ARTICLE

JOURNAL`S UNIVERSITY OF BABYLON FOR ENGINEERING SCIENCES (JUBES)

مـــجلـة جـــامعة بـــابـل للعلـــوم الهندسية

Vol. 32, No. 3. \ 2024 ISSN: 2616 - 9916

Classifying of Diabetes Disease in Women Based on Support Vector Machine and Random Forest Algorithms

Randa shaker Abd-Alhussain[1] **Zina Abd Al Hussein Saleh**[2]

[1] *Presidency of the University of Babylon, Department of Missions and Cultural Relations, University of Babylon, Babylon, Iraq*

Email: randashaker1984@gmail.com

[2] *Department of Studies and Planning, Presidency of the University of Babylon, Babylon, Iraq*

Email: zina.badi@uobabylon.edu.iq

Received:	13/2/2024	Accepted:	7/5/2024	Published:	13/6/2024

Abstract

Diabetes is a dangerous illness. It is indicated by blood sugar levels and/or glucose levels. Diabetes is a chronic illness that can lead to global health crisis, but there are things that can done to help control these crises. The primary energy source that get from the food on a daily basis in people with diabetes is blood sugar, or glucose. The insulin hormone created by the pancreas, assisting in the uptake of glucose from the blood into the cells so that it can be utilized as an energy source for daily tasks. Glucose remains in the blood when the body produces insufficient amounts of insulin, which can cause a number of health issues like heart attacks and strokes. There are numerous forms of diabetes, the most prevalent being type1 and type2. Type1 is typically diagnosed in children and young adults, whereas type 2 is typically found in middle age or older population. The objective of the project is to develop a system that can accurately classify patients as either diabetic or not by combining the results of various machine learning techniques with algorithms like Support Vector Machine and Random Forest. Machine learning is a scientific field where machine learning is derived from human experience. The model that best predicts diabetes is the one whose accuracy as a percentage is determined by calculating the accuracy of the model using each method. According to the experimental findings, RF and SVM have accuracy of 100% and 89% respectively and the precision of 100% and 91% respectively. Also, the recall (sensitivity) of RF and SVM are 100% and 95%, respectively.

Keywords: Diabetes disease, SVM, RF.

1. Introduction

A condition known as diabetes occurs when the body can't metabolize glucose, or blood sugar, which causes blood sugar levels to increase dangerously high. One hormone that aids in blood sugar regulation is insulin. Individuals with diabetes can have diabetes of one type, which is caused by inadequate synthesis of insulin or diabetes of two type, which is caused by an inability to respond appropriately to insulin. Two type of diabetes accounts for about 90% of all diagnosed cases [1]. Between 1990 and 2010, the proportion of individuals with diabetes more than tripled, and the number of new cases increased yearly [2]. It is estimated that 80–85% of two type diabetes cases are caused by obesity[3], The World Health Organization study shows that, since 1975, obesity rates have nearly tripled globally[4]. These findings support the theory that there is a direct correlation between obesity and the higher occurrence of two types of diabetes. Instantaneous access to a wide range of purchases, communication, education, and entertainment has all been made available by technology.

People are eating more meals high in fat and calories and exercising less as a result of new technological advancements that offer quick entertainment, education, connection, and shopping [5]. This is one of the causes of the global obesity epidemic.

Thanks to recent technology advancements that allow for rapid entertainment, education, communication, and shopping, people are eating more fat and calorie laden meals and exercising less [5]. This is among the reasons behind the obesity pandemic in the world.

The remainder of the document is structured as follows: previous research on the same issue is included in Section 2. The intricate specifics of the dataset that was used, the steps involved in getting the data ready for a machine learning model, and its methods that were employed are all covered in Section 3. Moreover, the suggested system mentioned in Section 4. Furthermore, Section 5 presents the outcomes of each technique used along with the corresponding accuracy of it. Section 6 offers the conclusion at the end.

2. Related Work

Since identifying heart problems and protecting patient data are important aspects of our work, this section assesses the literature that has been written in related fields:

Several machine learning techniques are applied to the dataset in this paper in order to do classification. With a 96% accuracy rate, logistic regression produces the best results of all. With a 98.8% accuracy rate, the AdaBoost classifier was the most successful model after pipeline application. The authors have seen a comparison of two different datasets' and machine learning systems' accuracy. It is clear from comparing this dataset to earlier ones that the model predicts diabetes more precisely and accurately[7].

The main goal of this paper was to create a model with the help of supervised learning techniques that might help doctors identify diabetes in patients early and improve their quality of life. Of the training methodologies utilized to train many models mentioned in the research, the Random Forest Classifier attained the highest accuracy of 82% using the Pima Indians dataset [6].

The IQ-OTH/NCCD dataset, a unique collection of CT scans for lung cancer, is used in this paper. This information was used to develop an automated approach for lung cancer diagnosis. Image processing consists of feature extraction, segmentation, and enhancement, is where this technology began. The photos were then classified using SVM once features were selected. There were three distinct kernels used to train the SVM, linear, RBF, and polynomial models. The high accuracy of 89% was obtained by combining two recovered features with SVM with a polynomial kernel and the GLCM and Gabor filter. The associated performance criteria show that the used approach is reliable in distinguishing between benign and malignant chest CT images, with scores at or above 90% for sensitivity and specificity [8].

Support Vector Machine classification algorithm and the Python programming language are used by the authors of this paper to classify the underprivileged group. The amount of data gathered was determined by the dependent variable, kind of dwelling, land area, income, and occupation. Processing, grading, labeling, and testing are done to the data yields a 97% accuracy rating [9].

It is evident from this article that the job was divided into two sections for processing. The Pima Indian Diabetes Dataset and the Localized Diabetes Dataset were gathered from Bombay Medical Hall and UCI, respectively, for the first section. The processing of the two datasets using two distinct methods, including the implementation and examination of certain classification techniques with two class issues for both the real and existent datasets, is covered in the second section. Simple classification is used in the first strategy, while feature reduction is used in the second strategy before classification techniques are applied. The four classification techniques applied in both strategies are as follows: features reduction, polynomial, RBF, linear, and sigmoid function kernels. With the exception of the RBF Kernel with PSO feature reduction approach, all kernel SVM algorithms improved in this case. For the Pima Dataset, every classification approach performs better than average, and for the Localized Diabetes Dataset, every classification method that is used yields positive results [10]

3. Materials And Methods

The primary methods and supplies used in this work are summarized in this section.

3.1 Feature Standardization

Before entering the data for the methods used, work was done on feature scaling, which involves changing values using one of two basic techniques: normalization or standardization. During normalization, the input values are changed to fall between 0 and 1. By using standardization techniques, the numbers are transformed to have a zero-centered value and a standard deviation of 1. The normalizing technique was utilized in this paper.

3.2 Support Vector Machine (SVM)

The Support Vector Machine algorithm is one type of supervised learning technique that are practical to solve regression or classification problems. Finding the hyperplane that divides the different classes in the training set as maximally as possible is the main objective of support vector machines, or SVMs. This is done by comparing the separation between each class's nearest data points and the hyperplane to find the hyperplane with the largest margin [11]. Once the hyperplane has been located, newly collected data can be classified through determining which side does it land on in the hyperplane.

SVMs are especially helpful in situations where the data has a large number of features and/or a distinct margin of separation [12]. A comparatively straightforward supervised machine learning approach for regression and/or classification is called Support Vector Machine (SVM). Although regression can occasionally benefit greatly from it as well, it is more suited for classification. SVM essentially locates a hyper plane that establishes a boundary between the various categories of data. On this two-dimensional hyperplane, there is nothing but a line. Each data item in the dataset is plotted using support vector machines in an N-dimensional space, where N is the number of features or attributes in the dataset. Next, the hyperplane that best divides the data is selected. On the other hand, there are many ways to handle issues with multiple classes. Assisting Vector Machine with different class issues in multi-class cases, SVM can be applied by building a binary classifier for every kind of input. Each classifier will yield one of two outcomes: either the data point is included in the class OR it is not as shown in below figure [12].

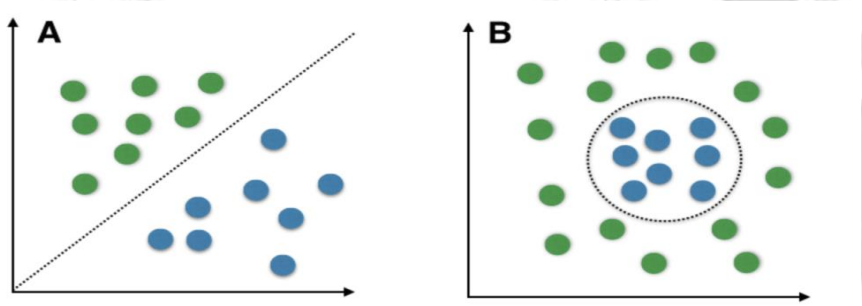

Figure1. A: Data that is linearly separable B: Incomparably dividend data[12]

For data that is linearly non-separable, kernelized SVM is utilized. When a set of data cannot be separated linearly in a single dimension. It is possible to convert this data into two dimensions, at which point it will exhibit linear separability. Each one-dimensional data point is mapped to a two-dimensional ordered pair in order to accomplish this. Because of this, data that can be easily transferred to a higher level even if it is not linearly separable in any dimension. in order to become separable linearly. This is a widespread and incredibly potent metamorphosis. All that a kernel is is a similarity metric between data points. When you have two data points in the original feature space and one in the newly transformed feature space, the kernel function in a kernelized SVM informs you how similar the two points [13]:

An extremely interesting observation is that SVM does not really need to transfer the data that supports the new high dimensional feature space. This is known as the kernel trick. The Kernel Trick: only inasmuch as they relate to similarity computations between point pairs in the higher dimensional feature space where the updated feature representation is implicit, can the kernelized SVM internally carry out these complex adjustments. In essence, the kernel of a kernelized support vector machine (SVM) is this similarity function. It is a kind of complex dot product mathematically. This suggests that SVM can be used in scenarios where the underlying feature space is complex or perhaps infinite-dimensional. Because of its complexity, the kernel technique is outside the purview of this paper. Key elements of kernelized Support Vector Classifiers (SVCs) [14]

1) **The Kernel:** Both the type of transformation and the type of data are taken into consideration while choosing the kernel. The (RBF) is the kernel by default.
2) **Gamma:** This factor determines the transformational swath of a single training sample, hence influencing the degree to which the decision borders encircle input space points. Points further apart are regarded as similar if the gamma value is minimal. As a result, there are more grouped points and smoother decision boundariesthough perhaps less accuracy. Points get closer together at larger gamma values, which could lead to overfitting [12].
3) **The 'C' parameter:** The quantity of regularization performed to the data is controlled by this parameter. Large values of C indicate low regularization, which leads to an extremely good fit (perhaps overfitting) in the training set. Stronger regularization is indicated by lower C values, which increases the model's error tolerance and could result in decreased accuracy.[12]

Advantages of Kernelized SVM:

1) They have excellent results across a variety of datasets.
2) They are adaptable. One can provide several kernel functions or define bespoke kernels for particular data types.
3) For both high and low dimensional data, they perform admirably. [12]

Disadvantages of Kernelized SVM:

1) As the size of the training set grows, efficiency (running time and memory use) declines.
2) It requires precise parameter tuning and input data normalization.
3) It does not offer a probability estimator that is direct.
4) It is difficult to understand the reasoning behind a prediction.[12]

3.3 Random Forest (RF)

It is applied to forecast and classify data. In order to improve a dataset's predictive performance, this classifier builds several decision trees on various dataset subsets. To be more precise, Random Forest does not rely only on one decision tree; instead, it averages the forecasts made by each tree or selects the predictions that receive the majority of votes [15]. By expanding the number of trees in the forest, more accuracy is attained and the overfitting issue is avoided. Furthermore, the algorithm is reliable. New data points typically only have a small impact on one tree, meaning that the algorithm as a whole is not greatly impacted [16]. The next stages complete this algorithm:

1) Select data samples at random from the provided training dataset.
2) Make a decision tree for each sample of training data.
3) Calculate the voting results by maximizing or average the decision trees.
4) Select as the desired outcome the forecasted outcome that garnered the most number of votes.[17]

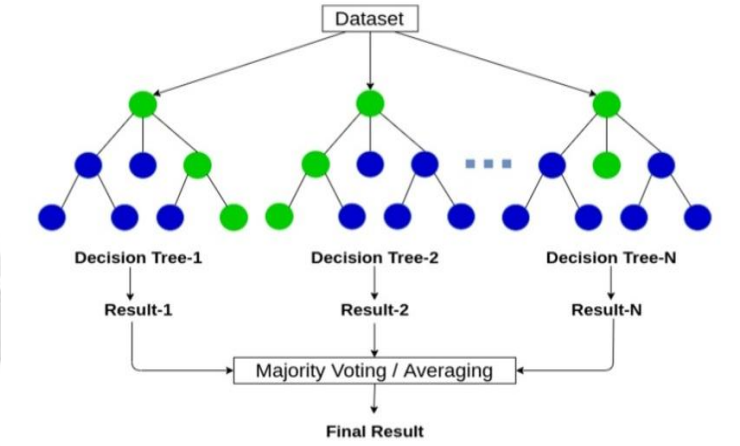

Figure 2. Structure of RF algorithm [17]

Fig.2 shows the stages of the RF algorithm.

3.4 Performance Metrics

Four evaluation criteria are used in this paper: sensitivity, accuracy, precision, and F1-score. The number of times a classifier accurately identified data throughout the whole dataset is the accuracy measure. The percentage of cases that were accurately classified to all cases that were correctly classified is how it is displayed. The ratio of correctly labeled positive instances to all cases that were either incorrectly or correctly identified as positive is known as the precision. Put differently, accuracy quantifies the number of occurrences that can be positively identified as affirmative. By dividing the entire number of positive instances including the cases that were incorrectly labeled as negative by the total number of positive occurrences, sensitivity calculates how well the system can classify positive situations. Combining precision and sensitivity data yields the F1-score. which evaluates the accuracy of a classifier. The four metrics mentioned above are generated using the following equations:[19][20]

Accuracy: It's the ratio of the total number of input samples to the number of accurate predictions, It is given as:

$$Accuracy = \frac{(TP1 + TN1)}{(TP1 + TN1 + FP1 + FN1)} \quad (2)$$

Precision: It is calculated by dividing the number of correctly predicted positive results by the number of positive results the classifier anticipated, It is expressed as:

$$\Pr ecision = \frac{TP1}{(TP1 + FP1)} \quad (3)$$

Recall: It is the number of correct positive results divided by the number of all relevant samples, In mathematical form it is given as:

$$Sensitivity = \frac{TP1}{(TP1 + FN1)} \quad (4)$$

Where the acronyms mentioned above can be clarified as follows:

- **True Positive (TP1)**is the positive states that are appropriately classified as such.
- **False Positive (FP1)**indicates the bad conditions that are mistakenly classified as good conditions.
- **True Negative (TN1)**shows the appropriate categorization of a negative diagnosis.
- **False Negative (FN1)**shows the positive situations that have been mistakenly categorized as negative.

F1_Score: It is employed to gauge the correctness of a test. The F1 Score represents the Harmonic Mean of memory and precision. F1 Score has a range of [0, 1]. It provides you with information about the robustness and precision of your classifier. In mathematics, it is given as:

$$F1\_score = \frac{2 * \Pr ecision * Sensitivity}{\Pr ecision * Sencitivity} \quad (5)$$

Standardization: results in a distribution that has a mean of 0 and a variance of 1

$$V_{new} = \frac{V_i - V_\mu}{V_\sigma} \quad (6)$$

Where V_i is a dataset's feature value, $V_{new}$is a scaled value for a feature, $V_\mu$is the feature values' mean, and $V_\sigma$is the feature value standard deviation [19].

4. Proposed System

This paper describes the design of a system intended to help diagnose and classify people with diabetes through the use of a variety of clinical symptoms. The performance of the random forest and support vector machine algorithms was also compared. The Random Forest algorithm was chosen due to its ability to examine frequencies of patient-related health data. Using this method, researchers can better understand the associations between various diabetes-related factors and develop treatment and care plans. The efficacy of this method in handling high

dimensional data with numerous, potentially complex elements is the cause of this. The SVM method is a common choice in health and medical data analysis to comprehend and forecast factors connected with diabetes since it separates data and creates precise prediction models based on the various components in the database.

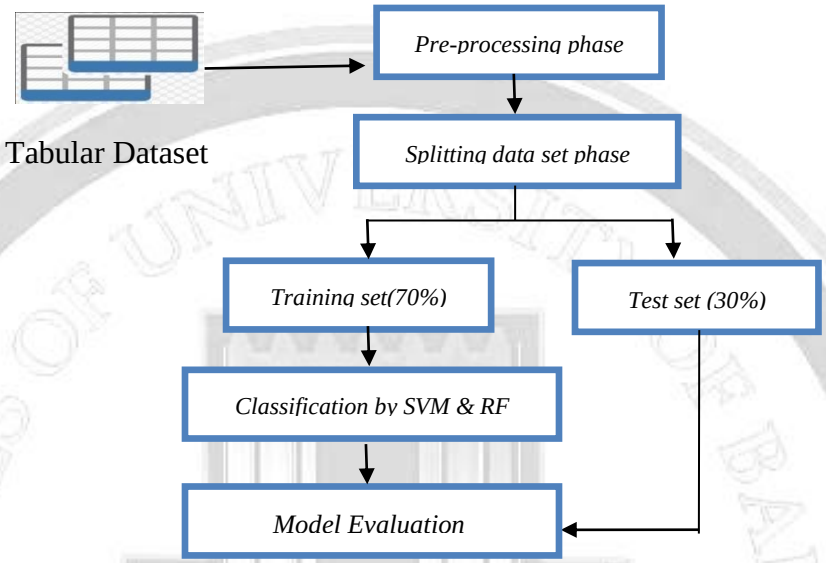

Figure 3.The architecture of the proposed system

4.1 Pre-Processing Phase

Preliminary data processing must be completed before to the classification procedure in order to achieve high accuracy in the classification of diabetes data. In our suggested system, we employ feature scaling as the pre-processing strategy, while there are other approaches as well. It was applied to cut down on complexity and time. Although there are other techniques for feature scaling, the standardization approach was employed in this paper. It is a method for equally dispersing the independent aspects of the data within a predetermined range. It controls wildly fluctuating amounts or values. The performance of the algorithm will be impacted if one of the feature scaling methods is not used and the data set has dissimilar values. This is because the algorithm will prioritize large values during training and ignore small values, which will have a significant impact on the accuracy of the results. Equation (6) is used to rescale a feature value as part of the standardization process, producing a distribution with a mean of 0 and a variance of 1.

4.2 Splitting Data Phase

In this phase the data will be divided two groups the training set, whose main goal is to make precise discoveries using machine learning algorithms, and the testing set, which is

employed to assess the effectiveness of the system. Essentially, a training set equal 70% of the data, while a testing set equal 30%.

4.3 Classification Data Phase

This phase is considered one of the most important phases in the proposed system, as two machine learning algorithms are used, namely RF and SVM, for the purpose of classifying the data set for women as to whether they have diabetes or not.

4.4 Model Evaluation Phase

Finally, the model evaluation stage, where it is necessary, after designing any model, to evaluate its work to determine its level of performance, using a number of well-known standards such as F1_score, sensitivity, accuracy, and precision.

5. Experimental Results And Discussion

The results of these phases are the main subject of analysis and assessment in this section. The dataset1 used in these stages includes a variety of features, including age, pregnancy, blood pressure, glucose, diabetes, and other factors that can significantly impact an individual's risk of developing diabetes disease, as shown in table. These are crucial elements that may significantly affect how diabetes in women is diagnosed and categorized. By examining these characteristics, precise models can be developed for diabetes classification and prediction, as well as for comprehending the connections between these variables and the illness. It should be noted that the system is implemented in Python.

Table 1. Features of the PIMA Indian dataset described

Attributes	Range	Description
Pregnancies	0-17	number of pregnancies
Glucose	0-199	In an oral glucose tolerance test, plasma glucose levels after two hours
Blood Pressure	0-122	Blood pressure diastolic (mm Hg)
BMI	0-67.1	Body mass index (kg of weight divided by m of height)^2)
Skin Thickness	0-99	Skin fold thickness (mm) of the triceps
Diabetes Pedigree Function	0.078-2.42	An algorithm that rates the risk of diabetes according to family history
Age	21-81	Years of age
Insulin	0-846	2-Hour serum insulin (mu U/ml)
Outcome	0-1	Class variable, diagnosis classes: 1 = diabetes diagnosis, 0 = good health

Table 2 shows the results of implementing the RF algorithm in classifying diabetes data. From the table, It is evident that when there were two trees, the RF algorithm performed at its peak, reaching 100%.

Table 2.The results of RF technique based on number of trees

Applied method	No. of trees	Accuracy	Precision	Sensitivity	F1_score
RF	1	0.97	0.98	0.95	0.96
	2	1	1	1	1

Table 2 shows the results of implementing the SVM algorithm on a diabetes data set, where the accuracy is 89%, which is less than the accuracy resulting from implementing the RF algorithm on the same data set.

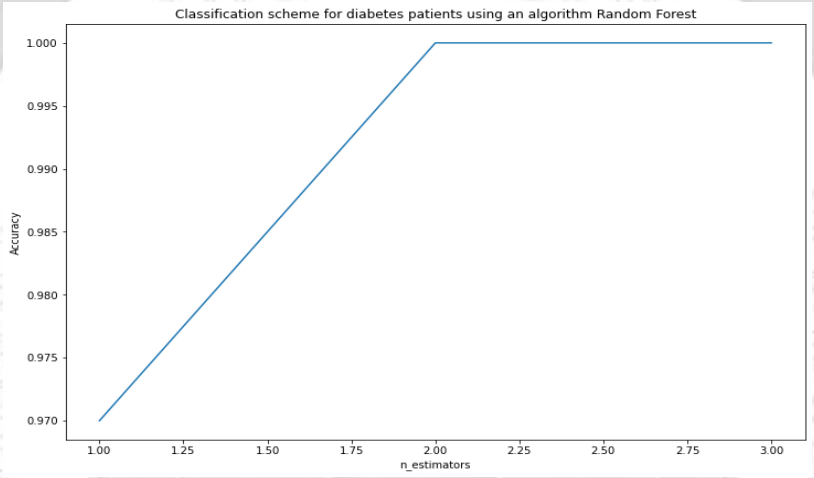

Figure 4. The Accuracy of RF where number of trees is 1,2

Fig. 4 shows The accuracy of RF depends on the number of trees that were used in this paper, as two possibilities for the number of trees were used, namely 1 and 2.

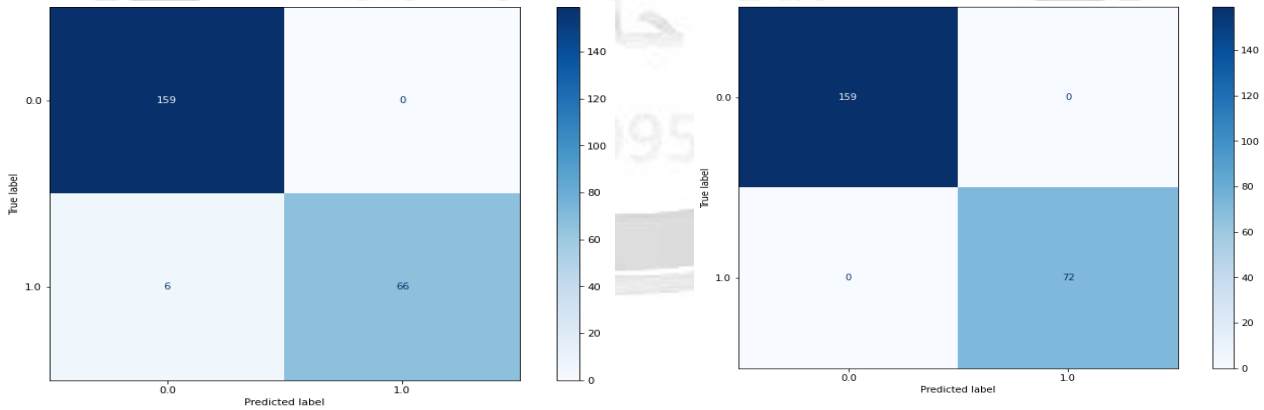

Figure 5. RF Confusion Matrix where no. of trees is 1

Figure 6. RF Confusion Matrix where no. of trees is 2

1 The dataset is available online at (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)

It is evident from the statistics above that the algorithm performed at its peak when there were two trees.

Table 3. The results of SVM technique

Applied method	Accuracy	Precision	Sensitivity	F1_score
SVM	0.89	0.91	0.95	0.92

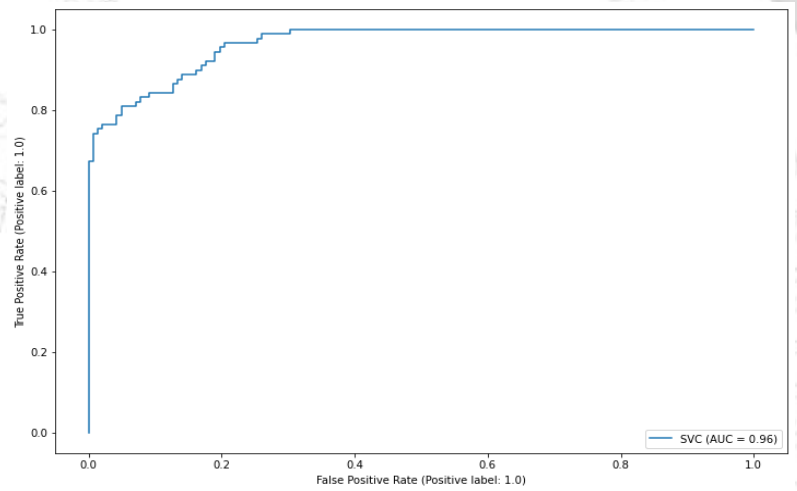

Figure 7. The accuracy of SVM algorithm

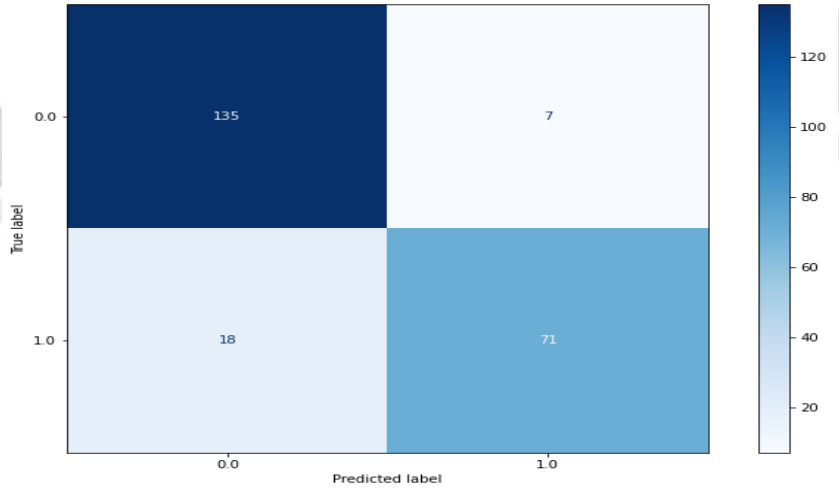

Figure 8. SVM Confusion Matrix

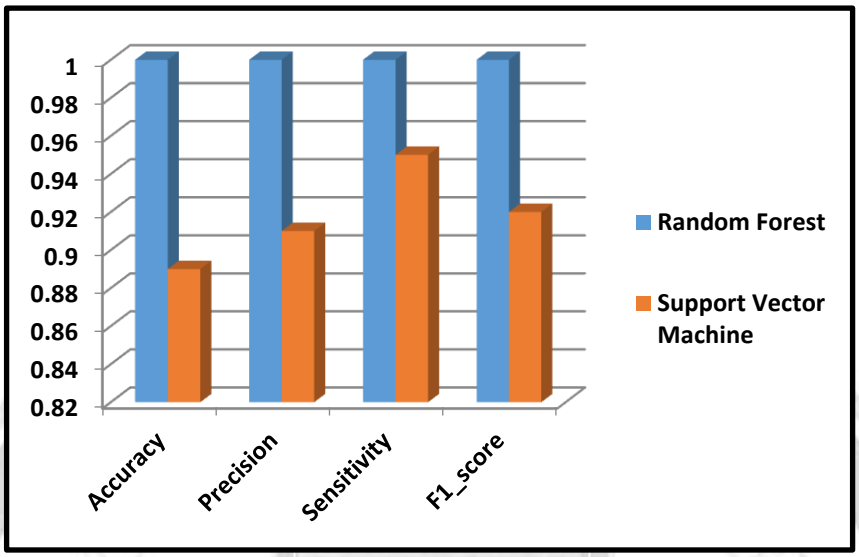

Figure 9. The performance comparison between RF and SVM

Fig.9 shows a comparison between the RF algorithm and SVM method based on the measurements used in this project and the same dataset is used. The figure shows the superiority of the RF algorithm in performance compared to SVM.

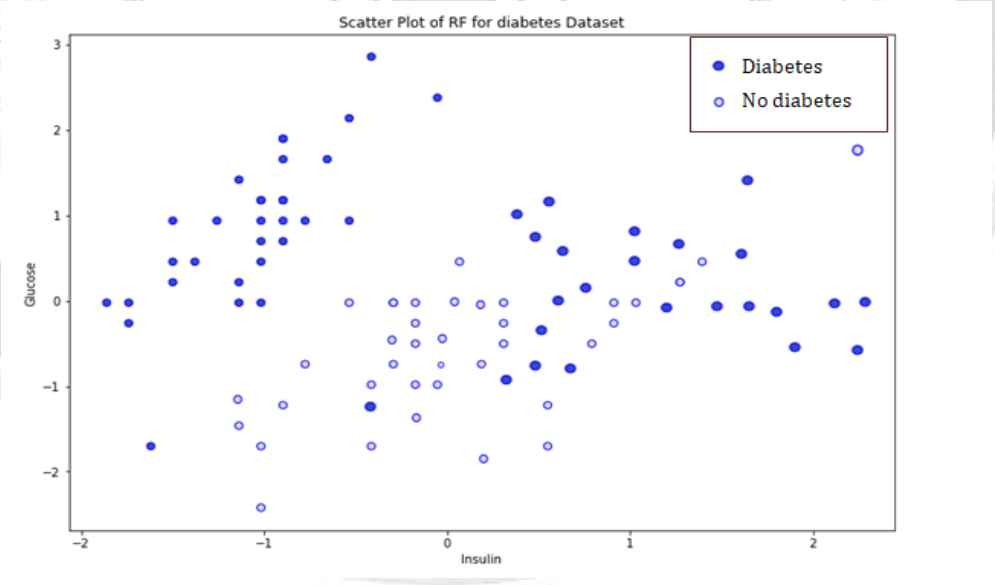

Figure 10. Scatter plot of the attributes of glucose and insulin after implementing the RF algorithm

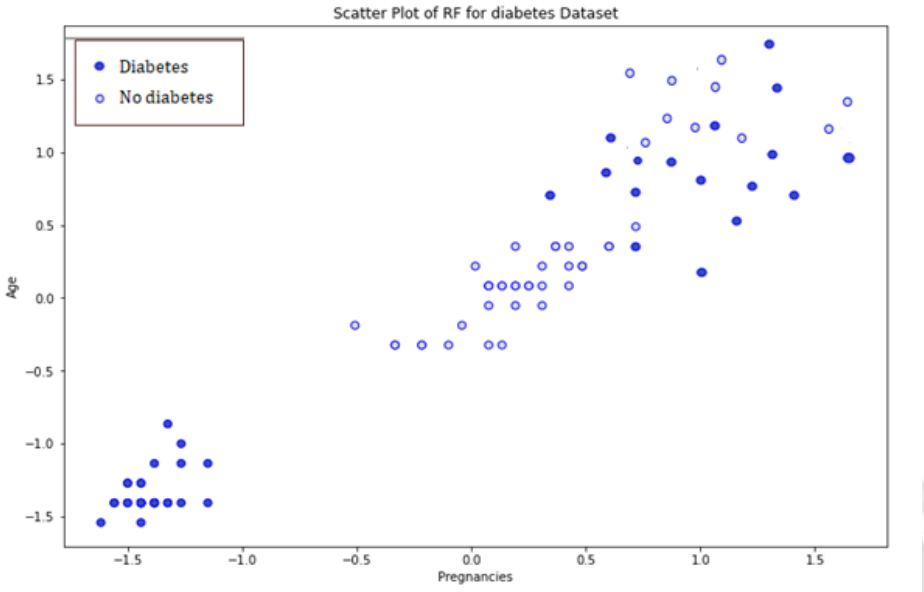

Figure 11. Scatter plot of the attributes of pregnancies and age after implementing the RF algorithm

Figures 10 and 11 show the scatter plot of diabetes cases (positive) and non-diabetic cases (negative) for the attributes of insulin and glucose also pregnancies and age after implementing the RF algorithm.

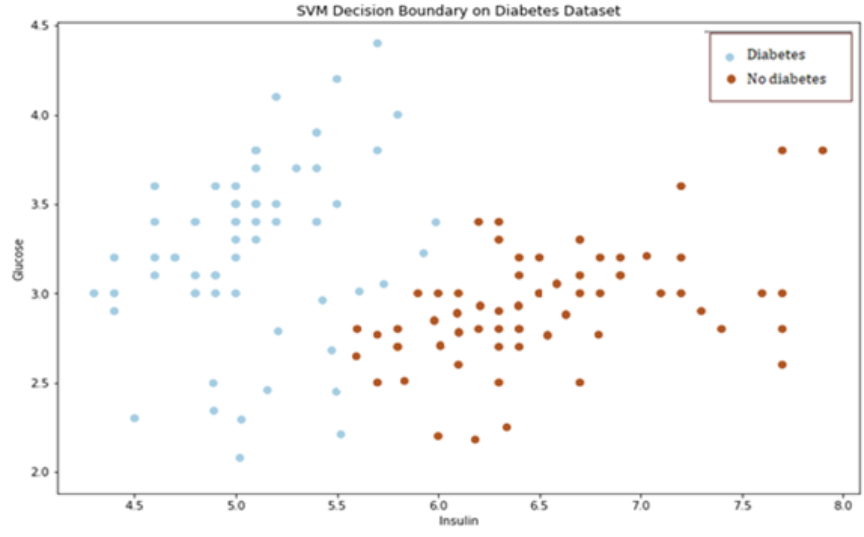

Figure 13. Scatter plot of the attributes of glucose and insulin after implementing the SVM algorithm

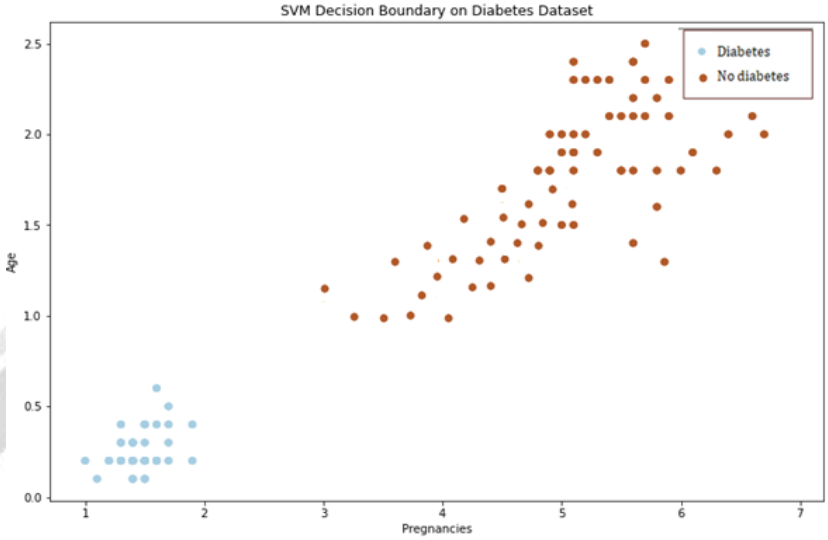

Figure 14. Scatter plot of the attributes of age and pregnancies after implementing the SVM algorithm

Figures 13 and 14 show the scatter plot of diabetes cases (positive) and non-diabetic cases (negative) for the attributes of insulin and glucose also pregnancies and age after implementing the SVM algorithm.

Table 4. Comparison table with some related work

References	Year	Dataset	Technique(s)	Accuracy
[6]	2022	www.kaggle.com	SVM and RF	SVM = 79% RF = 82%
[7]	2019	www.kaggle.com	RF	91%
Our envisioned work	2024	www.kaggle.com	SVM and RF	RF = 100% SVM = 89%

The findings of the current research are contrasted with those of various relevant publications in Table 4. This table includes five columns: the citation's number and the year it was published, the data set this reference utilized, the technique this reference used, and the accuracy this reference attained.

6. Conclusion

The main objective of the paper is to develop a model that uses supervised learning techniques that will assist medical professionals in early diagnosis of diabetes in order to improve the quality of life for their patients. After implementing the RF and SVM algorithms on the data set, RF classifier obtained the highest accuracy of 100% among the different training strategies presented in the research.

7. Future Work

It is possible to use several machine learning algorithms in future research and compare them to determine which one is the most accurate. This means that different datasets for men can be used to paper different diseases, such as cancer, heart disease, etc

References

[1] "Diabetes Fast Facts," [Online]. Available: https://www.cdc.gov/diabetes/basics/quick-facts.html.

[2] "Statistics and facts about type 2 diabetes," Medical news today, [Online]. Available: https://www.medicalnewstoday.com/articles/318472.

[3] "Diabetes and Obesity," The global diabetes community, [Online]. Available: https://www.diabetes.co.uk/diabetes-and-obesity.html.

[4] "Obesity and overweight," World Health Organization, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[5] "Obesity Rising: Can We Do Anything to Reverse This Deadly Trend?," health line, [Online]. Available: https://www.healthline.com/health-news/obesity-rising-can-we-reverse-this-deadly-trend#Complex-factors-behind-obesity.

[6] N. Abdulhadi and A. Al-Mousa, "*Diabetes Detection Using Machine Learning Classification Methods*," 2021 International Conference on Information Technology, ICIT 2021 - Proceedings, pp. 350–354, 2022.

[7] V. V Aishwarya Mujumdar, "*Diabetes Prediction using Machine Learning Algorithms*." Elsevier B.V., India, pp.292-299, 2019.

[8] H. F. Kareem, M. S. AL-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, "*Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset*," Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 3, pp. 1731–1738, 2021

[9] A. A. Aldino, A. Saputra, A. Nurkholis, and S. Setiawansyah, "*Application of Support Vector Machine (SVM) Algorithm in Classification of Low-Cape Communities in Lampung Timur*," Building of Informatics, Technology and Science (BITS), vol. 3, no. 3, pp. 325–330, 2021.

[10] D. K. Choubey, S. Tripathi, P. Kumar, V. Shukla, and V. K. Dhandhania, "*Classification of Diabetes by Kernel Based SVM with PSO*," Recent Advances in Computer Science and Communications, vol. 14, no. 4, pp. 1242–1255, 2019.

[11] A. T. O. Siti-Farhana Lokman1(&) and and S. M. Muhamad Husaini Abu Bakar, "*Blockchain-Based Image Sharing Application*," Communications in Computer and Information Science, vol. 1132 CCIS, pp. 46–59, 2020.

[12] "Introduction to Support Vector Machines (SVM)", geeksforgeeks/ [Online].Available:https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/

[13] R. I. Borman, F. Rossi, Y. Jusman, A. A. A. Rahni, S. D. Putra, and A. Herdiansah, "*Identification of Herbal Leaf Types Based on Their Image Using First Order Feature Extraction and Multiclass SVM Algorithm*," 2021 1st Int. Conf. Electron. Electr. Eng. Intell. Syst. ICE3IS 2021, p.5, 2021.

[14] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "*Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review*," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. IEEE, Canada, p.18, 2020.

[15] S. A. Salman, S. A. Dheyab, and Q. M. Salih, "Parallel Machine Learning Algorithms," *Mesopotamian J. Big Data*, p.4, 2023.

[16] K. J. G. S. Dmytro Chumachenko , Mykola Butkevych , Daniel Lode , Marcus Frohme and and Alina Nechyporenko, "*Machine Learning Methods in Predicting Patients with Suspected Myocardial Infraction Based on Short-Time HRV Data*." Sensors, Germany, p.18, 2022

[17] F. A.-S. Randa shaker Abd-Alhussain , Hadab Khalid Obayes, "*Secure Heart Disease Classification System Based on Three Pass Protocol and Machine Learning,*." Iraqi Journal for Computer Science and Mathematics, Iraq, p. 11, 2023.

[18] F. A.-S. Randa shaker Abd-Alhussain1 , Hadab Khalid Obayes2, "*Utilizing Synthetic Tabular Data Method to Improve Heart Attack Prediction Accuracy*," Al-Salam Journal for Engineering and Technology, Iraq, p.22, 2023.

[19] H. K. Obayes, F. S. Al-Turaihi, and K. H. Alhussayni, "*Sentiment classification of user's reviews on drugs based on global vectors for word representation and bidirectional long short-term memory recurrent neural network*," Indonesian Journal of Electrical Engineering and Computer Science. Iraq, p. 9, 2021.

ARTICLE
JOURNAL`S UNIVERSITY OF BABYLON FOR ENGINEERING SCIENCES (JUBES)
مـــجلـة جـــامعة بـــابـل للعلـــوم الهندسية
Vol. 32, No. 3. \ 2024 ISSN: 2616 - 9916

تصنيف مرض السكري لدى النساء استنادًا إلى آلة ناقل الدعم وخوارزميات الغابات العشوائية

رندة شاكر عبد الحسين[1] زينة عبد الحسين صالح[2]
[1] قسم البعثات والعلاقات الثقافية، رئاسة جامعة بابل
Email: randashaker1984@gmail.com
[2]قسم الدراسات والتخطيط، رئاسة جامعة بابل
Email: zina.badi@uobabylon.edu.iq

الخلاصة:

مرض السكري هو مرض خطير. يشار إليه بمستويات السكر في الدم و/أو مستويات الجلوكوز. مرض السكري هو مرض مزمن يمكن أن يؤدي إلى أزمة صحية عالمية، ولكن هناك أشياء يمكن القيام بها للمساعدة في السيطرة على هذه الأزمات. مصدر الطاقة الأساسي الذي يحصل عليه الأشخاص المصابون بالسكري من الطعام بشكل يومي هو سكر الدم أو الجلوكوز. هرمون الأنسولين الذي يفرزه البنكرياس، يساعد في امتصاص الجلوكوز من الدم إلى الخلايا بحيث يمكن استخدامه كمصدر للطاقة في المهام اليومية. يبقى الجلوكوز في الدم عندما ينتج الجسم كميات غير كافية من الأنسولين، مما قد يسبب عددًا من المشكلات الصحية مثل النوبات القلبية والسكتات الدماغية. هناك أشكال عديدة لمرض السكري، وأكثرها انتشارًا هو النوع الأول والنوع الثاني. يتم تشخيص النوع الأول عادة عند الأطفال والشباب، في حين يتم تشخيص النوع الثاني عادة عند منتصف العمر أو كبار السن. الهدف من المشروع هو تطوير نظام يمكنه تصنيف المرضى بدقة على أنهم مصابون بالسكري أو غير مصابين بالسكري من خلال الجمع بين نتائج تقنيات التعلم الآلي المختلفة مع الخوارزميات مثل آلة ناقل الدعم وخوارزميات الغابات العشوائية. التعلم الآلي هو مجال علمي حيث يستمد التعلم الآلي من الخبرة البشرية. النموذج الذي يتنبأ بمرض السكري بشكل أفضل هو النموذج الذي يتم تحديد دقته كنسبة مئوية من خلال حساب دقة النموذج باستخدام كل طريقة. وفقًا للنتائج التجريبية، تتمتع دقة RF وSVM بدقة 100% و89% على التوالي ودقة 100% و91% على التوالي. أيضًا، تبلغ نسبة استدعاء (حساسية) الترددات اللاسلكية و100% SVM و95% على التوالي.

الكلمات الدالة: مرض السكري، SVM، RF.