li, C., Zhang, L., Fang, S. (2022). EntrapNet: A blockchain-based verification protocol for trustless computing. IEEE Internet of Things Journal, 9(11), 8024-8035. doi:10.1109/JIOT.2021.3124007

[26] Liew, S. W., Sani, N. F. M., Abdullah, M. T., Yaakob, R., Sharum, M. Y. (2019). An effective security alert mechanism for real-time phishing tweet detection on twitter. Computers and Security, 83, 201-207. doi:10.1016/j.cose.2019.02.004

[27] Liu, D. -., Geng, G. -., Zhang, X. -. (2022). Multi-scale semantic deep fusion models for phishing website detection. Expert Systems with Applications, 209 doi:10.1016/j.eswa.2022.118305

[28] Liu, D. -., Geng, G. -., Jin, X. -., Wang, W. (2021). An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment. Computers and Security, 110 doi:10.1016/j.cose.2021.102421

[29] Liu, X., Fu, J. (2020). SPWalk: Similar property oriented feature learning for phishing detection. IEEE Access, 8, 87031-87045. doi:10.1109/ACCESS.2020.2992381

[30] Magdy, S., Abouelseoud, Y., Mikhail, M. (2022). Efficient spam and phishing emails filtering based on deep learning. Computer Networks, 206 doi:10.1016/j.comnet.2022.108826

[31] Mughaid, A., AlZu'bi, S., Hnaif, A., Taamneh, S., Alnajjar, A., Elsoud, E. A. (2022). An intelligent cyber security phishing detection system using deep learning techniques. Cluster Computing, doi:10.1007/s10586-022-03604-4

[32] Mvula, P. K., Branco, P., Jourdan, G. -., Viktor, H. L. (2022). COVID-19 malicious domain names classification[formula presented]. Expert Systems with Applications, 204 doi:10.1016/j.eswa.2022.117553

[33] Nagunwa, T., Kearney, P., Fouad, S. (2022). A machine learning approach for detecting fast flux phishing hostnames. Journal of Information Security and Applications, 65 doi:10.1016/j.jisa.2022.103125

[34] Orunsolu, A. A., Sodiya, A. S., Akinwale, A. T. (2022). A predictive model for phishing detection. Journal of King Saud University - Computer and Information Sciences, 34(2), 232-247. doi:10.1016/j.jksuci.2019.12.005

[35] Panda, P., Mishra, A. K., Puthal, D. (2022). A novel logo identification technique for logo-based phishing detection in cyber-physical systems. Future Internet, 14(8) doi:10.3390/fi14080241

[36] Rao, R. S., Vaishnavi, T., Pais, A. R. (2020). CatchPhish: Detection of phishing websites by inspecting URLs. Journal of Ambient Intelligence and Humanized Computing, 11(2), 813-825. doi:10.1007/s12652-019-01311-4

[37] S., E. R., Ravi, R. (2020). A performance analysis of software defined network based prevention on phishing attack in cyberspace using a deep machine learning with CANTINA approach (DMLCA). Computer Communications, 153, 375-381. doi:10.1016/j.comcom.2019.11.047

[38] Sahingoz, O. K., Buber, E., Demir, O., Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357. doi:10.1016/j.eswa.2018.09.029

[39] Sameen, M., Han, K., Hwang, S. O. (2020). PhishHaven - an efficient real-time AI phishing URLs detection system. IEEE Access, 8, 83425-83443. doi:10.1109/ACCESS.2020.2991403

[40] Song, F., Lei, Y., Chen, S., Fan, L., Liu, Y. (2021). Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers. International Journal of Intelligent Systems, 36(9), 5210-5240. doi:10.1002/int.22510

[41] Taha, A. (2021). Intelligent ensemble learning approach for phishing website detection based on weighted soft voting. Mathematics, 9(21) doi:10.3390/math9212799

[42] Tang, L., Mahmoud, Q. H. (2022). A deep learning-based framework for phishing website detection. IEEE Access, 10, 1509-1521. doi:10.1109/ACCESS.2021.3137636

[43] Vaitkevicius, P., Marcinkevicius, V. (2020). Comparison of classification algorithms for detection of phishing websites. Informatica (Netherlands), 31(1), 143-160. doi:10.15388/20-INFOR404

[44] Wang, S., Khan, S., Xu, C., Nazir, S., Hafeez, A. (2020). Deep learning-based efficient model development for phishing detection using random forest and BLSTM classifiers. Complexity, 2020 doi:10.1155/2020/8694796

[45] Web page Phishing Detection Dataset. (2021). Kaggle. Recuperado 25 de septiembre de 2022, de https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset/code

[46] Wei, W., Ke, Q., Nowak, J., Korytkowski, M., Scherer, R., Woźniak, M. (2020). Accurate and fast URL phishing detector: A convolutional neural network approach. Computer Networks, 178 doi:10.1016/j.comnet.2020.107275

[47] Wen, T., Xiao, Y., Wang, A., Wang, H. (2023). A novel hybrid feature fusion model for detecting phishing scam on ethereum using deep neural network. Expert Systems with Applications, 211 doi:10.1016/j.eswa.2022.118463

[48] Yang, P., Zhao, G., Zeng, P. (2019). Phishing webs ite detection based on multidimensional features driven by deep learning. IEEE Access, 7, 15196-15209. doi:10.1109/ACCESS.2019.2892066

[49] Yang, R., Zheng, K., Wu, B., Wu, C., Wang, X. (2021). Phishing website detection based on deep convolutional neural network and random forest ensemble learning. Sensors, 21(24) doi:10.3390/s21248281

[50] Zamir, A., Khan, H. U., Iqbal, T., Yousaf, N., Aslam, F., Anjum, A., Hamdani, M. (2020). Phishing web site detection using diverse machine learning algorithms. Electronic Library, 38(1), 65-80. doi:10.1108/EL-05-2019-0118

[51] Zhu, E., Chen, Y., Ye, C., Li, X., Liu, F. (2019). OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network. IEEE Access, 7, 73271-73284. doi:10.1109/ACCESS.2019.2920655