ISSN 2347 - 3983
Volume 11. No.6, June 2023
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter031162023.pdf
https://doi.org/10.30534/ijeter/2023/031162023

A Machine Learning Approach for Prediction of Diabetes Mellitus

Chilupuri Anusha[1]
[1]Assistant Professor, Department of Electronics and Communication Engineering,
G.Narayanamma Institute of Technology and Science,
Hyderabad, India.
chilupuri.anushareddy@gmail.com

Received Date: April 27, 2023 Accepted Date: May 18, 2023 Published Date : June 07, 2023

ABSTRACT

Diabetes Mellitus is among chronic diseases and lots of people are suffering with this disease. It may cause many complications and have a high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. There is no doubt that this alarming figure needs great attention. With the rapid development of Machine Learning, machine learning has been applied to many aspects of medical health. There are several Machine learning algorithms that are used to perform predictive analysis in various fields. Predictive analysis in healthcare is a challenging task but ultimately can help practitioners make data informed about a patient's health and treatment. In this project, for experiment purposes, we have taken a dataset which is originally from the National Institute of diabetes and digestive and kidney diseases. All patients here are females at least 21 years old of Pima Indian heritage. By studying the dataset, we must find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly.

The objective of this project is to diagnostically predict whether the patient has diabetes or not, based on certain diagnostic measurements included in the dataset. We have proposed a diabetes prediction model for better classification of diabetes by applying some popular machine learning algorithms namely, Logistic Regression, Random Forest Algorithm and KNN Algorithm to predict Diabetes.

Key words: Dataset, KNN Algorithm, Logistic Regression, Machine learning, Random Forest Algorithm, Diabetes Mellitus.

1. INTRODUCTION

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of kinds of diseases like heart attack, blindness, kidney diseases etc . The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit for a day or more to get their reports. But with the rise of Machine Learning approaches, we have the ability to find a solution to this issue [1].

Diabetes mellitus is an endless infection portrayed by hyperglycemia. It might cause numerous inconveniences. As per the developing bleakness as of late, in 2040, the world's diabetic patients will achieve 642 million, which implies that one of the ten grown-ups later is experiencing diabetes. There is no uncertainty this disturbing figure needs extraordinary consideration [2]. The World Health Organization has assessed 12 million passes around the world, consistently because of Heart maladies. A large portion of the passouts in the United States and other created nations are expected to cause cardiovascular maladies. The early visualization of cardiovascular illnesses can help in settling on choices on way of life changes in high hazard patients and thus decrease the intricacies [3]. This exploration means to pinpoint the most significant/hazard elements of coronary illness just as anticipate the general hazard utilizing calculated relapse. Machine Learning has been connected to numerous parts of medicinal wellbeing. In this project, we utilized Logistic regression, Random Forest and KNN Algorithm to anticipate diabetes mellitus.

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

2. LITERATURE SURVEY

The Pima Indians are genetically predisposed to diabetes, and it was noted that their diabetic rate was 19 times that of a typical town in Minnesota [4]. The National Institute of Health (NIH) originally owned the Pima Indian diabetes Database (PIDD). The number of patients (n) in the database n=768 each with 9 attribute variables. Out of the nine conditional attributes, six are due to physical examination, the rest of the attributes are chemical examination. Of these 9 attributes, there are eight inputs and the last one being the output. The goal is to use the first 8 variables to predict attribute values of the 9th variables.

Diabetes is an ailment caused due to the extended level of sugar obsession in the blood. In this paper, discussing various classifiers, a decision support system is proposed that uses the Logistic Regression algorithm base classifier for classification [5]. Moreover, Random tree, KNN have additionally executed as a base classifier for Logistic Regression calculation for exactness confirmation. The exactness got for Logistic Regression calculation with choices
stump as a base classifier is 82.72%, which is more noteworthy contrasted with that of Random tree, KNN and Artificial intelligence is having more effect is machine realizing, which creates calculations ready to take in examples and choice standards from information [6]. Machine learning calculations have been implanted into information mining pipelines, which can consolidate them with established measurable techniques, to remove learning from information. In the wake of having managed to miss information by methods for irregular woods and having connected appropriate methodologies to deal with class unevenness, we have utilized Logistic Regression, Random Forest, KNN Algorithms. With the stepwise component choice to foresee the beginning of retinopathy, neuropathy, or nephropathy, at various time situations, at 3, 5, and 7 years from the main visit at the Hospital Center for Diabetes (not from the conclusion). Considered factors are sexual orientation, age, time of determination, weight file (BMI), glycated haemoglobin (HbA1c), hypertension, and smoking propensity. Lust models, custom fitted as per the complexities, gave an exact up to 0.838. Diverse factors were chosen for every complexity and time situation, prompting models simply to mean the clinical practice [7]. Patients with diabetes should ceaselessly screen their blood glucose levels and modify insulin measurements, endeavoring to keep blood glucose levels as near typical as would be prudent. Blood glucose levels that veer off from the typical range can prompt genuine here and now and long-haul intricacies [8]. A programmed expectation shows that cautioned individuals of fast approaching changes in their blood glucose levels would empower them to make a preventive move. In this paper, we depict an answer that uses a bland physiological model of blood glucose progression to produce enlightening highlights for a regression display that is prepared with tolerant information.

3. PROPOSED SYSTEM DESIGN

The new model beats diabetes specialists at foreseeing blood glucose levels and could be utilized to envision right around a fourth of hypoglycaemic occasions 30 min ahead of time. Diabetes mellitus is a standout amongst the most genuine wellbeing challenges in both creating and created nations. As per the International Diabetes Federation, there are 285 million diabetic individuals around the world. This aggregate is relied upon to ascend to 380 million in 20 years. Because of its significance, an outline of a classifier for the recognition of Diabetes ailment with ideal cost and better execution is the need of the age. The Pima Indian diabetic database at the UCI machine learning research facility has turned into a standard for testing information mining calculations to see their expectation exactness in diabetes information arrangement. The proposed strategy utilizes Random Forest, a machine learning technique as the classifier for analysis of diabetes.

The machine learning strategy centers around arranging diabetes illness from a high dimensional therapeutic dataset. The trial came about to demonstrate that a vector machine can be effectively utilized for diagnosing diabetes illness [9]. The point of this examination is to the finding of diabetes illness, which is a standout amongst the most viral infections in the restorative field utilizing Generalized Discriminant Analysis (GDA) and Least Square KNN. Likewise, we proposed another course learning framework in light of Generalized Discriminant Analysis.

The proposed framework comprises two phases. The 1st stage, we have utilized Generalized Discriminant Analysis to discriminate highlight factors amongst sound and patient (diabetes) information as a pre-preparing process. The 2nd stage, we have utilized Random Forest so as to order the diabetes dataset. KNN acquired 78.21% grouping precision.
The proposed framework called GDA–LR got 82.05% order exactness utilization. The acquired order exactness is 82.05% and it is exceptionally encouraging contrasted with the beforehand detailed grouping strategies. A proper and thorough literature survey concludes that there are various methods that can be used to predict diabetes. Some of these approaches are Logistic Regression, Random Forest, and K-Nearest Neighbors.

3.1 System Architecture

The training data set is fed to the system as input which will be initially preprocessed. Data preprocessing is the phase where the raw data will be transformed into a meaningful and understandable format. Figure 1 shows the Architecture of Diabetes Prediction System.

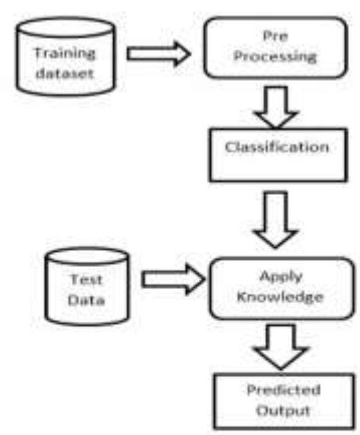

Figure 1: Architecture of Diabetes Prediction System

The pre-processed data will be later classified using the best classification mechanism. Then the classification will be compared with the test data in order to classify it accurately using some distance measures. The final classified data will be converted to data patterns using intelligent methods. The obtained patterns will be evaluated for accuracy and correctness. The identified patterns will be represented as knowledge in the required form as output.

3.2 Flowchart

Figure 2 displays the flowchart of the proposed system. Data Extraction for the proposed system. Raw Data will be preprocessed. Processed file will be divided into 85% of training data and 15% of test data. Model is created for the training data and saves the model. Results will be obtained for the test data using the saved model.

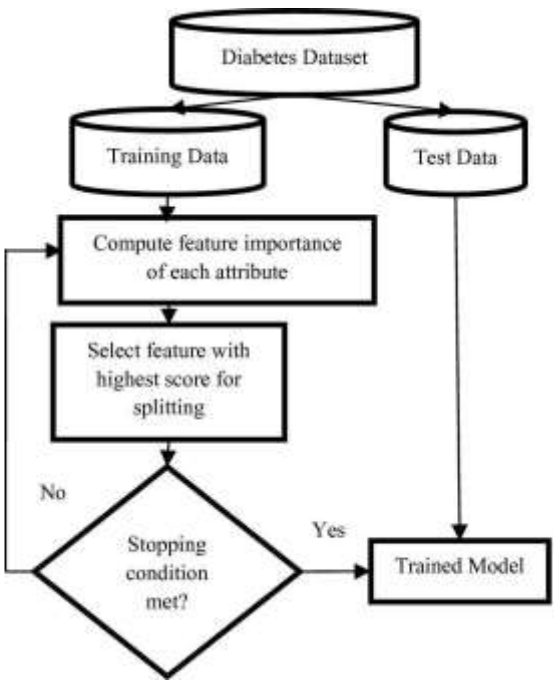

Figure 2: Flowchart of Diabetes Prediction System

Classification of Data of the proposed system. For the problem definition data will be collected and preprocessed. Model will be created for the gathered data and data access, data sampling, data transformation will be done. After creating the model and evaluating and interpreting it. Created model is applied for external applications. Next the Values for each attribute assigned based on importance for the input dataset. Process and generate value. If the value is equal to one then the patient is diabetic otherwise the patient is non diabetic.

3.3 Analysis

In this final phase, we will test our classification model on our prepared image dataset and also measure the performance on our dataset. To evaluate the performance of our created classification and make it comparable to current approaches, we use accuracy to measure the effectiveness of classifiers.
After model building, knowing the power of model prediction on a new instance, is a very important issue. Once a predictive model is developed using the historical data, one would be curious as to how the model will perform on the data that it has not seen during the model building process. One might even try multiple model types for the same prediction problem, and then, would like to know which model is the one to use for the real-world decision-making situation, simply by comparing them on their prediction performance (e.g., accuracy).

To measure the performance of a predictor, there are commonly used performance metrics, such as accuracy, recall etc. First, the most used performance metrics will be described, and then some famous estimation methodologies are explained and compared to each other. Performance Metrics for Predictive Modeling In classification problems, the primary source of performance measurements can be calculated by finding the accuracy.

4. IMPLEMENTATION

At the most basic, machine learning uses programmed algorithms that receive and analyse input data to predict output values within an acceptable range. As new data is fed to these algorithms, they learn and optimize their operations to improve performance, developing ‘intelligence’ over time.

Logistic Regression

Logistic Regression is a classification algorithm for categorical variables like Yes/No, True/False, 0/1, etc. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no) [10].

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Random Forest Algorithm

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, a random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting [11]. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. Working of Random Forest Algorithm is shown in figure 3.

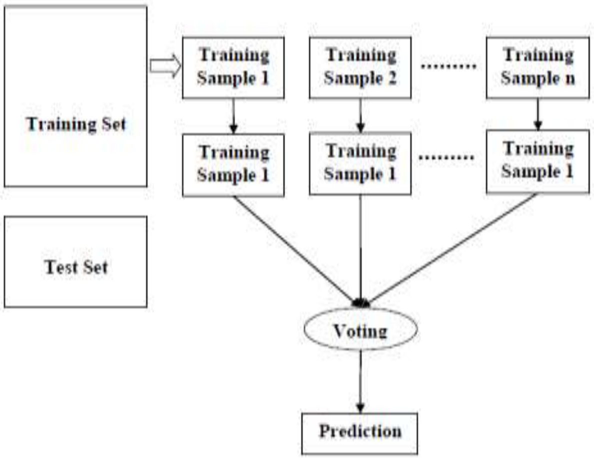

Figure 3: working of Random Forest algorithm

KNN Algorithm

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. We are often notified that you share many characteristics with your nearest peers, whether it be your thinking process, working etiquettes, philosophies, or other factors. As a result, we build friendships with people we deem like us. The KNN algorithm employs the same principle. Its aim is to locate all of the closest neighbours around a new unknown data point in order to figure out what class it belongs to. It's a distance-based approach. KNN calculates the distance from all points in the proximity of the unknown data and filters out the ones with the shortest distances to it. In order to correctly classify the results, we must first determine the value of K. We don't have a particular method for determining the correct value of K. Here, we will try to test the model's accuracy for different K values. The value of K that delivers the best accuracy for both training and testing data is selected. It is recommended to always select an odd value of K. The impact of selecting a smaller or larger K value on the model is shown in Figure 4.

Larger K value: The case of underfitting occurs when the value of k is increased. In this case, the model would be unable to correctly learn on the training data.

Smaller k value: The condition of overfitting occurs when the value of k is smaller. The model will capture all the training data, including noise. The model will perform poorly for the test data in this scenario.

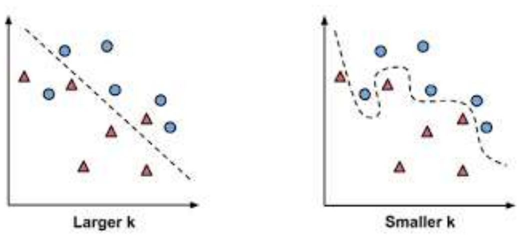

Figure 4: Impact of smaller K and Larger K

Dataset Attributes

There are 9 columns in the Dataset. Namely, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and Outcome. Here, the pregnancies column signifies how many times they were pregnant, Glucose being their Plasma Glucose Concentration on an (GTT), Blood Pressure being their diastolic Blood Pressure (mmHg) and so on and so forth.
Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test
Blood Pressure: Diastolic blood pressure (mm Hg)
Skin Thickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
Diabetes Pedigree Function: Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
Age: Age (years)
Outcome: Class variable (0 if non-diabetic, 1 if diabetic)
The following features have been provided to help us predict whether a person is diabetic or not:

Data Pre-processing
This phase of the model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So, we imputed missing values for a few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI, and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values. Also, the Diabetes Pedigree Function is a function which provides us with the likelihood of diabetes based on family history. The outcome column one target variable or class variable, which says if they had diabetes (1) or did not (0). We want to see what is the relation between some high-risk factors like the age, pregnancies and so on. we do not really need the skin thickness. Also, the Pedigree Function is one another value which is important but for the sake of simplicity Just moving forward we need to check if there is a fault in the dataset because while looking in the CSV file, we saw some values of BMI which are less than 10, it is not possible logically.

We also can see some outliers in the data that are close to 0 BMI which is a discrepancy, hence what we can do is to get rid of all the rows which have BMI less than 10.(BMI chart starts with 19), Similarly we see a lot of 0 values, needless to say, they are wrong data. all these values are zero in the entry hence we'll delete them to get cleaner data and for glucose, blood pressure etc can't be 0-20, well because it is not possible. Hence, we do a search across the columns of the data frame to find out if there's any such discrepancy.

This data consists of diagnostic information about some patients who have been tested for diabetes. Scroll to the right if necessary, and note that the final column in the dataset (Diabetic) contains the value 0 for patients who tested negative for diabetes, and 1 for patients who tested positive.
Steps of Implementation
Step 1: Import the Libraries and Load the Dataset is shown in figure 5.

Figure 5: Importing Libraries and Loading the dataset

Step 2: Dividing the input and output as shown in figure 6. The X variable contains all the columns from the dataset, except the outcome column, which is the target variable. The Y variable contains the values from the Outcome column. The X variable is our attribute set and Y variable contains corresponding values.

Figure 6: Dividing the input and output

Step 3: The final preprocessing step is to divide our data into training and test sets. The model selection library of Scikit-Learn contains the train_test_split method, which we'll use to randomly split the data into training and testing sets as shown in figure 7.

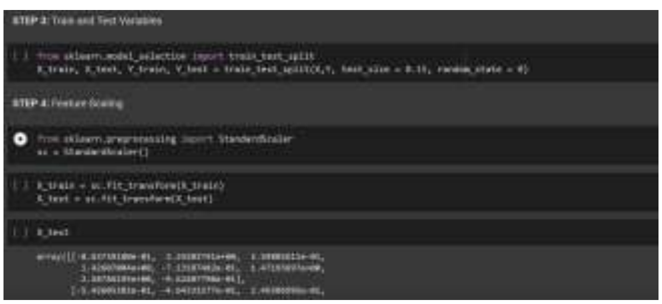

Figure 7: Training and Test variables, Feature Scaling

Step 4: Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. Feature scaling is essential for machine learning algorithms that calculate distances between data. The ML algorithm is sensitive to the "relative scales of features," which usually happens when it uses the numeric values of the features rather than say their rank. In many algorithms, when we desire faster convergence, scaling is a must like in ML.

Step 5: Build the Model.
Here various Algorithms are used to build a model

Logistic Regression
Figure 8 shows the training of Logistic Regression Model.

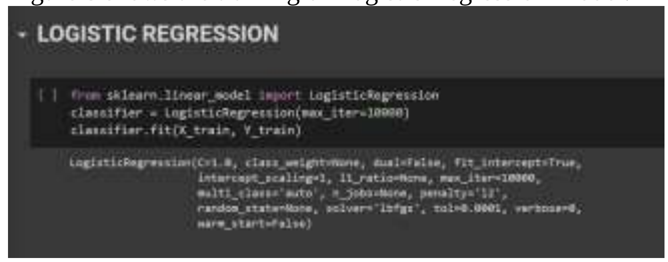

Figure 8: Logistic Regression Model Training

Random Forest Algorithm
Random forest Model training is shown in figure 9

Figure 9: Random forest Model Training

KNN Algorithm
Training in KNN model is shown in Figure 10.

Figure 10: KNN Model Training

5. RESULTS AND CONCLUSION

Output of the Logistic Regression algorithm when input list {} is given is as shown in Figure 11.

Figure 11: Output of Logistic Regression

Below Pie chart indicates percentage of Diabetic vs Non-Diabetic Individuals in the given dataset is displayed in figure 12.

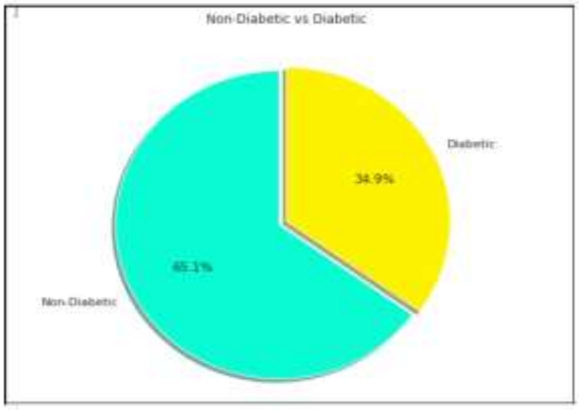

Figure 12: Non-Diabetic vs Diabetic

Accuracy of the Algorithms

After applying various Machine Learning Algorithms on the dataset, we got accuracies which was shown in figure 13, figure 14 and figure 15 for Logistic, Randon Forest and KNN algorithms respectively. and the values are tabulated in table 1 for quick comparison. Wherein Logistic Regression gives the highest accuracy of 82.7%.

Logistic Regression

Figure 13: Accuracy of Logistic Regression

Random Forest Algorithm

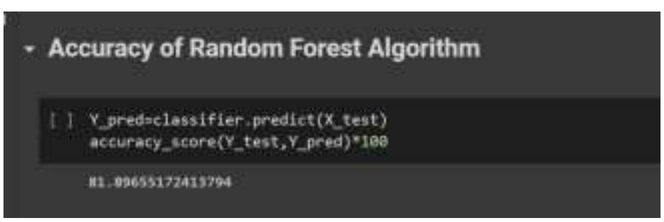

Figure 14: Accuracy of Random Forest

KNN Algorithm

Figure 15: Accuracy of KNN Forest

Table 1: Accuracy of Algorithms

Algorithms	Accuracy of the Algorithm
Logistic Regression	82.7%
Random Forest	81.8%
K-Nearest Neighbours	70.1%

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of those methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning methods in which KNN, Random Forest and Logistic Regression are used. And 82.7% classification accuracy has been achieved. The Experimental results can assist health care to take early predictions, make early decisions to cure diabetes and save humans life.

REFERENCES

[1] K. Vijiyakumar, B. Lavanya, I. Nirmala, and S. S. Caroline, *Random Forest Algorithm for the Prediction of Diabetes*.

[2] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.

[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front Genet*, vol. 9, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[4] IEEE Computational Intelligence Society, Institute of Electrical and Electronics Engineers, and B. IEEE World Congress on Computational Intelligence (2018 : Rio de Janeiro, *2018 IEEE Congress on Evolutionary Computation (CEC) : 2018 proceedings.*

[5] D. H. Kassem and M. M. Kamal, "Therapeutic efficacy of umbilical cord-derived stem cells for diabetes mellitus: a meta-analysis study," *Stem Cell Res Ther*, vol. 11, no. 1, Dec. 2020, doi: 10.1186/s13287-020-01996-x.

[6] S. M. Hasan Mahmud, M. A. Hossin, M. Razu Ahmed, S. R. H. Noori, and M. N. I. Sarkar, "Machine learning based unified framework for diabetes prediction," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2018, pp. 46–50. doi: 10.1145/3297730.3297737.

[7] R. Patil and S. Tamane, "A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 5, pp. 3966–3975, 2018, doi: 10.11591/ijece.v8i5.pp3966-3975.

[8] K. J. Rani, "Diabetes Prediction Using Machine Learning," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 294–305, Jul. 2020, doi: 10.32628/CSEIT206463.

[9] M. B. Digey, "DIABETES DETECTION USING CONVOLUTIONAL NEURAL NETWORKS (CNN) BASED MODEL," 2021.

[10] R. Agarwal, "Diagnosis of Diabetes Using Computer Methods: Soft Computing Methods f or Diabetes Detection Using Iris," 2017. [Online]. Available: https://www.researchgate.net/publication/318053905

[11] C. Anusha and K. Jyothi, "AN ARTIFICIAL INTELLIGENCE-BASED SPEAKING SYSTEM FOR DUMB PEOPLE," 2022. doi: 10.5281/zenodo.7940104.