Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Phishing Website Detection Based on Data Tuning Methods with PCA of Multidimensional Features by Machine Learning Algorithms

Qabeela Qassim Thabit *[1], Mohammed. Kadhim Alkhafaji [2] and Bayadir Abdulrazaq Issa [3]
[1]*Basra Education Directorate, Ministry of Education, Basrah, Iraq.*
[2] *Dep. of Electronic Techniques, Basrah Technical Institute, Southern Technical University, Basra, Iraq.*
[3] *Basrah Electrical Technical College, Southern Technical University, Basrah, Iraq.*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Introduction: Recently, especially in the last decade, cyber fraud crimes have increased in various ways to deceive victims, causing financial losses and losses in data and privacy of many users, so many researchers have focused on addressing cyber fraud problems to reduce them.

Objectives: For example, developing a mechanism to detect phishing operations in several ways, including old, traditional or modern methods, which are represented in using artificial intelligence models to detect behaviors or Internet users, whether suspicious or healthy, in addition, some reputable and solid sites have taken it upon themselves to monitor the database by collecting the methods used by hackers, including the famous Al-Mitre website.

Methods: we presented our work to detect whether a website is legitimate or a phishing website using seven algorithms from machine learning algorithms, which are Ada Boost classifier, Decision Tree classifier, Gradient Boosting, KNN algorithm, Logistic Regression, Random Forest classifier and Support Vector Machine.

Results: emply different implementation methods in the form of three experiments, and we concluded that all seven algorithms are almost good, which is Experiment C, in which we benefited from the advantages and disadvantages of previous experiments, and the accuracy obtained from the seven algorithms was as follows: 84.8, 96.6, 83.9, 98.7, 99.6, 99.9, 97.00, respectively.

Conclusions: It is clear from the above results that machine learning algorithms are a good and effective tool in combating malicious activities due to the algorithms' ability to detect phishing sites.

Keywords: Phishing Website, Ada Boost Classifier, Decision Tree Classifier, Gradient Boosting, KNN Algorithm, Logistic Regression, Random Forest Classifier and Support Vector Machine Algorithm, Python language

INTRODUCTION

Due to the significant increase in the use of the Internet and the unlimited number of platforms and their publishing directions from different fields via websites at the present time, the source of which is almost unknown in terms of security and reliability and whether it is arranged for actual beneficial purposes for the user or a means of catching victims, in addition to the fact that malicious attacks occur every day with new methods so that the user sometimes does not suspect that it is from an unreliable party, including phishing of websites, and this problem is considered the most common. The idea of deception may be to imitate a legitimate website or imitate a scientific or profitable website, and thus the goal is to deceive users and steal their sensitive information, the need has become urgent and necessary to find effective means of detection [1,2]. One of the major and ongoing threats in the digital world is phishing attacks, which are a source of concern for many users at the individual and organizational levels. The term phishing is generally applied to all types of cyber attacks, which are of several types, including email, text messages, phone calls, or sending an illegal link belonging to a hacker. This paper seeks to provide a solution to the link problem and distinguish its type by detecting it using machine learning algorithms [3,4]. In other words, the purpose of phishing is to steal the victims' sensitive data such as passwords and important confidential information such as bank statements. Phishing attacks can be defined as a cyber threat in which attackers are able to trick users by imitating legitimate, authentic websites [5]. After providing an introduction about the topic of paper in general, the remainder of this paper is organized as follows: -

The other sections of this paper are arranged as follows: Section II includes the work related to the research topic, Section III covers the basic information that represents the theoretical and practical background of the research from the details of the data used and an explanation of the machine learning algorithms and the metrics and standards that we draw to evaluate the performance of each model, while Section IV includes the research methodology and in it we addressed the three experiments completed in this paper, Section V was to present the results we obtained from the three experiments, Section VI discusses the results we obtained and compares them with previous results from the same work presented in this paper and finally the conclusions and future recommendations in Section VII.

LITERATURE REVIEW

Phishing, in general, is defined as a cyber-attack that aims to persuade or trick potential victims into revealing sensitive information such as passwords or their private data, including credit card numbers, etc. Online scammers do this by falsely impersonating a trusted person. An official entity or institution, requesting victims' data and demonstrating their sense of urgency. They are phishing attacks that aim to steal or destroy sensitive data by tricking people into revealing their personal data such as passwords, credit card numbers, etc. For this reason, many researchers have set out to find solutions to the problems in which victims fall easy prey to trolls, and as we mentioned, the losses are large and dangerous in this field. A research paper by Ali in 2017 [1] presented an application of some common machine learning techniques with effective and selected features to accurately detect phishing sites. In 2018, Mahajan, R., & Siddavatam, I [2] presented an attack identification model based on verifying the characteristics of suspicious and non-suspicious sites through a WHOIS database. While Ozgur Koray Sahingoz and others [6] in 2019 proposed a real-time anti-phishing system using seven algorithms based on natural language processing, this study enjoyed new literature from other studies, namely language independence, the use of a large volume of phishing data, legitimate data, and actual implementation. In a related context about the exposure of many users to phishing attacks, Volkamer, Renaud, Reinheimer, & Kunz, 2017 [7] presented the five main reasons why computer users fall victim to phishing:

- Most users do not have detailed knowledge about URLs.
- Web pages that can be trusted are not known or distinguished by users.
- Sometimes users or victims do not see the full address of the web page, due to redirection or hidden URLs.
- Sometimes due to rush or that users do not have much time to view the URL, or enter some web pages incorrectly.
- Also, the reason is often that users cannot distinguish between phishing web pages and legitimate ones.

In 2022, Mukta Mithra Raj & J. Angel Arul Jothi [8] proposed a website phishing detection technique based on URL features to predict whether a set of websites is phishing or not with high accuracy. For this purpose, they employed eight machine learning algorithms to classify those websites. In 2022, Almomani, A and others [9] made a contribution to this field of research based on extracting URL features, abnormal features, domain identity, as well as HTML and JavaScript features. To detect phishing sites, domain features were used as semantic features, which facilitated the work of Prosager in terms of making The classification process using these semantic features is more controllable and more efficient. The authors then used 16 machine learning models trained with the 10 semantic features that represented the most effective features and trained detection on two data sets from phishing web pages. In 2019, Chiew, K. L. et al. [10] Machine Learning-Based Phishing Detection introduced a novel framework for feature selection for a system known as Hybrid Ensemble Feature Selection (HEFS). Subsets of main features are generated in the first step of HEFS by utilizing a novel cumulative distribution function gradient (CDF-g) technique. These subsets of primary features are then input into a data perturbation set to generate subsets of secondary features. Using a function perturbation set, a set of primary features is derived in the second stage from subsets of secondary features. By taking into account hybrid features based on URLs and hyperlinks to achieve high accuracy depending on machine learning models without relying on any third-party systems, Guptta, S. et al. essentially provided an existing strategy to detect phishing sites in real-time in 2022 [11]. Phishing is a tactic used by scammers to deceive people online by pretending to be legitimate websites in order to get private data, including credit card details, usernames, passwords, and social security numbers.

A. K. Jain and B. B. Gupta [12] presented a novel method in 2016 for thwarting phishing assaults, which involved employing an automatically refreshed whitelist of trustworthy websites that a specific user had visited. Our suggested method has a high detection rate and quick access time. The browser alerts users not to reveal their personal information when they attempt to access a website that is not on the whitelist.

In order to stop users from opening dangerous URLs and protect user privacy, Jha, A. K. and colleagues [13] develop a tool in 2023 that assists in identifying and differentiating between phishing and trustworthy websites. Aside from additional strategies, the primary classification methods employed are Multinomial NB and Linear Regression. Support vector machines, random forests, and artificial neural networks.

METHODS

3.1. Dataset Details

Before starting with the methodology and automated science methods that we will apply in the next section of the paper, we will review the data set that was relied upon in training the algorithms, which is a set of data downloaded from the site Kaggle[14] Source. The database contains 32 columns, 31 of which are the attributes that the sites carry, and the last column represents the result of these attributes or named output, either the website is a phishing website or the website is a legitimate website. Here we have 11055 rows in this database, and **Figure 1.** shows the percentage of websites that are phishing website to a legitimate websites.

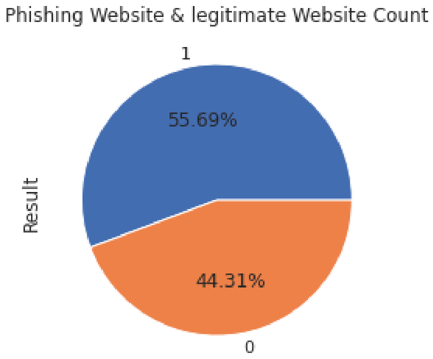

Figure 1. Exploratory data analysis

From Fig. 1, we can see that the data is balanced in proportions, with legitimate websites making up 55.69% of the data, and the remaining 44.31% being phishing websites. Table 1 below shows the database used for phishing data, where 31 columns represent the attributes and the last column shows the result for each row of attributes if the site is legitimate or phishing, and the table shows an important step that is taken to find out if there is missing data. The models used in this paper (training data in the three experiments after implementing the steps mentioned in each experiment) were tested with 70% for training and 30% for testing (in other words, 7737 training data to 3317 testing data) from the database used.

Table 1. Checking stage for missing data

No.	Feature Name	Missing No.
1	Index	0
2	UsingIP	0
3	LongURL	0

4	ShortURL	0
5	Symbol@	0
6	Redirecting//	0
7	PrefixSuffix-	0
8	SubDomains	0
9	HTTPS	0
10	DomainRegLen	0
11	Favicon	0
12	NonStdPort	0
13	HTTPSDomainURL	0
14	RequestURL	0
15	AnchorURL	0
16	LinksInScriptTags	0
17	ServerFormHandler	0
18	InfoEmail	0
19	AbnormalURL	0
20	WebsiteForwarding	0
21	StatusBarCust	0
22	DisableRightClick	0
23	UsingPopupWindow	0
24	IframeRedirection	0
25	AgeofDomain	0
26	DNSRecording	0
27	WebsiteTraffic	0
28	PageRank	0
29	GoogleIndex	0
30	LinksPointingToPage	0
31	StatsReport	0
32	class (output)	0

3.2. Machine Learning Algorithms

One of the branches of computer science that enters into the studies and research of artificial intelligence is machine learning, which has spread widely and distinctively recently because it has achieved a lot through its application in many fields such as computer vision, natural language processing, speech recognition, medicine, industry, agriculture and other applications of great importance that require decisive, fast, accurate and effective decisions at the same time. The results of training and prediction have shown this using several algorithms that were implemented on a set of databases collected by prestigious scientific institutions [15,16], while deep learning models have also developed significantly and rapidly in recent years and are also used in various research fields, whether health care, image encryption or other technical fields that are not limited to them [17]. The following is a brief explanation of the machine learning algorithms that were applied in this paper, which are as follows:

3.2.1. Adaptive Boosting (Ada Boost) Classifier

Coined in 1995 by Yoav Freund and Robert Schapire, Ada Boost, short for Adaptive Boosting, it is a statistical classification algorithm that won the 2003 Gödel Prize. It works in conjunction with many machine learning algorithms to achieve the highest possible level of performance. Ada Boost simply means that subsequent weak learners are adjusted in favor of those instances that were misclassified by previous classifiers. In some problems, it has an advantage over other algorithms because it can be less susceptible to the overfitting problem than other learning algorithms [18].

3.2.2. Decision Tree (DT) Classifier

It is a decision-making process in several stages. This is done with the help of a visual tool that helps in making decisions by asking questions at each stage and getting answers, and then asking questions again until a final result is obtained. This gives a tree-like progression, which is why it is called a decision tree due to the shape that resembles trees. It starts with a single node and then branches out into different branches, which in turn each node branches out into other nodes that represent decisions. What distinguishes the decision tree is that it is versatile and easy to interpret and can deal with any type of data. In addition to being easy to modify and update, it provides a complete study of the data through the progression of the nodes, but it is sometimes unstable and inaccurate and may not be suitable for complex calculations [19], in the following Figure 2. explain the structure of decision tree.

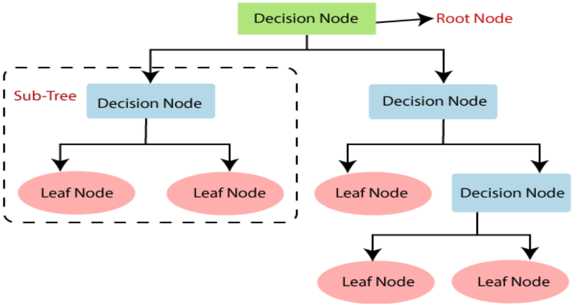

Figure 2. Decision tree structure details [20].

3.2.3. Gradient Boosting (GB) Classifier

Gradient boosting and adaptive boosting are both machine learning techniques that solve logistic regression and classification problems that produce a prediction model in the form of an ensemble of weak prediction models, usually as an ensemble of decision trees. They both increase the efficiency of simple or weak models to obtain better performance in the machine learning algorithm, but gradient boosting is more responsible for adding an improvement to the gradient while adaptive boosting is about the voting weights. So, in short, gradient boosting works specifically to increase accuracy by reducing the loss function (i.e. reducing the error value which is the difference between the actual value and the expected value) and making this loss the target for the next iteration and so the algorithm works. In contrast, adaptive boosting works to increase accuracy by giving more weight to the target that was misclassified by the model [21, 22, 23]. in the following Figure 3. explain the Flow Chart of GB.

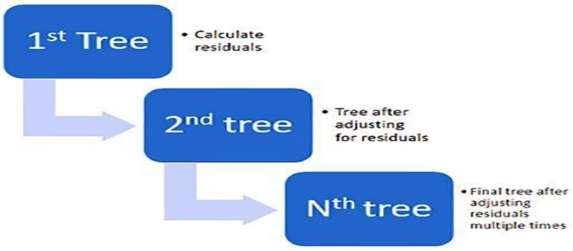
1st Tree
• Calculate residuals
2nd tree
• Tree after adjusting for residuals
Nth tree
• Final tree after adjusting residuals multiple times

Figure 3. Flow chart of GB [23].

3.2.4. K-Nearest Neighbors (KNN) Algorithm

The nearest neighbor algorithm is a supervised learning algorithm that works on the principle that all data points that are close to each other fall into the same class. It is considered a non-parametric supervised learning classifier, and its principle is based on using proximity to make classifications or predictions about the grouping of an individual data point. Due to its simplicity, it is one of the most widely used learning algorithms in all research fields. The strengths of this algorithm are also its ease of interpretation and short computation time [24], in the following Figure 4. show K-Neareast neighbors algorithm.

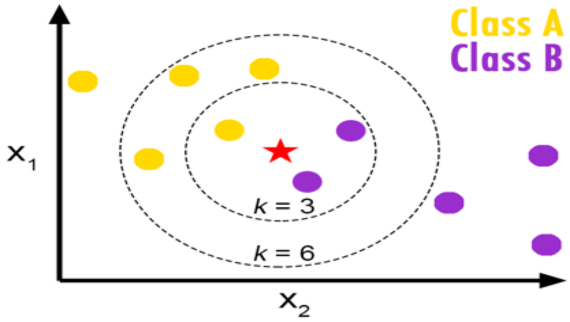

Figure 4. K-Nearest neighbors algorithm [17].

3.2.5. Logistic Regression (LR) Classifier

Logistic regression is a mathematical principle based on data analysis within the application of logarithmic probability function optimization, through which relationships can be found between two variables of data factors. This relationship is useful in predicting the value of one of the factors based on the values given to the other factor. Logistic regression can give us only two values, for example 0 and 1 or yes and no, and in this case it is called binary classification, or it gives several values as a result of the prediction, and in this case it is called multi-valued prediction (classification) [25, 26]. Classification is accomplished with a binary dependent variable (0/1, -1/1, or represented by true/false) and a binary independent variable or deprivation by using the sigma function.

Y=1/(1+e^(-x))   (1)

3.2.6. Random Forest (RF) Classifier

In fact, a group of decision trees form the Random Forest algorithm, which is a supervised machine learning algorithm that works based on a technique called Bagging, which is a preliminary concept for everyone. Its working principle is based on merging the predictions from all trees to make the final prediction. On the other hand, the Random Forest is a powerful algorithm that is widely used in the fields of medicine, agriculture, industry, and analysis of large data and studying its properties due to its ability to deal with large data with missing values in multiple dimensions [27], in the following Figure 5. explain Random forest algorithm components.

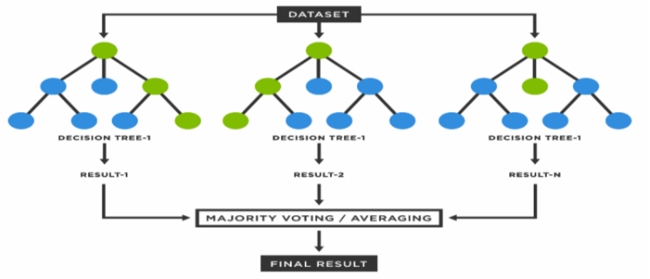

Figure 5. Random forest algorithm components [20].

3.2.7. Support Vector Machine (SVM) Algorithm

An important and powerful supervised machine learning algorithm used in many tasks including linear or nonlinear classification, regression, and even outlier detection tasks, but it is best suited for classification. The basic work of the support vector machine algorithm involves finding the best level among a number of features up to N that can separate groups of data points into different classes at multiple levels. The training is done until it obtains a level where the margin between the closest points of the different classes is as large as possible. What distinguishes support vector machines is a variety of tasks, such as image classification, text classification, handwriting identification, spam detection, face detection, gene expression analysis, and anomaly detection. With the ability to manage high-dimensional data and nonlinear relationships, support vector machines are adaptable and effective in a wide variety of applications [28,29], in the following Figure 6. explain Support vector machine algorithm components.

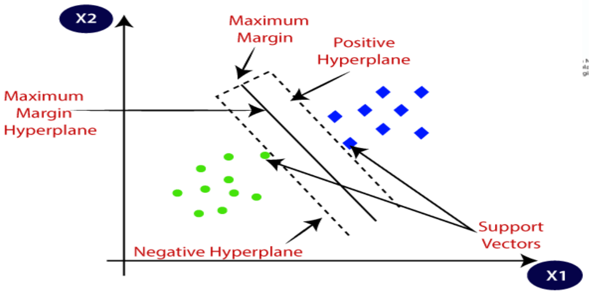

Figure 6. Support vector machine algorithm components [20].

3.3. Evaluation Metrics

In the course of the work, several performance evaluation metrics of the algorithms implemented in this paper were calculated. Usually, several classification metrics are used that are more important than others to evaluate the performance, effectiveness, and efficiency of the implemented model of machine learning algorithms, and they are as follows in Table 2 below [17,30,31]:

Table 2. The classification metrics

Metric Name	Definition	No. of Equation
Accuracy	(TP+TN)/(TP+FP+FN+TN)	(2)
F1-Score	2(Recall*Precision)/(Recall+ Precision)	(3)
Recall	TP/(TP+FN)	(4)
Precision	TP/(TP+FP)	(5)

TN: method predicted phishing websites as true phishing websites.

TP: method predicted normal as (legitimate websites), true predicted.

FN: method predicted the legitimate websites as (phishing websites), false predicted.

FP: method predicted phishing websites as (legitimate websites), false predicted

Also, statistical criteria are used in this paper to come up with a result through the principal component analysis of the data (PCA), which was added in Experiment C. The criteria used are R-squared error(RSE), root mean square error (RMSE), and mean absolute error (MAE) to improve the performance and effectiveness of the seven machine learning algorithms achieved. As shown in Table 3, the equations of the criteria are [21]:

Table 3. The statistical criteria

Criteria Name	Definition	No. of Eq.
RSE	RSE=1- ((Explaine-Dvariation)\(Total-Variation))	(6)
RMSE	$RMSE = \sqrt{\sum (y_{obs} - y_{pred})^2 / n}$	(7)
MAE	$MAE = \sum (\Vert y_{obs} - y_{pred} \Vert)/n$	(8)

METHODOLOGY

To accurately classify various types of URLs, this section describes our suggested research methodology for carrying out experiments using seven well-known machine learning algorithms: Adaptive Boosting Ada Boost Classifier, Decision Tree Classifier DT, Gradient Boosting GB, K-Nearest Neighbors Algorithm KNN, Logistic Regression LR, Random Forest Classifier RF, and Support Vector Machine Algorithm SVM. Our research approach uses the different steps in three experiment shown in Figure 7. to identify malicious URLs by following several mechanisms in implementing the detection of sites and these mechanisms are implemented on a set of experiments. The following subsections provide a detailed description of each experiment performed.

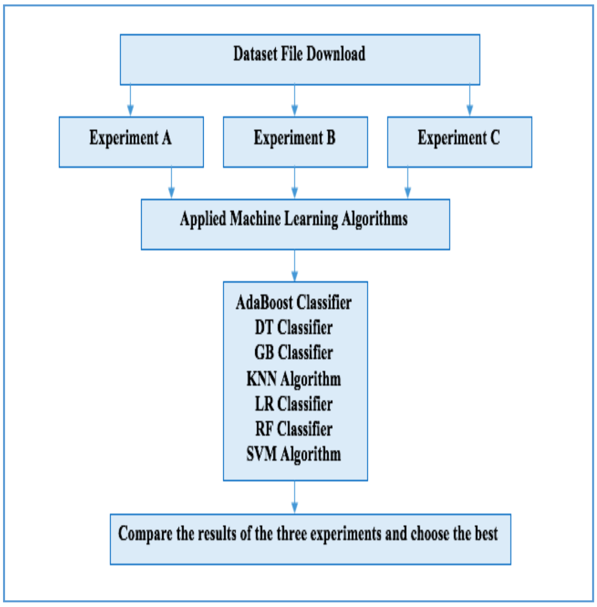

Figure 7. Procedure of work

Experiment A: In this experiment, the following steps were taken:

- Importing a data set from the Kaggle website
- Exploration data analysis (in terms of data that is needed, for example, in most cases, the number column is removed, and also detecting whether there is missing data and addressing that problem)
- Dividing the data into a training set and a test set (the data is divided in a certain proportion between training and testing)
- Feature engineering (choosing the features that affect the percentage of detecting the site as fraudulent or legal)
- Expanding the scope of features
- Building machine learning models

Experiment B: This experiment was conducted to lay the foundations for creating a website classifier using machine learning, starting from collecting data, pre-processing it in terms of missing values and ineffective values, cleaning the data such as removing stop words and extracting features, and applying ensemble methods that serve to benefit from the strengths of different models. We will create an ensemble model using a voting classifier to combine the predictions of its best models. Then, we will take a balanced scale between the upper and lower bounds of the data so that the algorithms work smoothly and efficiently and train the model to adjust the hyperparameters. The goal of this experiment is to reach rates in detecting and classifying fraudulent websites with high accuracy.

Experiment C: The methodology followed in the experiment in this paper is the following steps:

1- Download the data collected from the Kaggle website and import it into the program and prepare it.

2- Download the dataset that contains different cases of phishing sites that are of two types: suspicious or fraudulent and legitimate sites.

3- Follow an exploratory analysis of the downloaded data (EDA) to obtain a detailed breakdown of the structure and contents of this data.

4- Data visualization: Take a visualization of the data from several aspects

□ Check if there is missing data

□ Check if there is ineffective data

□ What is the ratio of data that represents phishing sites to the ratio of legitimate sites, the ratio of sites if there is not a large difference between the number of legitimate sites and phishing sites, an imbalance occurs in training the algoritwerthms.

5- Measuring features) Feature engineering & Feature scaling (and encoding categorical variables, where the data is pre-processed to ensure compatibility with machine learning algorithms

6- Facilitating the evaluation of the training model: We divide the dataset into 70 % training and 30% test sets to do this.

7- Implement various machine learning algorithms after performing principal component analysis PCA of data, such as Ada Boost Classifier, Decision Tree Classifier, Gradient Boosting, KNN Algorithm, Logistic Regression, Random Forest Classifier, and Support Vector Machine Algorithm to train predictive models.

8- Comparing result models:

The performance of the trained models is then evaluated using appropriate metrics, such as accuracy, precision, recall, and F1 score. Comparing the effectiveness of machine learning algorithms in different experiments in accurately classifying phishing sites and determining which one was the best.

5. Result Obtained and Discussion of The Results

In this section, we present the results we obtained in the three experiments whose procedures were mentioned in the previous section, where in the three experiments the data was split by 80% for training data and 20% for testing data in Tables 4, 5, 6, respectively. We analyze the performance of each machine learning algorithm using Accuracy, F1-Score, Recall and Precision metrics.

Table 4. Experiment A

Algorithm	Accuracy	F1-Score	Recall	Precision
Ada Boost	0.678	0.653	0.634	0.661
DT	0.872	0.856	0.869	0.877
GB	0.705	0.692	0.694	0.710
KNN	0.636	0.483	0.323	0.959
LR	0.518	0.554	0.652	0.537
RF	0.898	0.912	0.931	0.904
SVM	0.869	0.865	0.863	0.869

Table 5. Experiment B

Algorithm	Accuracy	F1-Score	Recall	Precision
Ada Boost	0.798	0.776	0.770	0.792
DT	0.893	0.889	0.896	0.905
GB	0.760	0.751	0.753	0.778
KNN	0.973	0.980	0.980	0.975
LR	0.928	0.939	0.954	0.928
RF	0.976	0.968	0.977	0.978
SVM	0.921	0.907	0.938	0.922

Table 6. Experiment C

Algorithm	Accuracy	F1-Score	Recall	Precision
Ada Boost	0.848	0.829	0.853	0.853
DT	0.966	0.964	0.964	0.963
GB	0.839	0.847	0.839	0.827
KNN	0.987	0.981	0.979	0.992
LR	0.996	0.990	0.991	0.989
RF	0.999	0.997	0.997	0.999
SVM	0.97	0.95	0.95	0.96

COMPARISON OF RESULTS WITH PREVIOUS WORKS

In this section, we have included Table 7, which compares the results of previous works related to the work presented in this paper. Through examination, we found that the results we obtained are better in all the algorithms achieved, unlike previous works that were good in one algorithm and weak in another.

Table 7. Comparing our best results with previous works

Ref.	Used Algorithm	Accuracy
[2]	DT, RF,SVM	97.11, 97.14, 96.51
[5]	LR, RF	94.93, 96.92
[6]	DT, Ada Boost, KNN, RF,	97.02, 93.24,95.67,97.98
[9]	Ada Boost, DT, GB, KNN, LR,RF, SVM	93.89, 99.12,95.39,96.46,92.00,99.06,95.10
[10]	RF	94.27
[28]	SVM	86.69
[32]	RF	99.31
[33]	Ada Boost, DT, KNN, LR,RF	87.00,100,96.70,79.00,95.00
Proposed Work	Ada Boost , DT, GB, KNN, LR, RF, SVM	84.8, 96.6, 83.9,98.7,99.6,99.9,97.00

CONCLUSION

The ease of access provided by the use of the Internet around the world has provided many positives for society and individuals, such as the speed of access to information, sending and receiving files, and enjoying many services. However, with this rapid development, many negatives have emerged, the most important of which are the risks that users are exposed to by ill-intentioned people who send links to fraudulent sites aimed at hacking devices and obtaining user data. Fraudsters exploit the lack of knowledge of users and create sites that make the user believe that they are legitimate and useful sites. In order to reduce these problems that occur, machine learning algorithms have emerged as a good and effective tool in combating malicious activities due to the ability of algorithms to detect phishing sites. In this paper, we presented seven models of machine learning algorithms, which are Ada Boost, Decision Tree, Gradient Boosting, KNN, Logistic Regression, Random Forest and Support Vector Machine with three experiments, each experiment specializes in an important part of data processing, but we reached results that are the best among the three experiments, which is Experiment C, in which we benefited from all the advanced positives in the previous experiments. It is important here to mention the possibility of conducting extensive training by combining machine learning and deep learning in the near future.

CONFLICT OF INTEREST

There was no conflict of interest declared by the authors.

REFRENCES

[1] Ali, W. (2017). Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection. International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications, 8(9). https://doi.org/10.14569/ijacsa.2017.080910

[2] Rishikesh Mahajan & Irfan Siddavatam. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 181(23), 45–47. https://doi.org/10.5120/ijca2018918026

[3] Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing Attacks: a recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3. https://doi.org/10.3389/fcomp.2021.563060

[4] Hamadouche, S., Boudraa, O., & Gasmi, M. (2024). Combining Lexical, Host, and Content-based features for Phishing Websites detection using Machine Learning Models. ICST Transactions on Scalable Information Systems. https://doi.org/10.4108/eetsis.4421

[5] Shaiba, H., Alzahrani, J. S., Eltahir, M. M., Marzouk, R., Mohsen, H., & Hamza, M. A. (2022). Hunger Search Optimization with Hybrid Deep Learning Enabled Phishing Detection and Classification Model. Computers, Materials & Continua/Computers, Materials & Continua (Print), 73(3), 6425–6441. https://doi.org/10.32604/cmc.2022.031625

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

[7] Volkamer, M., Renaud, K., Reinheimer, B., & Kunz, A. (2017). User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. Computers & Security, 71, 100–113. https://doi.org/10.1016/j.cose.2017.02.004

[8] Raj, M. M., & Jothi, J. a. A. (2022). Website phishing detection using machine learning classification algorithms. In Communications in computer and information science (pp. 219–233). https://doi.org/10.1007/978-3-031-19647-8_16

[9] Almomani, A., Alauthman, M., Shatnawi, M. T., Alweshah, M., Alrosan, A., Alomoush, W., Gupta, B. B., Gupta, B. B., & Gupta, B. B. (2022). Phishing website detection with semantic features based on machine learning classifiers. International Journal on Semantic Web and Information Systems/International Journal on Semantic Web and Information Systems, 18(1), 1–24. https://doi.org/10.4018/ijswis.297032

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

[11] Guptta, S. D., Shahriar, K. T., Alqahtani, H., Alsalman, D., & Sarker, I. H. (2022). Modeling hybrid Feature-Based phishing websites detection using machine learning techniques. Annals of Data Science, 11(1), 217–242. https://doi.org/10.1007/s40745-022-00379-8

[12] Jain, A. K., & Gupta, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Multimedia and Information Security, 2016(1). https://doi.org/10.1186/s13635-016-0034-3

[13] Jha, A. K., Muthalagu, R., & Pawar, P. M. (2023). Intelligent phishing website detection using machine learning. Multimedia Tools and Applications, 82(19), 29431–29456. https://doi.org/10.1007/s11042-023-14731-4

[14] https://www.kaggle.com/datasets/shubham9696/dectecting-phishing-website-using-machine-learning.

[15] Huang, G., Zhu, Q., & Siew, C. (2006). Extreme learning machine: Theory and applications. Neurocomputing, 70(1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

[16] FY, O., JET, A., O, A., O, H. J., O, O., & J, A. (2017). Supervised Machine Learning Algorithms: Classification and comparison. International Journal of Computer Trends and Technology, 48(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

[17] Thabit, Q. Q., Issa, B. A & Dawood, A. I. Detection of Pneumonia by Combining Transfer Learning Models with Data Regularization Based on Deep Learning Methods, Lecture Notes in Networks and Systems, In book: Proceedings of Data Analytics and Management, 129–145, April 2025. DOI: 10.1007/978-981-96-3358-6_11.

[18] Walker, K. W., & Jiang, Z. (2019). Application of adaptive boosting (Ada Boost) in demand-driven acquisition (DDA) prediction: A machine-learning approach. The Journal of Academic Librarianship, 45(3), 203–212. https://doi.org/10.1016/j.acalib.2019.02.013

[19] Thabit, Q. Q., Dawood, A. I. & Issa, B. A. Early Prediction of Stroke Based on Deep and Machine Learning Based on Different Datasets Mathematical Modelling of Engineering Problems, 11(12), 3499-3508. DOI: 10.18280/mmep.111228.

[20] Omari, K. (2023). Comparative Study of Machine Learning Algorithms for Phishing Website Detection. International Journal of Advanced Computer Science and Applications, 14(9), 417–425. https://doi.org/10.14569/ijacsa.2023.0140945

[21] Khan, M. S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University - Computer and Information Sciences, 34(8), 4773–4781. https://doi.org/10.1016/j.jksuci.2021.06.003

[22] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[23] N. Dahiya , B. Saini, and H.D. Chalak, "Gradient boosting-based regression modelling for estimating the time period of the irregular precast concrete structural system with cross bracing," Journal of King Saud University – Engineering Sciences, in press.

[24] Dewi, A. M. S. I., & Dwidasmara, I. B. G. (2020). Implementation of the K-Nearest Neighbor (KNN) algorithm for classification of obesity levels. JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), 9(2), 277. https://doi.org/10.24843/jlk.2020.v09.i02.p15

[25] Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A. C., Joseph, K. S., & Allen, V. M. (2018). Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC Pregnancy and Childbirth, 18(1). https://doi.org/10.1186/s12884-018-1971-2

[26] Thabit, Q. Q., Fahad, T. O., & Dawood, A. I. (2022). Detecting Diabetes Using Machine Learning Algorithms. IEEE. https://doi.org/10.1109/iiccit55816.2022.10010408

[27] Hutchinson, S., Zhang, Z., & Liu, Q. (2018). Detecting Phishing Websites with Random Forest. In Springer eBooks (pp. 470–479). https://doi.org/10.1007/978-3-030-00557-3_46

[28] Kulkarni, A., & L, L. (2019). Phishing Websites Detection using Machine Learning. International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications, 10(7). https://doi.org/10.14569/ijacsa.2019.0100702

[29] Mandadi, A., Boppana, S., Ravella, V., & Kavitha, R. (2022). Phishing website detection using machine learning. 2022 IEEE 7th International Conference for Convergence in Technology (I2CT). https://doi.org/10.1109/i2ct54291.2022.9824801

[30] Dawood A.I., Thabit Q.Q., Fahad T.O. (2023) Thyroid disease prediction with machine learning algorithms. Eurasian Res Bull 18:229–237. https://geniusjournals.org/index.php/erb/article/view/3777.

[31] Thabit Q.Q. (2023) Deep and machine learning for improving breast cancer detection. Eng Technol J 08(12). https://doi.org/10.47191/etj/v8i12.06

[32] Rao, R. S., & Pais, A. R. (2018a). Detection of phishing websites using an efficient feature-based machine learning framework. Neural Computing & Applications, 31(8), 3851–3873. https://doi.org/10.1007/s00521-017-3305-0

[33] Selvakumari, M., Sowjanya, M., Das, S., & Padmavathi, S. (2021). Phishing website detection using machine learning and deep learning techniques. Journal of Physics. Conference Series, 1916(1), 012169. https://doi.org/10.1088/1742-6596/1916/1/012169