Machine Learning Algorithms for Detecting Diabetes

Makayla dela Cruz
Department of Computer Science and Physics
Rider University
Lawrenceville, United States
delacruzmak@rider.edu

Kevin Perez
Department of Computer Science and Physics
Rider University
Lawrenceville, United States
perezespink@rider.edu

Tyler Duell
Department of Computer Science and Physics
Rider University
Lawrenceville, United States
duellt@rider.edu

Md Ali
Department of Computer Science and Physics
Rider University
Lawrenceville, United States
mdali@rider.edu

***Abstract* - Diabetes is the most common chronic disease which poses a threat to human life. There is no cure for the disease but early detection of diabetes is needed in order to prevent further complications. The goal of this research is to utilize machine learning algorithms to analyze the factors which contribute to a diabetes diagnosis, recognize patterns within those factors, and to find the most accurate machine learning algorithm to detect these outcomes. Machine learning is already widely used to detect diabetes as seen in the related works section. Through the research of various methods used to perform this experiment, we were able to take a different approach by comparing our methodologies and final results with the results of others in this field.**

First, the "PIMA Indian Diabetes" data set was preprocessed by filling in missing values, then the data set was split into a 70% training and 30% testing set. Next, certain features are extracted in order to improve the accuracy of the model. Specifically, the outcome feature was focused on to evolve the prediction accuracy. To analyze this, a heat map was used to visualize feature correction. The machine learning algorithms utilized in this experiment were Gaussian Naive Bayesian, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression; visualizations such as a confusion matrix and ROC curve showing AUC were implemented to analyze the accuracy of the models. Overall, it was concluded that the Gaussian Naive Bayesian prediction was the most accurate at 83.33%. Through this research, it was also determined through the data that pregnancy has one of the strongest correlations to a diabetes diagnosis, even more than insulin and blood pressure levels.

***Keywords* - diabetes, machine learning, algorithm, insulin, pregnancy, Naive Bayes, Random Forest, KNN, Logistic Regression**

I. INTRODUCTION

One of the most common chronic diseases which pose a heightened threat to human health and wellbeing is diabetes There are three distinct types of diabetes, Type 1, Type 2, and Gestational diabetes. In general, diabetes can be characterized by hyperglycemia; elevated blood glucose levels which is caused by the lack of or misuse of insulin production in the pancreas. As a result, increased blood sugar stays in the bloodstream which can lead to serious health problems. These serious health issues include long term damage and ultimate failure of different organs, especially the heart, eyes, kidneys, nerves, and blood vessels. Specifically, loss of vision, weight fluctuation, cardiovascular complications, susceptibility to certain infections, Charcot joints, and foot ulcers are only some of the health issues someone with diabetes may face. In addition to diabetes, those diagnosed with this chronic disease are more likely to also be diagnosed with cerebrovascular, atherosclerotic, cardiovascular, and peripheral disease. Overall, diabetes has become increasingly common in people's lives. In fact, in the past 20 years, the number of people being diagnosed with diabetes worldwide has more than doubled. Therefore, by 2040, it is predicted that more than 642 million people worldwide will be diagnosed with diabetes, in other words, one in ten adults will suffer from this chronic disease. There is currently no cure for any type of diabetes; thus, it is crucial to accurately detect this chronic disease earlier in life in order to control the symptoms easier. There are three main types of diabetes. [1,2, 3,4]

Type 1 diabetes is caused by an autoimmune reaction which prevents the pancreas from making insulin; a hormone needed in order for the body to process sugar and glucose which enters the cells and produces energy. Without insulin, sugar builds up in the bloodstream and does not enter the cells, causing blood glucose levels to rise. Type 1 diabetes is typically diagnosed during two specific peaks, childhood or adolescence between ages 4 and 7 or 10 and 14 years old; however, it can also develop in adults later in life. Overall, approximately 5-10% of people who have diabetes have type 1 and there is currently no way to cure it since the factors which trigger the disease are unknown. However, treatment can be utilized in order to manage the sugar in the blood using insulin to prevent complications [5, 6].

Type 2 diabetes accounts for more than 90% of people with diabetes. This chronic disease develops similar to type 1 diabetes, the pancreas does not produce enough insulin

which can cause blood glucose levels to increase. It is usually diagnosed later in life, especially after the age of 35. However, type 2 diabetes can also be diagnosed in childhood in children with obesity. People with Type 2 diabetes are at high risk for microvascular and macrovascular complications. There is also no cure for type 2 diabetes; however, a healthier lifestyle, diabetes medications, or insulin therapy can help control blood sugar levels. [7,8]

Gestational diabetes is defined as glucose intolerance of varying degrees which develops during pregnancy in individuals who have never been previously diagnosed with diabetes. This type of diabetes causes high blood sugar which can complicate the pregnancy and the health of the baby. Furthermore, it can increase the likelihood of developing type 2 diabetes later in life as well as obesity, and cardiovascular disease for both the mother and baby. [9,10].

As mentioned before, there is no cure for diabetes. However, with earlier detection, individuals diagnosed with this chronic disease may be able to prevent further complications which come with a diabetes diagnosis. This research aims to analyze the factors which contribute to a diabetes diagnosis, how closely they are related, and the overall outcome as a combination of these factors. The goal of this research is to utilize machine learning algorithms in order to analyze the factors which contribute to a diabetes diagnosis, detect patterns within the data set, and to better predict if someone has diabetes based on specific features.

Machine learning methods are already widely used for predicting a diabetes diagnosis; usually providing desirable results. As a preliminary experiment, we utilized the data set, "PIMA Indian Diabetes" from the data repository, Kaggle, originally from The John Hopkins University which was donated to the National Institute of Diabetes and Digestive and Kidney Diseases. This data set consisted of variables such as numbers of pregnancies, glucose levels, blood pressure, skin thickness, insulin, BMI, age, and the diabetes pedigree function. The outcome of these features is labeled as 0 or 1, 0 meaning the patient does not have diabetes and 1 meaning the patient does have diabetes.

In the initial analysis of the data, data cleansing techniques were used in order to fill missing values with the mean value for each feature. Then, the data was split into a training and testing set to train the machine learning algorithms to accurately predict whether or not a patient has diabetes
based on their features. In order to compare the prediction accuracy of each model, the Logistic regression, Classification, Naive Bayes, Random forest, and K-Nearest Neighbors machine learning algorithms were utilized. A confusion matrix, heatmap, and ROC curve were also used to visualize the performance of these models in accurately predicting whether or not a patient had diabetes. Overall, the Gaussian Naive Bayes machine learning algorithm produced the most accurate results.

II. RELATED WORK

In the analysis of a diabetes data by Sisodia and Sisodia in [11], the machine learning methods naive bayes, SVM, and decision trees were used to evaluate diabetes in pregnant women, just like in our experiment. This particular experiment was done on the Pima Indians Diabetes Database (PIDD) from the UCI machine learning repository, the same dataset we are using. In order to evaluate these methods, Precision, Accuracy, F-Measure, and Recall. ROC curves verify the results and Naive Bayes performed the best on this dataset with an accuracy of 76.3%.

Data visualization was excellent in this paper, presenting the quality of the algorithm's work in the forms of confusion matrices and classification reports, in the formats of both tables and bar graphs. We believe this experiment was done very well and their visualizations as well as method selection were incredible. While we also intend to use Naive Bayesian classification in part of our analysis, it would be smart to modify the analysis method slightly to try and increase the accuracy above 76.3%.

Dr. Kamrul Hasal [12] discusses a method of detecting diabetes through machine learning techniques involving a focus on mitigation of missing, invalid, and outlying data. In this experiment, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbor, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed, and the mean was used in place of missing values during preprocessing. This was reported to boost performance significantly, even more than when removing them with the outliers, which were already being excluded. The report showed that the XB algorithm actually outperformed the Naive Bayesian algorithm as well, with a phenomenal accuracy of 94.6%.

This paper is incredibly thorough and has a lot of strengths. There are a great number of different machine learning algorithms tested on their dataset, and the analysis of their performance comes in the form of both data-filled tables and helpful graphical representations. A myriad of preprocessing techniques and 5-fold-cross-validation were performed as well, which gave a lot of insight into how we should treat our data in our final experiment. In terms of weaknesses, this paper only analyzes one dataset. While it is the same dataset that we plan to use, and their analysis of the data was stellar, the accuracy of the MLPs could likely be improved even further with more data available. Combining this dataset with another in our research could compensate for this in our project and improve our results.

Experimentation in [13] by Aishwarya Mujumdar et al. uses another large number of algorithms in determining what the most effective approach to future data prediction was. Out of 12 tested algorithms for the same PID dataset, the logistic regression model came out on top as the most accurate with a great rating of 96% accuracy, and the least accurate was the Perceptron algorithm with a rating of 76%. Overall, their preprocessing and normalization of the data before the ML model was applied increased accuracy for ML models across the board, bringing them all to an acceptable level. When applying these models to the additional dataset used in the paper for experimentation, accuracy was even better, reaching

a minimum of 90% accuracy from the bagging algorithm and a maximum of 96% again from the logistic regression model. These researchers also pipelined the 6 most accurate models, and found that this gave the AdaBoost classifier the new highest accuracy of 98.8%. Overall, the research by this group was incredibly thorough and made great use of the many different algorithms at their disposal to show which models performed the best under different conditions. The theory behind the preprocessing and analysis techniques was also included, which was helpful in understanding how the data was being manipulated. However, the weak points lie in the fact that there were barely any visualizations for the data analysis. A graph and trend line was included for interpretation of the accuracy table, as well as a confusion matrix, but all of the other provided data was displayed in text. Additionally, a larger variety of post validation techniques could have been used, as only 3 were employed in this paper.

Jobeda Jamal Khanam and Simon Foo [14] apply 7 machine learning models for the prediction of diabetes on the same Pima Indian Diabetes Dataset, including a neural network (NN) consisting of two hidden layers. As most accurate out of all the applied models, it was able to reach a relatively high prediction accuracy of 88.6%. A strong point from this aspect of the paper is that Weka and a Jupyter Python notebook were used for the analysis and preprocessing, which can easily be replicated. The amount of visualizations provided for this analysis was also great, and performing 10-fold cross validation helps to ensure the reliability of the report. Overall, with this easily adaptable preprocessing and post analysis code, this paper showcases how effective a neural network machine learning algorithm can be for solving the problem of diabetes prediction. However, similar logistic regression, naive bayes, random forest, and ANN models achieved a much lower accuracy percentage (>80% for all except ANN, which achieved 88.57%) than they had in [13], suggesting that their methods are tailored more towards NN prediction.

The experiment in [15] conducted by S.Saru and S. Subashree shares a lot of similarities with [14], employing the use of Weka in order to preprocess the data. This paper also uses the Pima Indian Database, fitting in with our experiment and the rest of the referenced works. However, it also uses the bootstrapping resampling technique to enhance the accuracy before building their naive bayesian, decision tree, and KNN, SVM, and random forest learning algorithms. This bootstrapping was shown to significantly increase the accuracy of the decision tree and KNN algorithms, causing the jump from 78.43% and 69.93% to 94% and 93.7% respectively. Many of their algorithms are ones that we will be applying in our final experiments as well, and their effective result implies that bootstrapping before the application of our algorithms may greatly benefit us as well.

This paper falls short in the visualization section, only providing textual displays of the results. We intend to focus on this section of our paper to avoid this problem, and the effectiveness of their findings would be much higher if the authors of [15] had included graphical representations in their analysis as well.

IV. DATASET & EXPERIMENTATION

A. Experiment Setup

Our experiment will be contrasting the result of multiple machine learning algorithms in predicting the development of diabetes using a number of features from the PIMA Indian Dataset. As shown in Fig. 1, we will begin by collecting our data, performing preprocessing, extract important features, select our models, and visualize the results

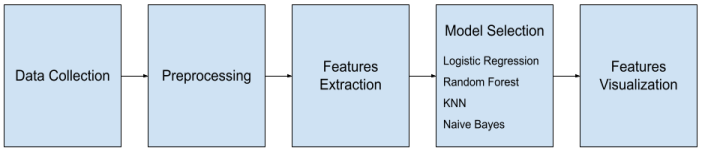

Fig. 1. Experiment Setup

B. Data Collection

The dataset we decided to use is from Kaggle.com and is named “PIMA Indian Diabetes”, or the PID dataset. It comes from The John Hopkins University which was donated to the National Institute of Diabetes and Digestive and Kidney Diseases. It exclusively contains data from pregnant women that are at least 21 years old who belong or are of Pima Indian Heritage. It contains various data features such as Pregnancies, the number of times pregnant, Glucose levels, Blood Pressure, Skin Thickness, Insulin levels, Body Mass Index (BMI), Diabetes Pedigree function, age and the Outcome (whether or not they are diabetic). We chose to use this dataset because of the large number of represented features, which we predicted would increase the accuracy of our algorithms.

C. Preprocessing

The main way we preprocessed the data was by filling in any spots containing missing values with the mean value for the column. Replacing missing or invalid values with the mean for the column helps us give the algorithms more chances to learn and predict, because there will be less instances that need to be excluded. There was not anything that we needed to do to normalize the data as the data structure does not require it.

We decided to split the dataset into a 70/30 ratio for our ML models.. 70% would be used for training and the remaining 30% for testing. We decided on these numbers to allocate a decent amount of data for testing without leaving too small of a portion for testing. Our data has one specific characteristic in terms of dimensionality. The outcome feature is very special because we can use this to evolve our prediction accuracy. This is because we have the outcome for Insulin levels, Body Mass Index and level of glucose independently, allowing us to identify the most important features in our data.

D. Features Extraction

Since feature extraction can improve model accuracy, we wanted to focus on the most important features. Those were initially thought to be insulin, glucose, skin thickness, and BMI. We thought this because insulin and glucose have a direct correlation with one another. Insulin is what regulates the glucose levels in your blood and if insulin levels are low then glucose cannot be regulated typically meaning the patient will be at an increased risk of becoming diabetic. Skin thickness and BMI were also thought to be important because a higher BMI increases the risk of becoming diabetic, and skin thickness increases in diabetic patients. Data on both of these will help our algorithms get a better picture of what features a diabetic person has.

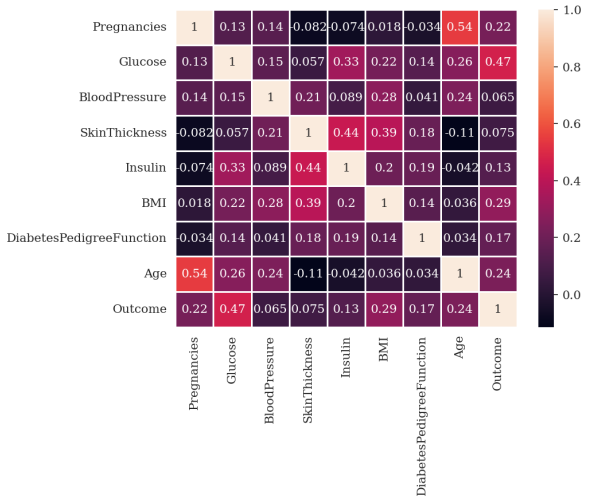

Fig. 2. Heat Map for Feature Correlation

Figure 2 shows the correlation between all of the features in our dataset, with glucose, BMI, age, pregnancies, and diabetes pedigree function being the most significant, in that order. This was surprising to us, as insulin shows relatively low significance when in relation to the outcome, and pregnancy is actually much more significant. This caused us to shift our view on the most important features in our dataset.

E. Model Selection

The models that we chose to use in our experiment in order to evaluate our data are the K-Nearest-Neighbor, random forest, logistic regression, and Gaussian Naive Bayes algorithms. We chose these models in order to explore different angles of analyzing our data and see which approach is most effective. We chose the KNN algorithm because our large number of features would allow us to see how different features would be grouped with the provided outcome, and which ones were insignificant in the prediction process. The random forest and logistic regression models were selected because of their effectiveness in regression analysis. The Gaussian Naive Bayes algorithm was selected because of its scalability to large numbers of features and requires a smaller amount of training data.

After running the experiment for each of the algorithms, we were able to obtain two low accuracy scores, one decent score, and one satisfactory score. Our KNN algorithm received an accuracy score of 65.36% only slightly behind the 69.69% achieved by our logistic regression algorithm. Our random forest algorithm came next with a much better accuracy of 77.71%, predicting diabetes correctly the majority of the time. Lastly, our Gaussian NB algorithm gave us a satisfactory score of 83.33%. We believe this algorithm worked the best because of its ability to adapt to a large number of features and obtain a good level of accuracy with less data than other algorithms.

F. Features Visualization

We calculated our highest prediction score using the Gaussian Naïve Bayes (GaussianNB) algorithm. Gaussian NB required us to fit our previous trained x and y values using the ".fit()" method in order to get an accurate score prediction of 83.33%. After this we were able to execute the predict command to show the current prediction for the data. From there we created several visualizations. All of the below visualizations will be for the Gaussian NB algorithm, as it was our best score.

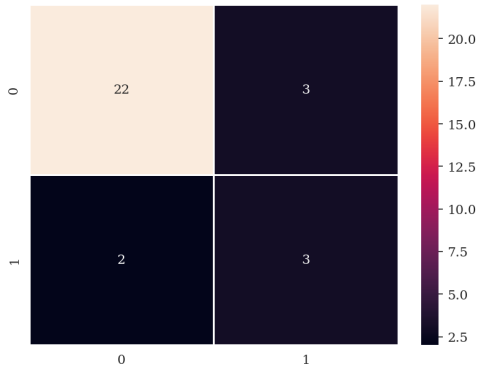

Fig. 3. Confusion Matrix

The confusion matrix (fig. 3) backed up the prediction score we received from GaussianNB, showing that we were able to get the majority of predictions correct. The top left of the confusion matrix shows the number of true positives (TP), while the bottom left shows the number of true negatives (TN). These are all of our correct predictions. The remaining two quadrants represent the incorrect predictions, the top right being false positives (FP) and bottom right being false negatives (FN).

We also created a Receiver operating characteristic curve (ROC) to display the accuracy. The ROC curve shows how well our prediction fared. (fig. 4) The closer the AUC or area under the curve score is to 1, the more accurate it is. If any line fell below the dotted line we would have to reevaluate, and the blue line represents the mean. We were able to get the best score with 4 fold ROC, which came up to

0.88. This is a great score, but the mean falls closer in line with our prediction accuracy overall with a score of 0.82.

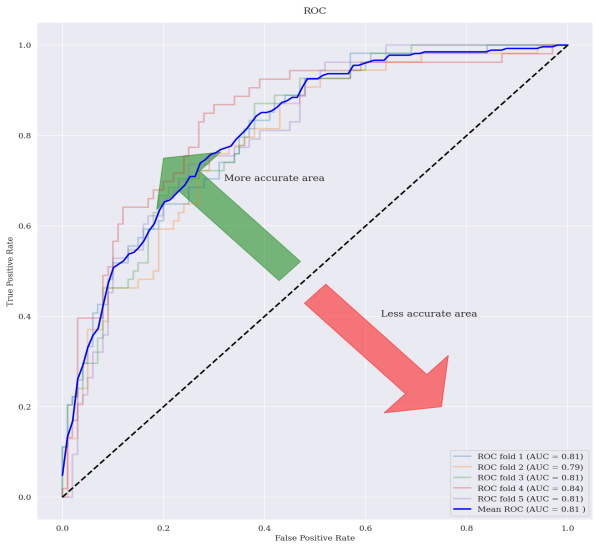

Fig 4. ROC curve showing AUC

The final visualization created was a classification report showing recall, precision, and f1-score (fig. 5). The recall value tells us what proportion of actual positive values were correctly identified. It is found using the formula, where TP indicates a true positive prediction and FN indicates a false negative prediction:

$$Recall = TP / (TP + FN)$$

The next value, precision, provides a similar value, showing what proportion of positive identifications were actually correct out of all positive predictions. This is done using the following formula, with FP indicating a false positive:

$$Precision = TP / (TP + FP)$$

The last value is f1-score, which functions as a combination of the previous two scores. It is the harmonic mean of the precision and recall, and found using the following formula:

$$F1\text{-}Score = 2(precision * recall) / (precision + recall)$$

The F1-score comes to an acceptable 0.81 for TP predictions and 0.63 for TN predictions, showing that our algorithm was much more accurate when guessing yes than no. This is likely due to the distribution of the data and there being more positive cases of diabetes than negative, thus leading to the algorithm getting much more experience with positive cases than negative ones.

	precision	recall	f1-score	support
0	0.82	0.80	0.81	493
1	0.62	0.65	0.63	245
accuracy			0.75	738
macro avg	0.72	0.73	0.72	738
weighted avg	0.75	0.75	0.75	738

Fig. 5. Gaussian NB classification report

V. CONCLUSION & FUTURE WORK

As can be seen from the algorithm's results, Gaussian Naive Bayes Prediction was able to achieve a relatively high accuracy of 83.33% compared to our other models. We believe that the KNN, logistic regression, and random forest models suffered in comparison to the Gaussian Naive Bayes model because of the size of our analyzed data. We believe that two options could have improved the model's accuracy. The first of these would be changing the amount of data sets we include, possibly incorporating an additional two. They would also need to be focused on women, sharing many of the same features as our dataset, or the correlation values for each feature and the prediction value would not be correct. The second option is changing the proportion of data used for training and testing. We currently have 70% dedicated to training our algorithms, with the remaining 30% reserved for testing. Increasing the training portion to 80% of our dataset and decreasing the testing portion to 20% could potentially improve the accuracy of this model as more data would be available to find relationships before having to make predictions. A possible combination of these solutions would also be beneficial, where the data set is trained entirely on our PID data and then applied on another for prediction. However, the features between the two must be consistent in order to have an accurate result.

Our Gaussian Naive Bayes Model performed relatively well given the limited amount of data. The analysis of this model is interesting to look at in particular as it shows how pregnancy has the fourth strongest correlation with the prediction, even more so than insulin levels and blood pressure. This model could still be improved, however, as the f-1 score comes in only at 0.81 for 0 and 0.63 for 1. The same solutions proposed for improving the accuracy of our logistic regression model would likely improve the accuracy of this model as well, especially training entirely on one dataset while testing on another with related features.

Our related works were still able to achieve a success rate greater than ours, some breaching 90% compared to our 83.33%, giving some insight to how future works could be conducted. Some algorithms that others have shown to be just as or more effective than our Gaussian Naive Bayesian approach include K-Nearest-Neighbor (KNN) and Support Vector Machine (SVM) algorithms, as well as the XB boost algorithm and neural networks. However, many of these experiments did not include the same features, focusing on different combinations of features to make predictions. Taking a collection of these databases and eliminating inconsistencies

within the features represented could result in a much more accurate prediction rate for any of the mentioned models, as well as bringing more light onto which features have the most weight in determining whether or not a person is diabetic. This work could also be extended to include only men, as they cannot become pregnant and thus may have very different correlations between their features and outcome predictions. A diversification of the data in terms of the regions it is collected from could also have an interesting impact on the prediction accuracy, as different cultures and environments have very different rates of disease development.

VI. REFERENCES

[1] "What is diabetes?," *Centers for Disease Control and Prevention*, 07-Jul-2022. [Online]. Available: https://www.cdc.gov/diabetes/basics/diabetes.html. [Accessed: 21-Mar-2023].

[2] American Diabetes Association. "Diagnosis and classification of diabetes mellitus." Diabetes care 34.Supplement_1 (2011): S62-S69.

[3]Zimmet, Paul Z., et al. "Diabetes: a 21st century challenge." The lancet Diabetes & endocrinology 2.1 (2014): 56-64.

[4] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[5]Atkinson, Mark A., George S. Eisenbarth, and Aaron W. Michels. "Type 1 diabetes." The Lancet 383.9911 (2014): 69-82.

[6] D. Daneman, "Type 1 diabetes," *The Lancet*, 09-Mar-2006. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0140673606683414. [Accessed: 30-Apr-2023].

[7] DeFronzo, Ralph A., et al. "Type 2 diabetes mellitus." Nature reviews Disease primers 1.1 (2015): 1-22.

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

[9] Buchanan, Thomas A., and Anny H. Xiang. "Gestational diabetes mellitus." The Journal of clinical investigation 115.3 (2005): 485-491.

[10] McIntyre, H. David, et al. "Gestational diabetes mellitus." Nature reviews Disease primers 5.1 (2019): 47.

[11] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.

[12] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857. [Accessed: 22-Mar-2023]

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

[14] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." ICT Express 7.4 (2021): 432-439.

[15] Saru, S., and S. Subashree. "Analysis and prediction of diabetes using machine learning." International journal of emerging technology and innovative engineering 5.4 (2019).