Computer-aided system for extending the performance of diabetes analysis and prediction

Conference Paper · August 2021

DOI: 10.1109/ICSECS52883.2021.00091

5 authors, including:

Saydul Akbar Murad
Universiti Malaysia Pahang
18 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Zafril Rizal M. Azmi
Universiti Malaysia Pahang
21 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Nusrat Jahan Prottasha
Daffodil International University
20 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Md Kowsher
Stevens Institute of Technology
64 PUBLICATIONS **184** CITATIONS
SEE PROFILE

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) | 978-1-6654-1407-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSECS52883.2021.00091

Computer-aided system for extending the performance of diabetes analysis and prediction

Saydul Akbar Murad
Faculty of Computing
Universiti Malaysia Pahang
Malaysia
saydulakbarmurad@gmail.com

Zafril Rizal M Azmi
Faculty of Computing
Universiti Malaysia Pahang
Malaysia
zafril@ump.edu.my

Zaid Hafiz Hakami
Faculty of Computer Science & Information Technology
Jazan University, Jazan, Saudi Arabia
zhhakami@jazanu.edu.sa

Nusrat Jahan Prottasha
Department of Computer Science and Engineering
Daffodils International University
Bangladesh
jahannusratprotta@gmail.com

Md Kowsher
Department of Applied mathematics
Noakhali Science and Technology University
Bangladesh
ga.kowsher@gmail.com

***Abstract*- Every year, diabetes causes health difficulties for hundreds of millions of individuals throughout the world. Patients' medical records may be utilized to quantify symptoms, physical characteristics, and clinical laboratory test data, which may then be utilized to undertake biostatistics analysis to uncover patterns or characteristics that are now undetected. In this work, we have used six machine learning algorithms to give the prediction of diabetes patients and the reason for diabetes are illustrated in percentage using pie charts. The machine learning algorithms used to predict the risks of Type 2 diabetes. User can self-assess their diabetes risk once the model has been trained. Based on the experimental results in AdaBoost Classifier's, the accuracy achieved is almost 98 percent.**

Keyword's- Diabetes, AdaBoost Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier, Bernoulli NB, MLP Classifier and Impact Learning, Cloud Computing.

I. INTRODUCTION

Diabetes is a metabolic condition that occurs when a person's blood sugar levels are too high. It is caused by either the body's inability to create enough insulin or the cells' inability to react to the insulin that is generated [11]. Diabetic diagnosis necessitates a precise evaluation of diabetes facts. As a result, diabetes data analysis for diabetes control has become a fascinating study topic [16]. Diabetes affects over 376 billion people worldwide, with women accounting for the bulk of those affected [12]. According to the World Health Organization's medical study, this number is predicted to reach near or beyond 490 billion by 2030 [1]. When blood sugar levels are higher than usual (4.4 to 6.1mmol/L), a person is said to have diabetes. Insulin is a hormone generated by the pancreas, which is in charge of supplying glucose to the body's cells. A diabetic patient's body either produces insufficient insulin or is unable to utilize it effectively.

This research has focused on Type-2 diabetes. We have tried to find out by which reason people are most affected or less affected and give the prediction of diabetes by evaluating some features such as age, genetic history, etc. [19]. For this research, we have used a large dataset where data is collected from diabetes patients and non-patient people. To analyze this dataset, we have used several machine learning classifiers including AdaBoost Classifier, Random Forest Tree Classifier, K-Neighbors Classifier, Bernoulli NB, MLP Classifier and Impact Learning. This research will help people to be conscious about diabetes.

The following section discussed about the related work. In section III we discussed the methodology involved and section IV is about result and discussion. Finally, Section V and VI discussed future work and conclusion respectively.

II. RELATED WORK

Leyer et al. research attitudes of each person who is affected by type-1 diabetes to solve hypoglycemia [1]. Žilinskienė, Jolanta, et al. tries to find a connection between mothers' parenting style (PS) and their children with type I diabetes (T1DM) disease management [2]. Jain, Bhavini, et al. presented a prediction of diabetes using symptoms of a person and showed the changing characteristics [3]. Abaker, Ali A. et al. established a model to decide whether a diabetes patient should be admitted into hospital or home [4]. Sahoo, Priyabrata. Generated a model for predicting diabetes based on important attributes [5]. Saxena et al. tries to improve our knowledge of diabetes mellitus onset prediction [6]. Li, Jun,

978-1-6654-1407-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ICSECS52883.2021.00091

et al. develop a noninvasive diabetic's risk prediction model based on tongue characteristics fusion and estimate the risk of pre diabetics and diabetics using machine learning techniques [7]. Abokhzam et al. proposed a technique based on the ML Grid Search algorithm is suggested for efficiently and successfully diagnosing Diabetes Mellitus [8]. Gupta, Deepak, et al. investigates and evaluates several machine learning (ML) algorithms that can aid in predicting diabetes risk at an early stage and improving diabetes medical diagnosis [9]. Wadhwa et al. uses a huge amount of multimodal patient data used in this study to perform correlations between BMI, blood pressure, glucose levels, diabetic pedigree function, and skin thickness in patients with diabetes of various ages [10]. In addition, Kowsher et al showed the procedures of best medication detection of diabetes with a good accuracy.

Unlike the above-mentioned works, this research analyzed the rate of causes and also performed the forecasting model to generate the diabetes condition of patient.

II. METHODOLOGY

This research uses five major steps to get the prediction result: corpus building, label verification, data prepossessing, training data using ML algorithms, and predictions. The data quality is maintained by the large number of samples and adequate diverse variation.

A. Data Sources

We gathered data for the corpus by using social media platforms where individuals offer their valued ideas. We generated two Google forms to collect data by asking various questions. There is a form for patients and another for non-patients, but all fields are the same. Total number of collected data are 4069 where non diabetes patients are 2232 and diabetes patients are 1837. There are 11 columns which are age, weight, height, family history, late night sleep habit, late wake up habit, exercise, sleep after eating, addiction, sex and heart diseases.

B. Labeling and verification

Table I shows the collected data from patient and non-patient. Value 1 is for non-patients and 0 is for patients.

TABLE I: LABELING DATA OF COLLECTED DATASET

Age	Weight	height	DH	LNSH	LWH	Exercise	SAE	Addicted	Leveling (Non-p 1, P 0)
23	60	5.2	0	1	1	1	0	0	1
24	68	5.6	0	1	1	0	1	0	1
23	38	5.1	1	1	1	0	1	0	1
20	84	5.7	1	1	1	0	0	0	1
25	64	5.4	0	0	0	0	1	0	1

DH = Diabetes History, LNSH = Late night sleep habit, LWH = Late wake up habit, SAE = Sleep after eating

C. Data Pre-Processing

For data preprocessing we have used mean and median. A computation is used to find the mean, which is the same as the average value of a data collection. Divide the total number of values in the data set by the total number of numbers in the data set.

D. Mean Equation

The total of all the data in a data set is divided by the count n to get the mean $\bar{x}$ as shows in (1).

$$\text{Mean} = \bar{x} = \frac{\sum_{i=0}^{n} X_n}{n} \quad (1)$$

When numbers are given in ascending or descending order, the median is the number in the center. ˇ

E. Median Equation

If the total number of observations (n) is an odd number, then the formula is given in (2):

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} observations \quad (2)$$

If the total number of the observations (n) is an even number, then the formula is given in (3):

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} observation + \left(\frac{n}{2}+1\right)^{th} observation}{2} \quad (3)$$

F: ML Algorithm

In this work we used six algorithm and these algorithms are AdaBoost Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier, Bernoulli NB, MLP Classifier and Impact Learning.

AdaBoost Classifier: Ada-boost, also known as adaptive boosting, is a kind of ensemble boosting [18]. It combines many classifiers to improve classifier accuracy. Adaboost's core principle is to establish the weights of classifiers and train the data sample in each iteration, so that reliable predictions of uncommon observations may be made.

The following step for AdaBoost Classifier:

1. Adaboost randomly selects a training subset.
2. It trains the AdaBoost machine learning model repeatedly by picking the training set based on the last training's correct prediction.
3. It gives incorrectly categorized observations a larger weight so that they have a better chance of being categorized in the next iteration.

4. It distributes weight to the trained classifier in each iteration based on the classifier's accuracy. The more precise classifier will be given more weight.
5. This approach is repeated until all of the training data fits perfectly or until the maximum number of estimators is achieved.
6. Perform a "vote" across all of the learning algorithms you created to categories' them.

Random Forest Classifier: RFC [14] is made up of a large number of Decision Trees. Each node of the decision tree asks a question about the data, with the branches representing alternative replies. A hundred decision trees are combined in a random forest approach. The popularity of RFC models stems from their great accuracy and inexpensive computational costs.

Random forest is a supervised learning method that may be used in classification and regression models. However, it is mostly employed to solve classification issues.

The following step for random forest algorithm:

1. Select a random sample from the dataset.
2. Create a decision tree for each sample. Then take the predicted result from each tree.
3. Voting for each predicted result.
4. Final result will be predicted from the most voted prediction.

BernoulliNB: Another helpful nave Bayes model is Bernoulli Nave Bayes. This method is used for discrete data and is based on the Bernoulli distribution. The primary characteristic of Bernoulli Naive Bayes is that it only takes binary values such as true or false, yes or no, success or failure, 0 or 1, and so on. Text classification with the 'bag of words' model is an application of Bernoulli Nave Bayes classification. sklearn.naive_bayes is a Scikit-learn module. The Gaussian Nave Bayes method will be implemented by BernoulliNB.

Because we're dealing with binary numbers, consider 'p' as the chance of success and 'q' as the likelihood of failure, with q=1-p.

In the Bernoulli distribution, with a random variable 'X,'

$$P(x) = P[X = x] = \begin{cases} q = 1 - p & x = 0 \\ p & x = 1 \end{cases} \quad (4)$$

Here 'x' can only have two values: 0 or 1.

Rule for Bernoulli Naive Bayes Classifier is,

$$P(x_i \mid y) = P(i \mid y)\, x_i + (1 - P(i \mid y))(1 - x_i)$$

K-Nearest Neighbors Classifier: The K-Nearest Neighbor method [14], [15] is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data. It's also known as a lazy learner algorithm since it doesn't learn from the training set right away; instead, it saves the dataset and performs an action on it when it comes time to classify it. The K-NN algorithm can be used for both regression and classification, but it is most commonly used for classification problems. To calculate the Euclidean distance the equation is given in (5):

$$d_{Eucliden} = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (5)$$

We can describe KNN algorithm by following steps:

1. Choose K number of neighbors.
2. Measure Euclidean distance for K number of neighbors.
3. Select the K nearest neighbors from the calculation of Euclidean distance.
4. Count the number of data points in each category among these k neighbors.
5. Assign new data points to the category with the greatest number of neighbors.

MLP Classifier: A feed forward artificial neural network called a multilayer perceptron (MLP) is a type of feed forward artificial neural network (ANN). MLP classifier, which, as the name suggests, is linked to a Neural Network.Back propagation is a supervised learning technique used by MLP during training. MLP is distinguished from a linear perceptron by its numerous layers and non-linear activation. It can tell the difference between data that isn't linearly separable. The Perceptron is made up of two completely linked layers: an input layer and an output layer. MLPs have the same input and output layers, but as shown Fig. 1, they may have several hidden layers in between.

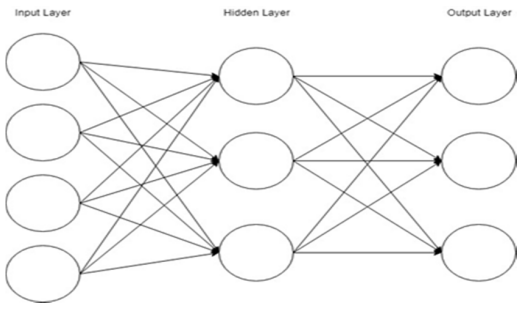

Fig. 1 Diagram of Multilayer

We can describe MLP algorithm by following steps:

1. The relevant libraries and modules are being loaded.
2. Reading the data and carrying out basic data checks.
3. Arrays for the features and the response variable are created.
4. The training and test datasets are being created.
5. The neural network model was built, predicted, and evaluated.

Impact Learning: Impact learning [17] is a supervised machine learning approach that uses examples to solve classification and linear or polynomial regression problems. It also helps with competitive data analysis in systems. This algorithm is special in that it can learn from a competition, which is the influence of separate characteristics.

The equation of Impact learning is given below:

$$\text{Imp} = (y' - (\frac{k\sum_{i=1}^{n} w_i x_i}{r - w_y k} + b))^{\frac{2}{N}} \quad (6)$$

IV. Result And Discussion

As shown in fig. 2, the accuracy, recall, specificity, precision, and F1 Scores of the six methods, AdaBoost Classifier, Random Forest Classifier, BernoulliNB, MLP Classifier, impact learning, and KNN, were compared. The number of accurate and erroneous classifications in each potential value of the variables being classified is used to determine the effectiveness of the classification model. From the results obtained. The Accuracy, Sensitivity, and Specificity, Precision, and F1 score are calculated using the following equations:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

The proportion of total relevant results accurately categorized by the algorithm is referred to recall or sensitivity.

$$\text{Recall} = \frac{TP}{TP+TN} \quad (8)$$

The precision of every record we projected was positive.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

The harmonic mean of precision and recall is the F1-score

$$\text{F1 Score} = \frac{2TP+TN}{2TP+FP+FN} \quad (10)$$

TABLE II: SHOWING SIX ALGORITHMS AND THEIR METRICS SCORE

Algorithm	Accuracy	Recall	Precision	F1
AdaBoost Classifier	0.98	0.99	0.96	0.97
Random Forest	0.97	0.95	0.95	0.95
KNN Classifier	0.83	0.52	0.55	0.51
BernoulliNB	0.85	0.5	0.42	0.46
MLP Classifier	0.86	0.61	0.74	0.63
Impact leraning	0.85	0.56	0.68	0.58

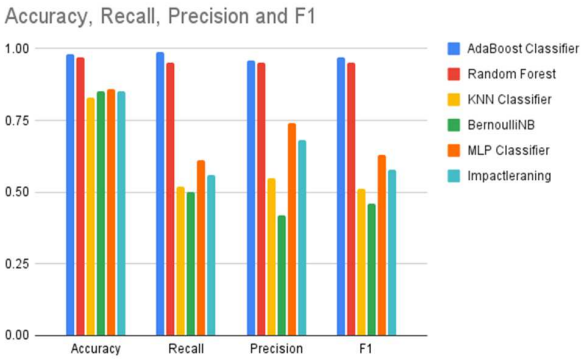

Fig. 2 Model Performance Metrics

The AdaBoost Classifier, as shown in Table II and Fig.2 has the maximum accuracy of 98 percent and on the other hand, another research paper by Mujumdar, Aishwarya, and V. Vaidehi got the accuracy 93 percent using same algorithm [21]. For Random Forest algorithm in this paper the accuracy is 97 percent and 91 percent in Mujumder et al. paper. But for KNN algorithm the accuracy is lowest. Its 83 percent is this paper and 90 percent in Mujumder et al. paper. The AdaBoost Classifier model was created to be more sensitive in predicting true positives, as measured by recall and F1 of 99 and 97 percent, respectively, according to the medical aim. The precision is 98 percent. The accuracy is only the lowest for KNN. For every algorithm we got a good accuracy.

Working Framework: Fig.3 depicts the suggested process for implementing machine learning models in the prediction node. To adjust for missing data, imputation is conducted first, followed by feature scaling to normalize the dataset's value range [20]. During training, feature selection approaches are used to eliminate duplicate features that do not contribute significantly to the prediction outcome and improve overall model fidelity. Finally, during the k-fold cross validation stage, the binary classifiers are fitted to the data, with all samples being utilized for training, validation, and testing, resulting in a more robust classifier.

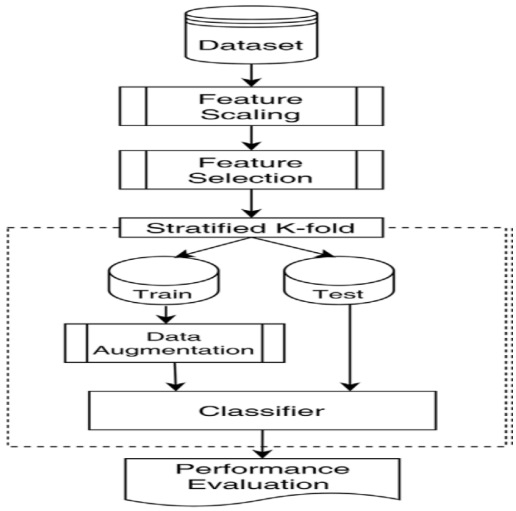

Fig. 3 Framework of the machine learning pipeline

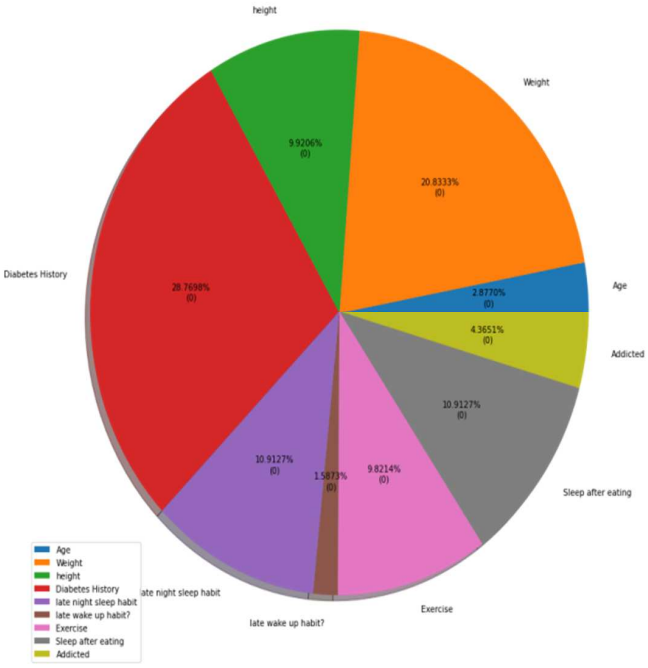

Fig. 4 The cause of diabetes is represented as a percentage.

Fig. 4 shows that, the majority of patients with diabetes are impacted by hereditary factors. The percentage is around 28 percent. Weight is another major factor for diabetes. The percentage is more than 20 percent. The percentage is almost similar for habit and exercise. It's around 10 percent or slightly higher than 10 percent.

V. Future Work

There are a number of study avenues that might be pursued in the future. Firstly, it is still unclear why type-1 diabetes occurs. Second, more variables that were not included in this study might be used to predict diabetic complications. Third, additional research should be done to determine the risk factors linked to diabetes complications. Finally, we'd want to apply this concept to additional chronic conditions.

VI. Conclution

One of the most pressing worldwide health concerns is detecting diabetes risk at an early stage. This research aims to develop a framework for predicting who is affected and the cause of diabetes type 2 is represented as a percentage. Six machine learning classification algorithms were used in this work, and the results were compared to several statistical metrics. Tests were run on data acquired using online and offline questionnaires that included many diabetes-related topics. The AdaBoost Classifier's accuracy in our dataset is 98 percent, which is the greatest among the others, according to the testing results. All the models achieved good results for various parameters such as accuracy, precision, etc. among the six distinct machine learning methods used. This result can be used to forecast any other illness in the future. Other machine learning algorithms to predict diabetes or any other illness are currently being researched and improved upon in this work.

Acknowledgment

The authors would like to thank the Ministry of Higher Education for providing financial support under Fundamental Research Grant Scheme (FRGS) No. FRGS/1/2019/ICT03/UMP/02/2 (University reference RDU1901194) .

Reference

[1] Leyer, Michael, and Dijana Iloska. "Analysing cognitive reasoning of individuals type 1 Diabetes Mellitus to resolve hypoglycaemia." Obesity Medicine 22 (2021): 100330.

[2] Žilinskienė, Jolanta, Linas Šumskas, Dalia Antinienė, and Jolita Jonynienė."Interaction of parenting style of mothers and paediatric diabetes management." Baltic Journal of Sport and Health Sciences 1.120 (2021): 22-31.

[3] Jain, Bhavini, Nandeshwari Ranawat, Pankaj Chittora, Prasun Chakrabarti, and Sandeep Poddar. "A machine learning perspective: To analyze diabetes." Materials Today: Proceedings (2021).

[4] Abaker, Ali A., and Fakhreldeen A. Saeed. "A Comparative Analysis of Machine Learning Algorithms to Build a Predictive Model for Detecting Diabetes Complications." Informatica 45.1 (2021).

[5] Sahoo, Priyabrata. "Primitive Diabetes Prediction using Machine Learning Models: An Empirical Investigation." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.11 (2021): 229-236.

[6] Saxena, Ankur, and Shivani Chandra. "Automated Diagnosis of Diabetes Mellitus Based on Machine Learning." Artificial Intelligence and Machine Learning in Healthcare. Springer, Singapore, 2021. 37-56.

[7] Li, Jun, Pei Yuan, Xiaojuan Hu, Jingbin Huang, Longtao Cui, Ji Cui, Xuxiang Ma et al."A tongue features fusion approach to predicting prediabetes and diabetes with machine learning." Journal of Biomedical Informatics 115 (2021): 103693.

[8] Abokhzam, Asma Ahmed, N. K. Gupta, and Dipak Kumar Bose. "Efficient diabetes mellitus prediction with grid based random forest classifier in association with natural language processing." International Journal of Speech Technology (2021): 1-14.

[9] Gupta, Deepak, Ambika Choudhury, Umesh Gupta, Priyanka Singh, and Mukesh Prasad."Computational approach to clinical diagnosis of diabetes disease: a comparative study." Multimedia Tools and Applications (2021): 1-26.

[10] Wadhwa, Shruti, and Karuna Babber. "Artificial Intelligence in Health Care: Predictive Analysis on Diabetes Using Machine Learning Algorithms." International Conference on Computational Science and Its Applications. Springer, Cham, 2020.

[11] Kowsher, Md, Farhana Sharmin Tithi, Tapasy Rabeya, Fahmida Afrin, and Mohammad Nurul Huda."Type 2 diabetics treatment and medication detection with machine learning classifier algorithm." Proceedings of International Joint Conference on Computational Intelligence. Springer, Singapore, 2020.

[12] Chatterjee, Sudesna, Kamlesh Khunti, and Melanie J. Davies. "Type 2 diabetes." The Lancet 389.10085 (2017): 2239-2251.

[13] Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." *Procedia Computer Science* 167 (2020): 706-716.

[14] Basu, Sanjay, Karl T. Johnson, and Seth A. Berkowitz. "Use of Machine Learning Approaches in Clinical Epidemiological Research of Diabetes." *Current Diabetes Reports* 20.12 (2020): 1-19.

[15] Roobini, M. S., Y. Sai Satwick, A. Reddy, M. Lakshmi, D. Deepa, and Anitha Ponraj. "Predictive Analysis of Diabetes Mellitus Using Machine Learning Techniques." *Journal of Computational and Theoretical Nanoscience* 17.8 (2020): 3449-3452.

[16] Kowsher, Md, Anik Tahabilder, and Saydul Akbar Murad. "Impact-learning: a robust machine learning algorithm." *Proceedings of the 8th International Conference on Computer and Communications Management*. 2020.

[17] Hu, Gensheng, Cunjun Yin, Mingzhu Wan, Yan Zhang, and Yi Fang."Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier." *Biosystems Engineering* 194 (2020): 138-151.

[18] Larabi-Marie-Sainte, Souad, Linah Aburahmah, Rana Almohaini, and Tanzila Saba."Current techniques for diabetes prediction: review and case study." *Applied Sciences* 9.21 (2019): 4604.

[19] Ramesh, Jayroop, Raafat Aburukba, and Assim Sagahyroon. "A remote healthcare monitoring framework for diabetes prediction using machine learning." *Healthcare Technology Letters* 8.3 (2021): 45.

[20] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." *Procedia Computer Science* 165 (2019): 292-299.