A Comparative Analysis of Machine Learning Algorithms for Classification of Diabetes Utilizing Confusion Matrix Analysis

Maad M. Mijwil[*1] ID ✉, *Mohammad Aljanabi*[2] ID ✉

[1]Computer Techniques Engineering Department, Baghdad College of Economic Sciences University, Baghdad, Iraq.
[2]Department of Computer, College of Education, Aliraqia University, Baghdad, Iraq.
*Corresponding Author.

Abstract

Healthcare experts have been employing machine learning more and more in recent years to enhance patient outcomes and reduce costs. In addition, machine learning has been applied in various areas, including disease diagnosis, patient risk classification, customized treatment suggestions, and drug development. Machine learning algorithms can scrutinize vast quantities of data from electronic health records, medical images, and other sources to identify patterns and make predictions, which can support healthcare professionals and experts in making better-informed decisions, enhancing patient care, and determining a patient's health status. In this regard, the author opted to compare the performance of three algorithms (logistic regression, Adaboost, and naïve bayes) through the correct classification rate for diabetes prediction in order to ensure the effectiveness of accurate diagnosis. The dataset applied in this work is obtained from the Vanderbilt university institutional repository and is publicly available data. The study determined that three algorithms are very effective at prediction. Mainly, logistic regression and Adaboost had a classification rate above 92%, and the naive bayes algorithm achieved a classification rate above 90%.

Keywords: Algorithms, Classification, Confusion Matrix, Diabetes, Machine Learning.

Introduction

Diabetes is a dangerous disease that occurs as a result of an imbalance between blood sugar and the hormone insulin[1,2]. As a result of this disease, it is not possible to use sugar as it should, so blood sugar begins to spin freely in the blood. After a period of time, damage to the blood circulation in the body may occur and the disease may become fatal in the long run because it will seek to destroy the blood vessels and organs Fig 1[3,4]. So, diabetes is a malignant disease and is usually observed after stability in the body. There are a set of symptoms, including urinating more frequently than usual often at night, excessive thirst, sudden weight loss, severe hunger, lack of vision (blurred), skin problems (itchy skin), fatigue, and weakness. When insulin hormones are not secreted in a balanced manner, the body will make a serious effort to remove sugar from the blood, and thus the patient will feel exhausted and weak [5-7]. Weakness is one of the major symptoms of diabetes, as the body constantly loses water in the cells and is in intense activity, which can lead to fainting [8-10]. After that, the patient reaches the stage of continuous

and unreasonable depression, as well as tension, as this stage leads to the generation of sudden outbursts of anger, which is one of the most noticeable symptoms of diabetes. Moreover, there is another type of diabetes, known as hidden diabetes [11,12], where physicians and healthcare professionals define it as the first stage of diabetes, and the symptoms are unclear. Therefore, recognizing this disease at this stage and taking precautions will facilitate treatment and prevent the disease from developing and spreading in the body. The most notable signs of the onset of this stage and the most well-known are excess weight, continuous sweating, sleep disturbances, and the constant desire to eat sweets.
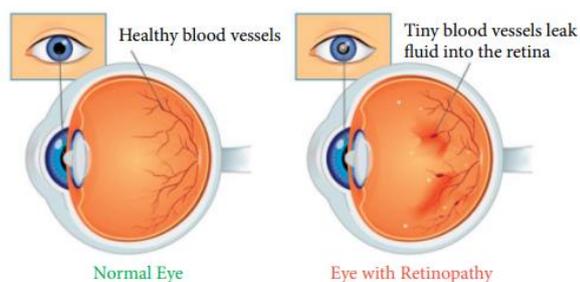
Figure 1. The eye of a healthy person vs. the eye of a person with diabetes [3].

In general, statistics indicate an increase in people with diabetes. For instance, in the International Diabetes Federation (IDF) annual report in 2021 about the statistics of this disease, as it was mentioned in this report that 537 million adults live with diabetes, and their ages are between 20-97 years; where 1 in 10 people worldwide are affected by this disease. These statistics are expected to increase by 2030 to 643 million people, and to more than 780 million people by 2045. Also, this report stated that this disease causes expenditures of at least 966 billion dollars annually, an increase of 316% over the past fifteen years [13]. In a study by Sharmaet

al. [14], they indicated statistically that the prevalence of type 2 diabetes had doubled in the United Kingdom from 2.39/1000 people in 2000 to 5.32/1000 people in 2013. The region's population is counted in the London Borough of Hackney (United Kingdom). It was found that 40% of them are from the ethnic minority group, as they discovered the highest prevalence of type 2 diabetes in the black and Asian groups [15-17]. Obesity is one of the most common causes of diabetes. Thus, physicians and healthcare workers advise people with this disease to constantly seek to lose excess weight in order to achieve a cure for diabetes or not to develop this dangerous disease for people who are not infected and have a hefty weight [18]. In addition, there are surgeries that help patients lose weight, and these operations have been widespread recently in many countries around the world. These are called bariatric surgeries. A study carried out by Rebelos et al. [19] evaluated the importance of these surgeries and their influence on weight loss for people with type 2 diabetes. The data of 312 obese people who underwent surgeries were analyzed, and their body weight was monitored for 1, 2, 3, 4 and 5 years in regular ambulatory visits to them and; the numbers are as follows 269, 312, 210, 151, 105, in each year, respectively. Their study found that there is a slight effect on reducing the level of disease after bariatric surgery. Understanding the mechanisms that lead to weight loss can prevent this disease. Abbott invented a device that monitors blood sugar levels in people with diabetes without needing backup finger prick tests and is approved by the US Food and Drug Administration [20]. There are many ways to monitor the level of sugar, glucose and insulin, as well as these mechanisms are linked with the smartphone through applications or personal computer Fig 2. Companies are constantly seeking to develop these devices and move away from traditional methods.

Figure 2. Blood sugar, glucose, and insulin monitoring devices [downloaded from Google].

Physicians and healthcare workers seek to take benefit of the capabilities of artificial intelligence

techniques in diagnosing medical datasets and making a proper decision regarding the patient's

health condition [21,22]. Artificial intelligence techniques have proven to have an influential role in health services in recent years due to insufficient human resources and the increased workload. However, the areas of application of these techniques in healthcare are very hourly and have the ability to analyze extensive data. Governments and organizations aspire to grow these techniques and rely on them in determining and making health administrative decisions. In addition, these techniques study datasets' behaviors and prepare performance reports for family medicine, hospitalization, surgery, diagnosis, prediction, etc. These techniques are gaining popularity and the satisfaction of many healthcare workers day by day. Considerable studies are constantly being conducted on the practices of these techniques and their effects in diagnosing medical datasets such as heart disease, COVID-19, cancer diseases, etc [23-27]. Concisely, artificial intelligence contributes to the development of health and administrative services, reducing the costs of surgeries and reducing human-caused errors. As mentioned above, one of the most important branches of artificial intelligence, which is machine learning, is used in the implementation of this work. Machine learning is an important part of artificial intelligence because it has a set of algorithms that have the ability to predict and diagnose all medical data and study its behavior and extract important features from it. The following are contributions to the current study:

- Obtaining a publicly available data set from Vanderbilt university, which is stored in the Vanderbilt biostatistics datasets site. This set includes data from 403 people diagnosed with and confirmed to have diabetes. A glycosylated hemoglobin >7.0 is often taken as a positive diagnosis of diabetes.
- In order to analyze this dataset, machine learning algorithms were used that have the ability to deal with this type of data. Notably, picked three different algorithms – logistic regression, Adaboost, and naïve Bayes – based on their practicality for this task and their established performance in similar contexts. This is to make a comparison between these algorithms through prediction practices and obtain the most profitable performance through metrics:

accuracy, recall, specificity, precision, F1-score, MCC, and AUC.
- Checking the performance of the applied models during the cross-validation value. Also, compare the practices of these algorithms through the effects of correct prediction and incorrect prediction of the testing data.

The rest of this article is organized as follows: Section Two briefly reviews some literature through the applied algorithms and the accuracy effect obtained. Section Three is divided into two parts; the first part discusses the dataset that is employed in this work is discussed, and the second part explains the algorithms that have been applied and their significance along with their major equations. Section Four describes the evaluation metrics used, evaluates the performance of the three models, mentions the results, and shows the most profitable performance. This article concludes with Section 5 and the last, with a set of conclusions about the behavior of these algorithms and conducting future work.

Literature Survey

This section reviews the literature on using machine learning algorithms to predict diabetes mellitus through dataset analysis. The review will concentrate on the algorithms employed and the consequent accuracy metrics. By highlighting these factors, we hope to gain a better understanding of the effectiveness of these algorithms in predicting diabetes mellitus. Kumar et al., [28] conducted a study using a Pima Indian diabetes dataset to evaluate the performance of multi-layer feedforward neural networks (MLFNN), random forests, and naïve Bayes algorithms. These algorithms are employed to handle missing values and enhance classification accuracy. The study indicated that the MLFNN algorithm performed satisfactorily, with a classification accuracy of over 84%. Mujumdar and Vaidehi conducted an investigation to predict diabetes utilizing a dataset of 800 records and ten attributes, including glucose level, number of pregnancies, skin thickness, blood pressure, insulin, BMI, job type, age and effect. To achieve this, they applied six machine learning algorithms: AdaBoost, gradient boost, random forest, logistic regression, extra trees, and linear discriminant analysis. After analyzing the results, they found that the AdaBoost classifier performed the best, reaching an accuracy of

over 98%. Their analysis emphasizes the possibility of machine learning algorithms in predicting diabetes and suggests practical insights into the effectiveness of distinct algorithms in this field [29]. Olisah et al. [30] conducted a study using machine learning algorithms to predict and diagnose patients with diabetes mellitus. They employed datasets from PIMA Indian and LMCH diabetes, involving six algorithms to the diagnostic diabetes patients' dataset. The study found that the 2GDNN model had the highest performance, reaching a classification accuracy of over 97%. In a study executed by researchers Khanam and Foo [31] on the application of machine learning algorithms in the prediction of diabetes. The dataset for this study includes more than 760 patients with nine traits taken from the UCI machine learning repository. These researchers were able to highlight the performance of the algorithms and get high accuracy of more than 88% achieved by the Artificial neural network algorithm. In addition, this study includes a complete analysis of the Pima Indian Diabetes (PID) dataset. Another study tested the ability of the learning algorithms to predict diabetes. Zou et al. [32] apply the application of three (decision tree, random forest and neural network) algorithms in predicting diabetes from a data group of more than 164,000 instances and 14 attributes collected from the hospital physical examination data in Luzhou, China. They concluded that the best performance is the random forest by achieving an accuracy of more than 80%. There is also literature on the use of machine learning to predict diabetes through a set of medical images, where Math and Fatimato conducted a study [33] using adaptive machine learning techniques to classify diabetic retinopathy images. They proposed a segment-based learning approach for detecting diabetic retinopathy and trained a convolutional neural network (CNN) to work with the dataset at a segment level. The results showed that the ROC curve exceeded 96%, and also more than 96% for both sensitivity and privacy, respectively. Other literature is summarized and compared with the current study in Section 4.

Experimental Setup

Intro to dataset

In this article, a diabetic dataset is collected from the Vanderbilt university institutional repository. The dataset comprises information on 19 variables associated with 403 patients. This dataset originally consisted of 1,046 individuals interviewed to study the prevalence of obesity, diabetes, and cardiovascular risk factors in central Virginia for African Americans. Through this survey, it was found that only 403 people are actually suffering from diabetes. The data utilized in this article is accessible to the public since the latest update, done by Frank E. Harrell Jr., was in December 2002 [34]. In addition, there are 15 out of 19 different variables associated with the factors mentioned above. In this study, 70% of the dataset has been set aside for use as training data, while the remaining 30% has been designated as testing data. The statistical characteristics of the dataset utilized are illustrated in Table 1. The dataset comprised a set of variables which encompassed age, weight, height, gender, cholesterol, glucose, high density lipoprotein, cholesterol/HDL ratio, glycosylated hemoglobin, location, frame, systolic blood pressure, diastolic blood pressure, waist, and hip ratio. Fig 3 illustrates a histogram of some of the variables. Moreover, the number of patients whose health status was assessed in this study is 403 patients, which includes 42% males (N = 169) and 58% females (N = 234). Fig 4 shows a significant correlation between the glucose and cholesterol variable. These variables (glucose and cholesterol) have a significant impact on human health and the incidence of diabetes. Wu et al. [35] conducted research analyzing the impact of β-carotene on humans with impaired glucose metabolism and rats with type 2 diabetes. The findings of the study suggest that elevated levels of glucose in the bloodstream can induce apoptosis of cardiomyocytes, cause dysfunction of the heart, inhibit the proliferation of fetal cardiomyocytes, and consequently contribute to the development of diabetes. Hodkinson et al. [36] carried out an analytical investigation to examine the influence of lipoprotein cholesterol on individuals who have heart disease or are vulnerable to cardiovascular disease. The study involved over 20,000 adults and encompassed 42 trials. According to the effects, cholesterol plays a

role in the development of heart disease and diabetes, highlighting its influence on human health. Fig 5 shows the correlations between all variables. According to the color scale, the correlation between the bars close to the red color is increased, while the correlation unhurriedly decreases in the colors towards blue. Machine learning algorithms are generally distinguished in their ability to deal with complex, high-dimensional data sets while integrating features and interactions among them to extract accurate results and determine the patient's condition. In addition, the effectiveness of machine learning algorithms in classifying diabetes depends on several factors, including features, data quality, and evaluation metrics employed. In other words, to get higher evaluation metrics effects, the data quality must be high with more features in order to train the machine learning algorithms sufficiently and improve performance.

Table 1. Dataset details

S. No.	Variables (Name & Label)	Range	Storage	Units	Mean ± Std.
1	Age	19-92	Double	Years	46.85±16.31
2	Weight	99-325 kg	Double	Pounds	177.59±40.34
3	Height	52-76 cm	Double	Inches	66.02±3.91
4	Gender	Male=1 Female=0	Integer	-	-
5	Cholesterol (*chol*)	78-443 mg/dl	Double	-	207.84±44.44
6	Stabilized Glucose (stab.glu)	48-385 mg/dl	Double	-	106.67±53.07
7	High Density Lipoprotein (hdl)	12-120 mg/dl	Double	-	50.44±17.26
8	Cholesterol/HDL Ratio (ratio)	1.5-19.3	Double	-	4.52±1.72
9	Glycosolated Hemoglobin (glyhb)	2.68-16.11	Double	-	5.58±2.24
10	Location	Buckingham=0, Louisa=1	Integer	-	-
11	Frame	Small=1, medium=2, large=3	Integer	-	-
12	Systolic Blood Pressure (bp.1s, 2s)	90-250 mmHg	Double	-	136.90±22.74
13	Diastolic Blood Pressure (bp.1d, 2d)	48-124 mmHg	Double	-	83.32±13.58
14	Waist	26-56cm	Double	Inches	37.90±5.72
15	Hip ratio	30-64	Double	Inches	43.03±5.65

All information is taken from the same repository
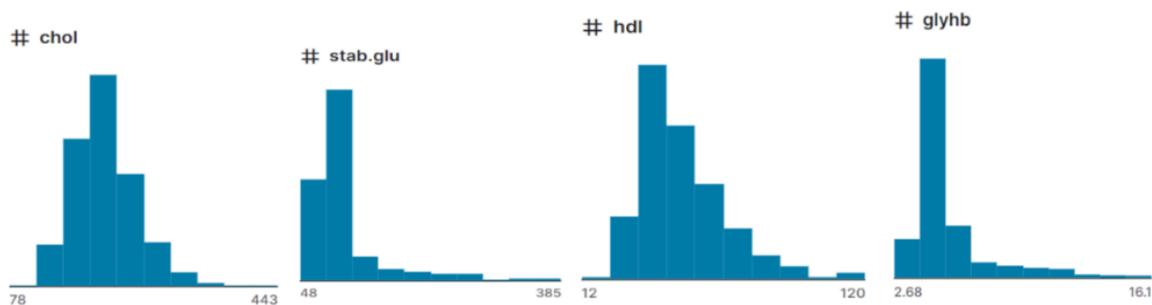
Figure 3. Histogram of some Variables
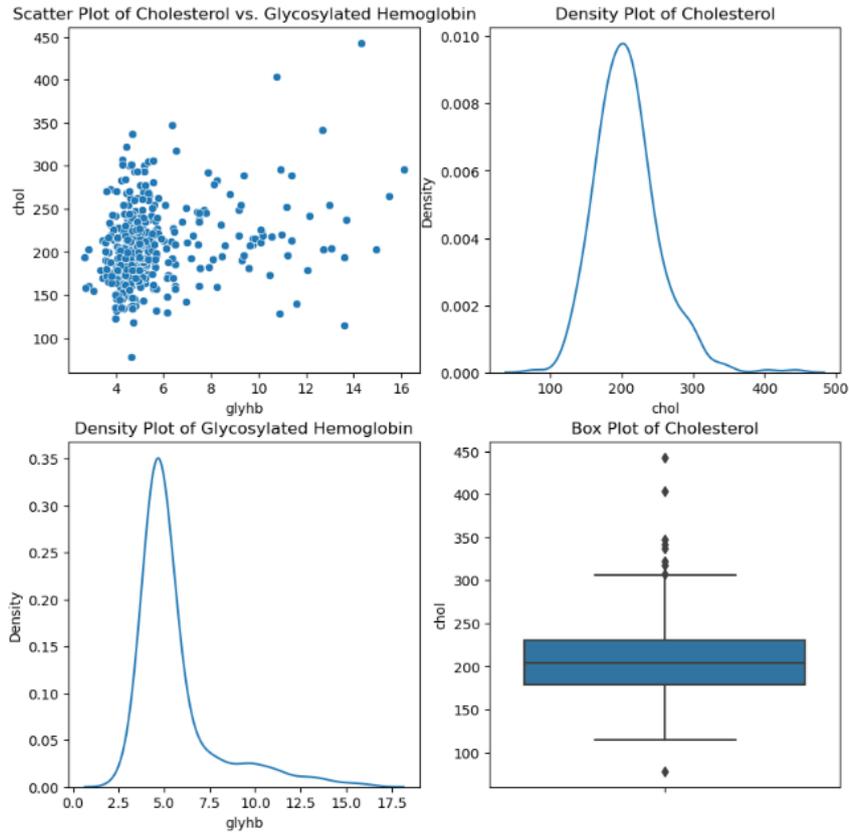
Figure 4. Correlation between the Glucose and cholesterol.
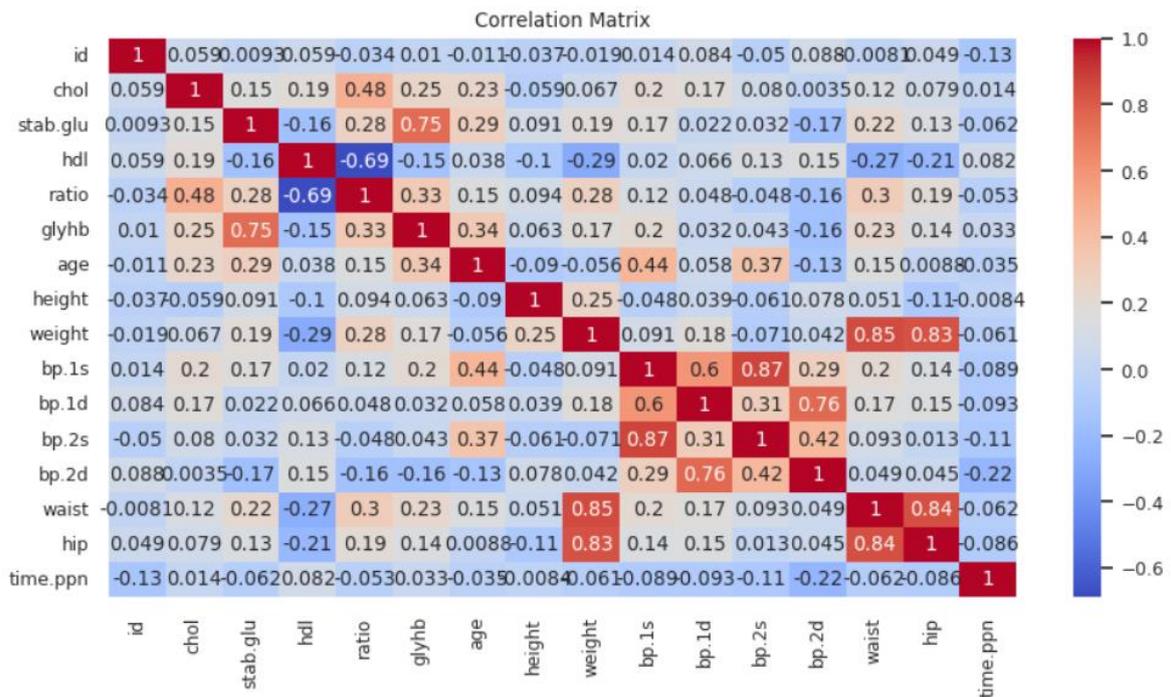
Figure 5. The correlations matrix between all variables.

The algorithms

Machine learning is commonly used to classify medical data as it can effectively manage vast amounts of data and extract pertinent information from it [37-40]. Medical data typically exists in different forms, such as images, texts, signals, and structured data. For this specific study, the focus is on text data, and three algorithms are operated to perform the training and testing processes and assess the performance of these algorithms. Directly, these algorithms will be described in a simplified manner as follows:

Logistic Regression algorithm

Logistic regression is one of the most popular machine learning algorithms concerned with binary classification. This algorithm involves predicting a binary outcome based on one or more input variables, which is why it was employed in this work. In diagnosing medical data, this algorithm is utilized to predict whether a patient will have a medical condition based on clinical variables and related variables. In other words, this algorithm is to model the probability of a binary outcome (e.g., the presence or absence of a disease) utilizing a logistic function. The logistic function takes a linear combination of input variables and plots it as a probability value between 0 and 1. This probability value can then be used to predict the binary outcome. Before employing the logistic regression algorithm on the medical dataset, specific preparation and processing steps must be carried out. These include addressing any missing values, standardizing input variables, and converting categorical variables into numerical ones. Once the data is prepared, this algorithm can be utilized to train on this data, and then the performance of the algorithm is measured based on a set of metrics such as accuracy, precision, recall, and F1-score. This algorithm assumes that the input variables are linearly related to the output variable and that the relationship is constant across the range of input values. In the event of any deviation from this assumption, the algorithm may behave more like a decision tree algorithm or neural network. Furthermore, this algorithm presupposes that the data are independent and similarly distributed, which may not be true for all medical datasets. During the work, the sigmoid function is used to convert the dot product of the input features and model weights (w) into an expected predicted value between 0 and 1. Eq. 1 represents the sigmoid function. While Eqs 2 and 3 are used to find the cost function and parameter update, respectively. Fig 6 shows a simplified schematic diagram of the logistic regression algorithm with a threshold value.

$$f(x) = \frac{1}{(1+e^{-z})} \ldots\ldots\ldots\ldots\ldots \quad 1$$

$$J(w) = \left(-1/m\right) * \left[y * \log\big(h(x)\big) + (1 - y)\log\big(1 - f(x)\big)\right] + \left(\frac{lambda}{2m}\right) * sum(w^2) \ldots\ldots \quad 2$$

$$w = w - alpha * dw \ldots\ldots\ldots.. \; 3$$

Where:

z is dependent variable, x_1, x_2, x_n are independent variable, and w_1, w_2, w_n are coefficient of slope.
$z = \; w_0x_0 + w_1x_1 + w_2x_2 + \cdots w_nx_n, \quad x_0 = 1$ & $w_0 = bais \; or \; intercept$
m is the no. of samples, y is the actual target, $lambda$ is regularization parameter.
$alpha$ is the learning rate. The model weights are updated iteratively until the cost function converges or the maximum number of iterations is achieved.
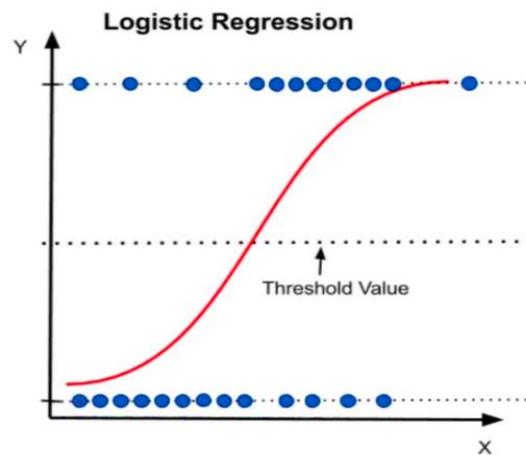
Figure 6. Logistic regression in binary dataset separation

Adaptive Boosting algorithm

AdaBoost (short for Adaptive Boosting) is one of the machine learning algorithms preferred for classifying the medical dataset. This algorithm is characterized by its group learning which combines many weak learners to create a strong classifier. A group of weak classifiers is trained on the medical dataset. These classifiers are in the form of decision

trees, logistic regression models, or any other type of classifier. Subsequently, a weight is assigned to each instance in the dataset. The weights are set to $^1/_n$, where n is the number of instances in the dataset. The algorithm estimates the error rate for each weak classifier. This rate is used to determine the weight of the classifier in the final group, and to recognize the classifiers with the lowest error rate and provide them with less weight. Besides, the weak classifier is trained on the updated dataset, and this strategy is repeated until all the weak classifiers are trained. Also, this algorithm aims to merge weak classifiers with one robust classifier by giving weights to each

weak classifier based on its performance situation. Algorithm 1 illustrates the steps of executing this algorithm with an explanation. This algorithm is applied in many medical applications due to its excellent ability to predict the probability of developing diseases, such as heart disease, breast cancer diagnosis, and Alzheimer's disease detection. That is why it is used in this work in the classification of diabetes. As with any machine learning algorithm, it is crucial to carefully evaluate the performance of the model on the specific medical dataset used, and to interpret the results in the context of medical knowledge and experience.

Algorithm 1: AdaBoost steps

Step 1: Input: Dataset $D = \{d_1, d_2, \ldots, d_n\}$,

Step 2: **Initialize weights**: set the weights of all instances in the training set to $w(i) = {^1/_n}$, where n is the number of instances in the dataset.

Step 3: For $t = 1 \to T$, where T is the no. of weak classifiers to be combined:

- **Train a weak classifier**: Train a weak classifier $h(t)$ on the training set using the current weights.
- **Compute the error**: Compute the weighted error of the weak classifier on the training dataset,
$$E(t) = sum(w(i) * I(y(i)! = h(t)(x(i))))$$
where $y(i)$ is the true label of the $i - th$ instance in the dataset, $x(i)$ is the feature vector of the $i - th$ instance, and $I()$ is the indicator function that returns 1 if the condition in the brackets is true and zero otherwise.
- **Compute the classifier weight**: Compute the weight of the current classifier,
$$alpha(t) = \frac{1}{2} * \log((1 - E(t)/E(t)))$$
- **Update the instance weights**: Update the weights of all instances in the training dataset,
$$w(i) = w(i) * \exp(-alpha(t) * y(i) * h(t)(x(i)))/Z$$
where Z is a normalization factor that ensures that the weights sum up to 1.

Step 4: **Output→ Compute the final classifier:** Compute the final classifier as a weighted combination of the weak classifiers, $H(x) = sign(sum(alpha(t) * h(t)(x)))$.

Step 5: **End** (the outcomes)

Notice: In the above equations, $h(t)(x)$ represents the prediction of the $t - th$ weak classifier on the input instance x, and $H(x)$ represents the prediction of the final classifier on the input instance x. The sign function returns +1 if the argument is positive or zero, and -1 otherwise.

Naïve Bayes algorithm

One of the most popular machine learning algorithms that is employed in data classification tasks, including the classification of medical datasets. This algorithm is based on Bayes' theorem, which is a fundamental concept in probability theory. Within the work steps, this algorithm supposes that the features are present in the dataset and are independent of each other, so this algorithm has been named naïve. The algorithm has demonstrated excellent performance across multiple applications and is extensively utilized in various scientific investigations, primarily for classifying medical datasets. Initially, this algorithm requires working on

a disaggregated medical records dataset. These records contain a set of features such as (age, gender, blood pressure, symptoms, etc.) and a corresponding label indicating the medical condition (such as disease, non-disease, or specific disease type). Next, the prior probability for each category is computed. This possibility is to show the occurrence of each category before looking at any data. For instance, to clarify, if a dataset contains 100 medical records and 30 of them are primarily classified as people with a particular disease, the prior probability for this category is 0.3. The next step is to calculate the conditional probability for each feature in each category. This step is done by calculating the

frequency of each feature in the records to belong to each category. To obtain the conditional probability, is accomplished by dividing this frequency by the total number of records with the category label. After getting the prior and conditional probabilities for each category and feature in the dataset, each class is calculated with a serious set of features. To achieve this, the method entails multiplying the category's prior probability with the conditional probabilities of every characteristic within that category. Subsequently, the result is standardized to guarantee that the probabilities sum up to 1. During the last stage, the newly generated record can be allocated to the category with the highest likelihood. Practically, this algorithm is proficient in the medical dataset's classification process. Regrettably, they do have limitations, particularly when the assumption of feature independence is not satisfied or when there are complex interactions between features. In these cases, more advanced machine learning techniques may be required. In this scenario, this algorithm has proven to have an outstanding performance in classifying small data. Eq 4 computes the probability

of a class given an instance or data point. Probability calculations are accomplished by Eq 5.

$$P(c|x) = P(x|c) * P(c) / P(x) \ldots\ldots\ldots 4$$

$$P(c|x) = P(c) * P(x_1|c) * P(x_2|c) * P(x_3|c) * \ldots * P(x_n|c) \ldots\ldots\ldots 5$$

Where:
$P(c|x)$ is the posterior probability of class c given the instance x.
$P(x|c)$ is likelihood which is the probability of instance x given class c.
$P(c)$ is the prior probability of class c.
$P(x)$ is the prior probability of instance x.
$P(x_i|c)$ is the conditional probability of feature x_i given class c.
n is the total number of features.
The main benefit of the above equations is to train the algorithm on the dataset and make predictions about new data points. This algorithm calculates the probability of each class for a unique data point and gives the class with the highest probability as the predicted class.

Results and Discussion

This section will cover the effects that were achieved by employing machine learning algorithms, specifically logistic regression, Adaboost, and naïve bayes. Subsequently, the performance of these algorithms will be considered and compared. The Python environment is used to implement this work, as it is a well-liked programming language in the field of machine learning. This language has a set of libraries (including TensorFlow, Keras, PyTorch, and Scikit-learn) that provide pre-built and customizable models for a wide range of machine learning tasks. Moreover, this language provides a wide range of tools, simple syntax, and cross-platform compatibility, allowing developers to create powerful machine learning applications quickly and efficiently. Using algorithms, the dataset is divided into 70% training data and 30% as testing data, with learning performed for each algorithm with no

scaling of the data. The compilation and validation process of each algorithm is repeated 10 times and the average values are taken into account. All algorithms using the training data utilize a cross-validation value ($k = 5$). The decision threshold is set at 0.5, which allows setting the sigmoidal output to classify the data into two classes, and the learning rate (λ) is 0.1. The training of the model stopped when the highest accuracy is achieved. After processing the dataset to eliminate duplicates and missing values, only 390 records remained, split into 273 records (70% training data) and 117 records (30% testing data) Fig. 7,8 illustrates a map of the working mechanism in implementation. In addition, Fig 9 illustrates that the confusion matrix was employed to identify four fundamental metrics for each algorithm. The primary objective of this matrix is to evaluate the algorithms' predictive capabilities.
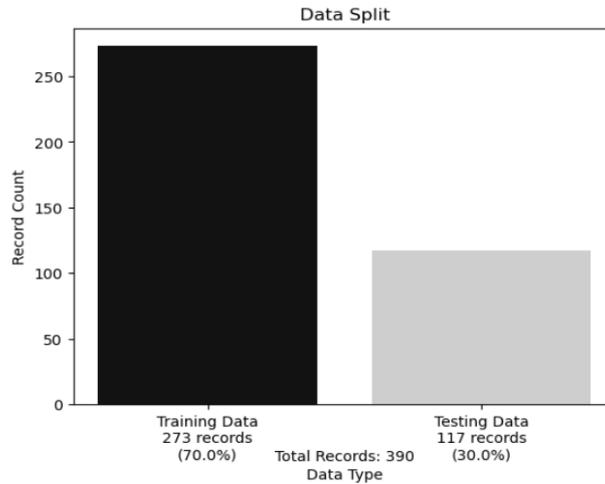
Figure 7. Split a dataset into 70% training and 30% testing data.
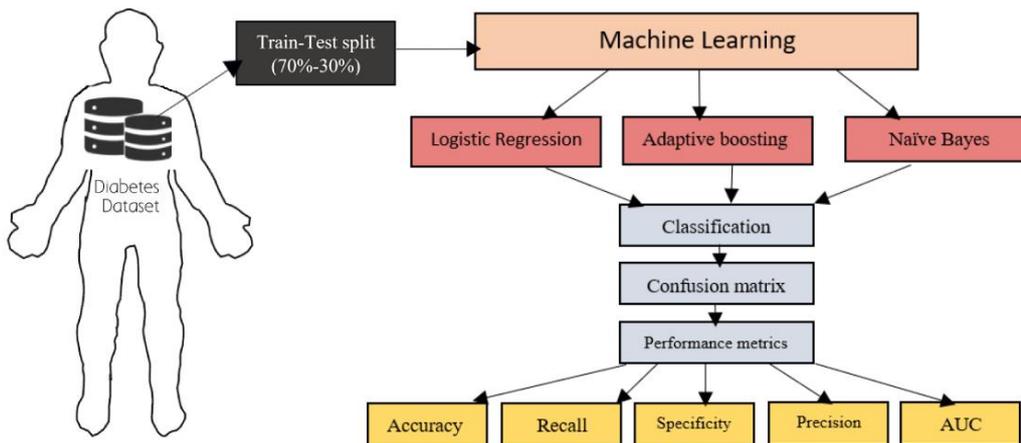
Figure 8. The working mechanism in implementation.

Classical mathematical formulas are utilized to consider the performance and predictive abilities of the algorithms by assessing their accuracy, recall, specificity, precision, F1-score, and MCC (See formulas 6 to 11). These formulas are essential tools for evaluating the performance of machine learning algorithms because they provide a quantitative method for evaluating the quality of predictions and comparing performance to determine which works better. Thus, the use of these formulas is crucial to ensure that the algorithms are both dependable and practical in their respective applications. Moreover, the AUC-ROC (Area Under the Receiver Operating Characteristic Curve) metric is used to measure the algorithm's ability to distinguish between positive and negative classes through the true positive rate (TPR) against the false positive rate (FPR) at

different threshold values. A high score on this metric means that the algorithm performs satisfactorily in distinguishing between positive and negative classes, while a low score indicates inadequate performance.

Figure 9. Confusion matrix form.

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \qquad 6$$

$$Recall = \frac{TP}{TN+TP} \qquad 7$$

$$Specificity = \frac{TN}{TN+FP} \qquad 8$$

$$Precision = \frac{TP}{TP+FP} \qquad 9$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad 10$$

$$Mcc = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(FN*TN)}} \qquad 11$$

Information:

Accuracy→ The percentage of models predicting correctly.

Recall→ The percentage of positive data is predicted to be positive.

Specificity→ The percentage of actual negatives accurately diagnosed.

Precision→ The percentage of data predictions as positives is correct.

F1-score→ The harmonic means of precision and recall.

Matthews correlation coefficient (MCC)→ is a calculation of the correlation between the predicted and actual classifications.

Formulas 11 and 12 are mainly employed to find the false positive ratio (FPR) and false negative ratio (FNR):

$$FPR = 1 - \left(\frac{specificity}{100}\right) \qquad 11$$

$$FNR = 1 - \left(\frac{sensitivity}{100}\right) \qquad 12$$

After conducting the tests, it was noticed that the algorithms' practices lead to excellent performance through the accuracy metric, which is the motivation for evaluating the performance of machine learning algorithms. The logistic regression algorithm demonstrated accurate classification by achieving an accuracy rate of 92.5%. Similarly, the AdaBoost algorithm achieved an accuracy rate of 92.5%, which was very close to the logistic regression algorithm. On the other hand, the Naïve Bayes algorithm exhibited very good accuracy, surpassing 90.5%, which is the best score. However, its performance did not match that of the logistic regression and AdaBoost algorithms. Logistic regression was able to correctly predict 108 individuals out of 117 of the testing data. The Naïve Bayes algorithm correctly predicted 105 individuals out of 117 of the testing data. The AdaBoost algorithm achieved correct predictions for 106 individuals out of 117 in the testing data, with a maximum of ten estimators chosen for the algorithm. The algorithms are trained to stop when the highest accuracy was attained. Regarding the algorithms' performance based on the AUC metric, the logistic regression algorithm displayed remarkable capability in distinguishing between classes. It earned a score of 85%, which was the highest among the algorithms based on this metric. The AdaBoost algorithm also proved its proficiency to distinguish between categories as it achieved a score of 80%. On the other hand, the naïve bayes algorithm exhibited the lowest performance in this regard, with a score of 70%. Now, all algorithm results will be displayed in the form of tables and figures. Figs 10 to 12 exhibit each algorithm's confusion matrix and the AUC. Table 2 provides an overview of the performance evaluation of all algorithms employed in this study, whereas Table 3 displays the correct and incorrect predictions of each algorithm. Finally, Fig 13 illustrates a comparison of the algorithms' performance based on Accuracy and AUC metrics. Table 4 compares the current study findings with a set of published literature.
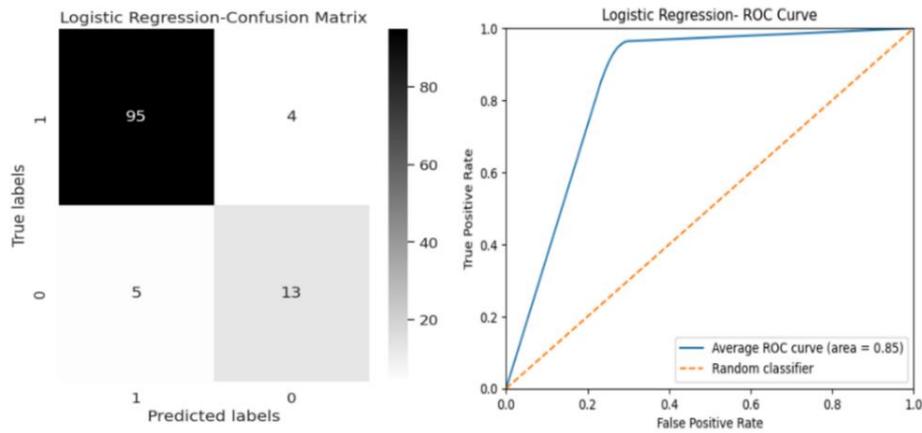
Figure 10. Logistic Regression: (a) confusion matrix and (b) ROC.
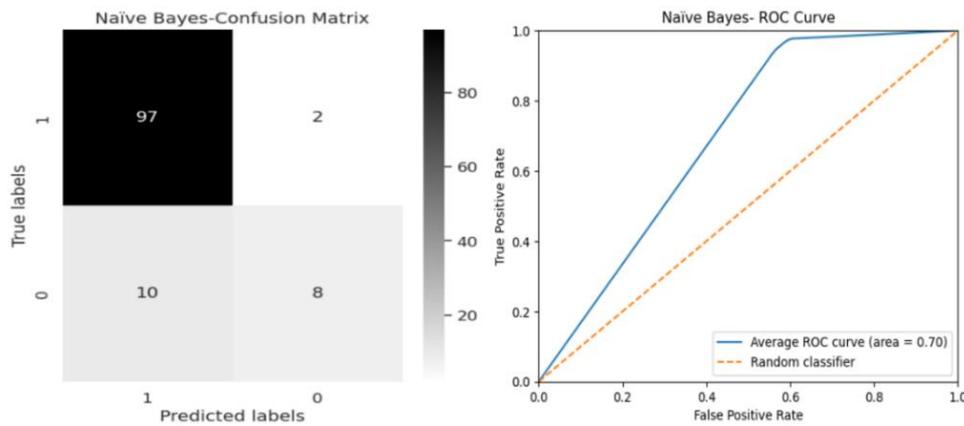
Figure 11. Naïve Bayes: (a) confusion matrix and (b) ROC.
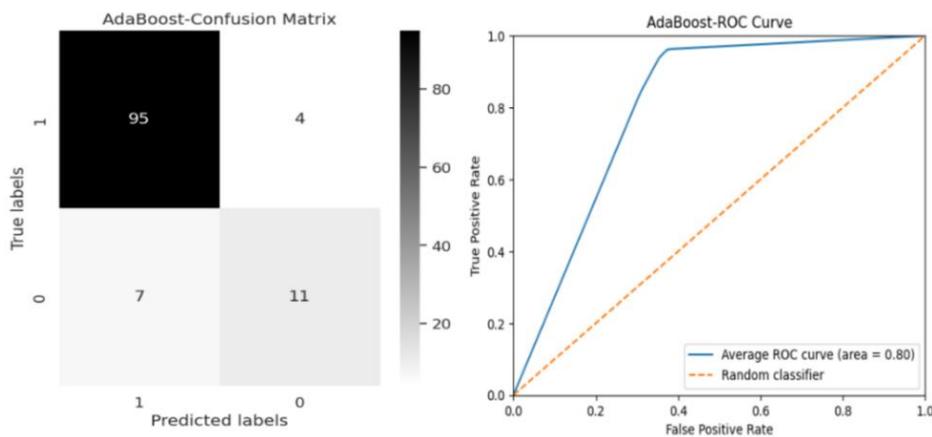
Figure 12. AdaBoost: (a) confusion matrix and (b) ROC.

Table 2. Performance results of all algorithms applied.

Evaluation metrics							
Algorithms	**Accuracy**	**Recall**	**Specificity**	**Precision**	**F1-score**	**MCC**	**AUC**
Logistic regression	**92.5%**	**95.9%**	**72.2%**	**95%**	**95.4%**	**77.6%**	**85%**
Naïve Bayes	90.5%	97.9%	44.4%	90.7%	94.2%	67.8%	71%
AdaBoost	92.5%	95.9%	61.1%	93.1%	94.5%	72.9%	80%
Boldface values indicate highest performance group.							

Through the table above, it is clear that the practices of the linear regression algorithm give effects in all metrics that are the highest among all other algorithms except for the recall metric, which denotes the percentage of data samples that are perfectly recognized as belonging to the positive category, as it achieved a result of more than 95 %, which is a remarkable effect. Still, it is not the highest compared to the Naïve Bayes algorithm, which gained a recall level of more than 97%.

Table 3. Correct and incorrect prediction for each algorithm

Algorithms	Prediction	
	Correct	**Incorrect**
Logistic regression	108	9
Naïve Bayes	105	12
AdaBoost	106	11
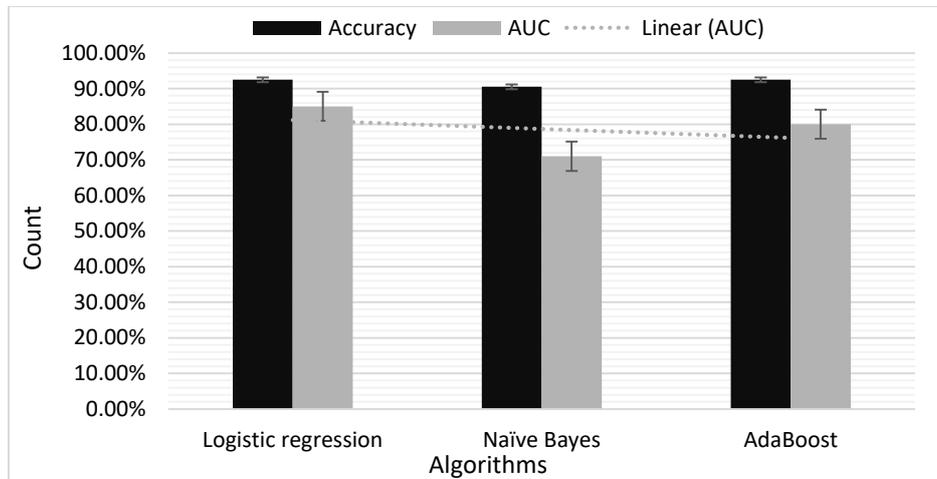
Figure 13. Algorithms performance by Accuracy and AUC metrics.

Table 4. Comparison of the proposed algorithms with other works.

Authors	Year	Algorithms	Best effect	Accuracy	Recall	Precision	Specificity	F1-score	Dataset
Zou et al., [32]	2018	DT, RF, and NN	RF	88.5%	97.4%	-	81.4%	-	Hospital
Maniruzzaman et al., [41]	2018	RF, LR, MI, PCA, AV, and FDR	RF	92.2%	95.9%	-	79.7%	-	UCI Repository
Alam et al., [42]	2019	ANN, RF, and K-MC	ANN	75.7%	75.0%	-	29.0%	-	UCI Repository
Sivaranjani et al., [43]	2020	SVM+PCA	SVM	81.4%	82%	83%	-	-	Kaggle
Hasan et al., [44]	2020	K-NN, DT, RF, AdaBoost, NB, and XGBoost	AdaBoost	95%	78.9%	-	93.4%	-	UCI Repository
Nadeem et al.,[45]	2021	SVM and ANN	Fusion of SVM-ANN	94.6%	89.2%	94.1%	97.3%	-	NHANES and Kaggle
Laila et al.,[46]	2022	AdaBoost, Bagging, and RF	RF	97.1%	97.1%	97.1%	-	97.1%	UCI Repository
Kaur and Kumari [47]	2022	SVM-linear, RBF, k-NN, ANN, and MDR.	SVM	89%	87%	88%	-	87%	UCI Repository
This work		LR, NB, and AdaBoost	LR	92.5%	95.9%	72.2%	95%	95.4%	Vanderbilt university repository

Conclusion

The primary purpose of this article is to reach the performance of the three algorithms and to determine the most suitable performance. The effects are reached based on two main metrics accuracy and AUC. The results indicate that the logistic regression algorithm achieved the highest level of accuracy. At the same time, the AdaBoost algorithm also showed a powerful performance with accuracy that is comparable to the logistic regression algorithm. As for the logistic regression algorithm, it had a high ability to distinguish between categories through the AUC metric, which is considered the highest compared to other algorithms. However, the accuracy metric remains the most important in machine learning, which limits the algorithm's ability to make correct predictions in patient data and assist in decision-making. Based on this analysis, it can be concluded that the logistic regression algorithm delivers the most acceptable level of performance. Indeed, all metrics analyzed of the logistic regression algorithm in this study are the most elevated compared to other algorithms. In contrast, the naïve Bayes algorithm showed the poorest performance among the algorithms, despite its high recall. This is evident from its low accuracy and AUC, indicating its inflexible ability to differentiate between classes. In the future, other models will be used to study their behavior in analyzing diabetic datasets and reaching performance between them.

Nomenclatures

Acronym	Description
AI	Artificial intelligence
AUC	Area under the ROC Curve
CART	Classification and Regression Tree
DT	Decision Tree
K-NN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LR	Linear Regression
LR	Logistic Regression
ML	Machine Learning
NB	Naïve Bayes
RF	Random Forest
ROC Curve	Receiver Operating Characteristic Curve
SVM	Support Vector Machine
UCI	University of California Irvine

Author's Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript by citing their references.
- The authors have signed an animal welfare statement.
- Ethical Clearance: The project was approved by the local ethical committee in Baghdad College of Economic Sciences University.

Author's Contribution Statement

M. M. M. and M. A. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Baghdad Science Journal

References

1. Ali A, Al-Awkally N M, Ahmer A, Tariq S, Mumtaz T, Shahbaz H, et al. Diabetes Mellitus and Renal Failure Effect on Intestinal Insulin. Brilliance: Res Artif Intell. 2022; 2(3): 102-106. https://doi.org/10.47709/brilliance.v2i3.1600

2. Sunday H G, Sadia A H, Ojo O G. Mechanisms of Diabetes Mellitus Progression: A Review. J Diab Neph Diab Mang. 2022; 1(1):1-5.

3. Mahesh T R, Kumar D, Kumar V V, Asghar J, Bazezew B M, Natarajan R, Vivek V. Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease. Comput Intell Neurosci. 2022; 2022(4451792): 1-9. https://doi.org/10.1155/2022/4451792

4. Laurent S, Agabiti-Rosei C, Bruno R M, Rizzoni D. Microcirculation and Macrocirculation in Hypertension: A Dangerous Cross-Link?. Hypertension. 2022; 79(3): 479–490. https://doi.org/10.1161/HYPERTENSIONAHA.121.17962

5. Yang K, Wang Y, Li Y, Chen Y, Xing N, Lin H, et al. Progress in the treatment of diabetic peripheral neuropathy. Biomed. Pharmacother. 2022; 148: 1-10. https://doi.org/10.1016/j.biopha.2022.112717

6. Chao A M, Wadden T A, Clark J M, Hayden K M, Howard M J, Johnson K C, et al. Changes in the Prevalence of Symptoms of Depression, Loneliness, and Insomnia in U.S. Older Adults With Type 2 Diabetes During the COVID-19 Pandemic: The Look AHEAD Study. Diabetes Care. 2022; 45(1): 74–82. https://doi.org/10.2337/dc21-1179

7. Harbuwono D S, Handayani D O T L, Wahyuningsih E S, Supraptowati N, Ananda, Kurniawan F, et al. Impact of diabetes mellitus on COVID-19 clinical symptoms and mortality: Jakarta's COVID-19 epidemiological registry. Diabetes Prim Care. 2022; 16(1): 65-68. https://doi.org/10.1016/j.pcd.2021.11.002

8. Iacopi E, Pieruzzi L, Riitano N, Abbruzzese L, Goretti C, Piaggesi A. The Weakness of the Strong Sex: Differences Between Men and Women Affected by Diabetic Foot Disease. Int J Low Extrem Wounds. 2021; 22(1): 19-26. https://doi.org/10.1177/1534734620984604

9. Gollapalli M, Alansari A, Alkhorasani H, Alsubaii M, Sakloua R, Alzahrani R, et al. novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM. Comput Biol Med. 2022; 147: 1-12. https://doi.org/10.1016/j.compbiomed.2022.105757

10. Giha H A, Sater M S, Alamin O A O. Diabetes mellitus tendino-myopathy: epidemiology, clinical features, diagnosis and management of an overlooked diabetic complication. Acta Diabetol. 2022; 59:871–883. https://doi.org/10.1007/s00592-022-01860-9

11. Marques M, López-Sánchez P, Tornero F, Gargantilla P, Maroto A, Ortiz A, Portolés J. The hidden diabetic kidney disease in a university hospital-based population: a real-world data analysis. Clin Kidney J. 2022; 15(10):1865–1871. https://doi.org/10.1093/ckj/sfac100

12. Yunka T T, Mogas S B, Zawdie B, Tamiru D, Tesfay Y. The Hidden Burden of Diabetes Mellitus in an Urban Community of Southwest Ethiopia. Diabetes Metab Syndr Obes: Targets Ther. 2020; 13: 2925–2933. https://doi.org/10.2147/DMSO.S269386

13. Diabetes reports. IDF Diabetes Atlas, 2021. [Cited 2023 Feb 15]. Available from: https://diabetesatlas.org/

14. Sharma M, Nazareth I, Petersen I. Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. BMJ Open. 2016; 6: 1-11. http://dx.doi.org/10.1136/bmjopen-2015-010210

15. Census 2021. Office for National Statistics. Available from: https://www.ons.gov.uk/census

16. City and Hackney Public Health Team Joint Strategic Needs Assessment: Adult Health and Diabetes, 2021. [Cited 2023 Feb 15]. https://www.cityhackneyhealth.org.uk/wp-content/uploads/2018/12/Diabetes-1.pdf

17. Marples O, Resca L, Plavska J, Hassan S, Mistry V, Mallik R, Brown A. Real-World Data of a Group-Based Formula Low Energy Diet Programme in Achieving Type 2 Diabetes Remission and Weight Loss in an Ethnically Diverse Population in the UK: A Service Evaluation. Nutrients. 2022; 14(15): 1-15. https://doi.org/10.3390/nu14153146

18. Brown A, McArdle P, Taplin J, Unwin D, Unwin J, Deakin T, et al. Dietary strategies for remission of type 2 diabetes: A narrative review. J Hum Nutr Diet. 2022; 35(1): 165-178. https://doi.org/10.1111/jhn.12938

19. Rebelos E, Moriconi D, Honka M, Anselmino M, Nannipieri M. Decreased Weight Loss Following Bariatric Surgery in Patients with Type 2 Diabetes. Obes Surg. 2022; 33: 179–187. https://doi.org/10.1007/s11695-022-06350-z

20. Johnson L A. FDA approves first blood sugar monitor without finger pricks, 2017. [Cited 2023 Feb 15]: https://www.statnews.com/2017/09/28/fda-approves-blood-sugar-monitor-without-finger-pricks/

21. Mijwil M M, Aljanabi M, ChatGPT. Towards Artificial Intelligence-Based Cybersecurity: The Practices and ChatGPT Generated Ways to Combat Cybercrime. Iraqi J Comput Sci Math. 2023; 4(1): 65-70. https://doi.org/10.52866/ijcsm.2023.01.01.0019.

22. Ismael A M, Şengür A. Deep learning approaches for COVID-19 detection based on chest X-ray images.

Expert Syst Appl. 2021; 164:114054. https://doi.org/10.1016/j.eswa.2020.114054

23. Mijwil M M. Deep Convolutional Neural Network Architecture to Detection COVID-19 from Chest X-ray Images. Iraqi J Sci. 2023; 64(5):2561-2574. https://doi.org/10.24996/ijs.2023.64.5.38

24. Nassif A B, Mahdi O, Nasir Q, Talib M A, and Azzeh M. Machine learning classifications of coronary artery disease. Int Joint Symp Artif Intell Nat Lang Process. 2018:1–6. http://doi.org/10.1109/iSAI-NLP.2018.8692942.

25. Ashraf S, Alfandi O, Ahmad A, Khattak A M, Hayat B, Kim K H, Ullah A. Bodacious-Instance Coverage Mechanism for Wireless Sensor Network. Wirel Commun Mob Comput. 2020; 2020(8833767): 1-11. https://doi.org/10.1155/2020/8833767

26. Ahmad A, Ullah A, Feng C, Khan M, Ashraf S, Adnan M, Nazir S, Khan H U. Towards an Improved Energy Efficient and End-to-End Secure Protocol for IoT Healthcare Applications. Secur Commun Netw. 2020; 2020(8867792): 1-10. https://doi.org/10.1155/2020/8867792

27. Shukur B S, Mijwil M M. Involving Machine learning as Resolutions of Heart Diseases. Int J Electr Comput Eng. 2023; 13(2): 2177-2185. http://doi.org/10.11591/ijece.v13i2.pp2177-2185

28. Kumar S, Bhushan B, Singh D, Choubey D K. Classification of Diabetes using Deep Learning. Int Confer Commun Signal Process. 2020: 1-6, Chennai, India. http://doi.org/10.1109/ICCSP48568.2020.9182293

29. Mujumdar A, Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. Procedia Comput Sci. 2019; 165: 292-299. https://doi.org/10.1016/j.procs.2020.01.047

30. Olisah C C, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Programs Biomed. 2022; 220: 1-12. https://doi.org/10.1016/j.cmpb.2022.106773

31. Khanam J J, Foo S Y. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021; 7(4): 432-439. https://doi.org/10.1016/j.icte.2021.02.004

32. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus with Machine Learning Techniques. Front Genet. 2018; 9(515): 1-10. https://doi.org/10.3389/fgene.2018.00515

33. Math L, Fatima R. Adaptive machine learning classification for diabetic retinopathy. Multimed Tools Appl. 2020; 80: 5173–5186. https://doi.org/10.1007/s11042-020-09793-7

34. Diabetic dataset. Vanderbilt biostatistics datasets site, 2023. [Cited 2023 Feb 15]. https://hbiostat.org/data/

35. Wu J, Zhou Y, Hu H, Yang D, Yang F. Effects of β-carotene on glucose metabolism dysfunction in humans and type 2 diabetic rats. Acta Mater Med. 2022; 1(1): 138-153. https://doi.org/10.15212/AMM-2021-0009

36. Hodkinson A, Tsimpida D, Kontopantelis E, Rutter M K, Mamas M A, Panagioti M. Comparative effectiveness of statins on non-high density lipoprotein cholesterol in people with diabetes and at risk of cardiovascular disease: systematic review and network meta-analysis. BMJ. 2022; 376(e067731): 1-13. https://doi.org/10.1136/bmj-2021-067731

37. Lee C S, Baughman D M, Lee A Y. Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images. Ophthalmol. Retina. 2017; 1(4): 322-327. https://doi.org/10.1016/j.oret.2016.12.009

38. Sidey-Gibbons J A M, Sidey-Gibbons C J. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019; 19(64):1-18. https://doi.org/10.1186/s12874-019-0681-4

39. Khanday A M U D, Khan Q R, Rabani S T. Detecting Textual Propaganda Using Machine Learning Techniques. Baghdad Sci J. 2021; 18(1),:199-209. https://doi.org/10.21123/bsj.2021.18.1.0199

40. Abdulmajeed A A, Tawfeeq T M, Al-jawaherry M A. Constructing a Software Tool for Detecting Face Mask-wearing by Machine Learning. Baghdad Sci J, 2022; 19(3):642-653. https://doi.org/10.21123/bsj.2022.19.3.0642

41. Maniruzzaman, Rahman J, Al-MehediHasan Suri H S, Abedin M, El-Baz A, Suri J S. Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. J Med Syst. 2018; 42(92): 1-17. https://doi.org/10.1007/s10916-018-0940-7

42. Alam T M, Iqbal M A, Ali Y, Wahab A, Ijaz S, Baig T I, Hussain A, Malik M A, Raza M M, Ibrar S, Abbas Z. A model for early prediction of diabetes. Inform Med Unlocked. 2019; 16: 1-6. https://doi.org/10.1016/j.imu.2019.100204.

43. Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. Int Conf Adv Comput Commun Syst. 2021: 1-6, Coimbatore, India. https://doi.org/10.1109/ICACCS51430.2021.9441935

44. Hasan K, Alam A, Das D, Hossain E, Hasan M. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access. 2020; 8: 76516-76531. https://doi.org/10.1109/ACCESS.2020.2989857

45. Nadeem M W, Goh H G, Ponnusamy V, Andonovic I, Khan M A, Hussain M. A Fusion-Based Machine Learning Approach for the Prediction of the Onset of Diabetes. Healthcare. 2021; 9(10): 1-16. https://doi.org/10.3390/healthcare9101393

46. Laila U E, Mahboob K, Khan A W, Khan F, Taekeun W. An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. Sensors. 2022; 22(14): 1-15.https://doi.org/10.3390/s22145247

47. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inform. 2022; 18(1/2): 90-100. https://doi.org/10.1016/j.aci.2018.12.004

تحليل مقارن لخوارزميات التعلم الآلي لتصنيف مرض السكري باستخدام تحليل مصفوفة الارتباك

معد محسن مجول[1] ومحمد خميس الجنابي[2]

[1]قسم هندسة تقنيات الحاسوب، كلية بغداد للعلوم الاقتصادية الجامعة، بغداد، العراق.
[2]قسم علوم الحاسبات، كلية التربية، الجامعة العراقية ، بغداد، العراق.

الخلاصة

يستخدم العاملين في الرعاية الصحية التعلم الآلي أكثر فأكثر في السنوات الأخيرة لتعزيز نتائج المرضى وخفض التكاليف عليهم. بالإضافة إلى ذلك، تم تنفيذ التعلم الآلي في مجالات مختلفة، بما في ذلك تشخيص الأمراض، وتصنيف مخاطر المرضى، واقتراحات العلاج المخصصة، وتطوير الأدوية. يمكن لخوارزميات التعلم الآلي أن تفحص كميات هائلة من البيانات من السجلات الصحية الإلكترونية، والصور الطبية، ومصادر أخرى لتحديد الأنماط والتنبؤات، والتي يمكن أن تدعم المتخصصين في الرعاية الصحية والخبراء في اتخاذ قرارات مستنيرة، وتعزيز رعاية المرضى، وتحديد حالة المريض الصحية. في هذا الصدد، اختار المؤلف مقارنة أداء ثلاث خوارزميات (الانحدار اللوجستي ، Adaboost ، بايزالساذجة) من خلال معدل التصنيف الصحيح للتنبؤ بمرض السكري من أجل ضمان فعالية التشخيص الدقيق. تم الحصول على مجموعة البيانات المطبقة في هذا العمل من مستودع مؤسسي بجامعة فاندربيلت وهي بيانات متاحة للجميع. استنتجت الدراسة أن ثلاث خوارزميات فعالة للغاية في التنبؤ. بشكل أساسي، كان معدل تصنيف الانحدار اللوجستي و Adaboost أعلى من 92٪ ، وحققت خوارزمية بايز الساذجة معدل تصنيف أعلى من 90٪.

الكلمات المفتاحية: الخوارزميات ، التصنيف، مصفوفة الارتباك، مرض السكري، التعلم الآلي.