Detection and Risk Analysis of Diabetes Using Machine Learning

Swapnil Karkhanis[1], Mohd. Wajahatraza Qayyumkhan Pathan[2], Yash Gandhi[3], Omkar Nalawade[4], Sarita Deshpande[5]

[1,2,3,4]*Student, Department of Information Technology, P.E.S.'s Modern College of Engineering, Maharashtra, India*
[5]*Head of Department, Department of Information Technology, P.E.S.'s Modern College of Engineering, Maharashtra, India*

---***---

Abstract - *Diabetes is a group of metabolic disorders which adversely affects the human body which impairs the body's ability to process blood sugar, also known as blood glucose. Detection of diabetes at an early stage and its reversal is the fastest and most efficient way to reduce the spread and growth of this disease. The current research is for developing an application for diabetes detection and risk analysis which implements a machine learning algorithm to analyze a patient's medical reports and conclude whether the patient is non-diabetic, diabetic or at risk and train a data model to analyze real-time inputs. The implemented machine learning model is trained on authentic medical reports provided by a diabetologist which contain values of various biomedical parameters such as HbA1c, fasting blood glucose, postprandial blood glucose, etc. This application is designed in such a way that it stores reports and diabetes detection results of all patients and provides the diabetologist with a rapid perception of their response to the current medical treatment which allows the diabetologist to modify the treatment and make it more effective.*

***Key Words*: Diabetes, Machine learning, Support Vector Machine (SVM), *K*-Nearest Neighbour (KNN), Artificial Neural Network (ANN), *K*-means Clustering Algorithm, Linear kernel, Radial Basis Function (RBF) kernel Diabetologist, Risk analysis, Blood glucose, HbA1c**

1. INTRODUCTION

Diabetes is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period of time.[3] Complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes which may occur if diabetes remains undetected. Diabetes can be classified into three different categories:

Type 1 diabetes: Type 1 diabetes was once called juvenile-onset diabetes as it often began in childhood. It is now called insulin-dependent diabetes as the pancreas fails to produce insulin, due to which a person has to take artificial insulin daily.[3]

Type 2 diabetes: Type 2 diabetes, also known as non-insulin-dependent diabetes or adult-onset diabetes occurs when the amount of insulin produced by the body is not enough or if the body is unable to effectively use the insulin produced by the pancreas.[3]

Gestational diabetes: Gestational diabetes is usually found in women during pregnancy as their bodies become less sensitive to insulin. This type of diabetes does not occur in all pregnant women and it usually resolves automatically after giving birth.[3]

The developed application utilizes the most effective machine learning algorithm for medically accurate diabetes detection and risk analysis. This application will also help diabetologists evaluate the effectiveness of the treatment a patient is currently undergoing to determine whether the treatment needs to be adjusted to improve the patient's condition.

2. LITERATURE REVIEW

The confidential dataset used for training the machine learning models was obtained from a renowned diabetologist whose clinic collects and stores reports of their patients through a mobile application. At the time of model training, the dataset contained 20,578 medical reports which contained medical parameters such as HbA1c, fasting blood glucose, postprandial blood glucose, urine sugar, urine ketones, insulin fasting, Vitamin D, Vitamin B12, etc. Out of these medical parameters, the most important parameters for diabetes detection are HbA1c, fasting blood glucose and postprandial blood glucose, which were considered while training the machine learning model.

HbA1c: HbA1c is a medical term which refers to glycated hemoglobin, which develops when hemoglobin joins with the glucose in the blood and becomes glycated.[4] By measuring HbA1c, diabetologists can understand the average blood sugar levels in a person's body over weeks or months. HbA1c is measured in percentage of glycated cells or mmol/mol (millimoles per mole).[5]

Fasting blood glucose: Fasting blood glucose is measured after a person fasts for eight hours by refraining from eating or drinking any fluid except for water. Fasting blood glucose is measured in mmol/l (millimoles per liter) or mg/dL (milligrams per deciliter).[6]

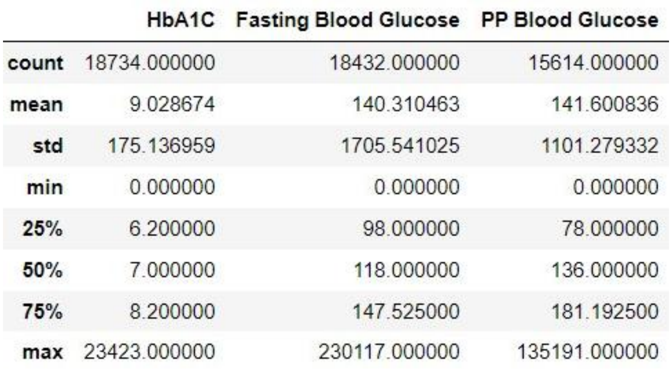
Postprandial blood glucose: Postprandial blood glucose is measured after a meal containing a set amount of carbohydrates, which shows how tolerant the body is to glucose. Postprandial blood glucose is measured in mmol/l (millimoles per liter) or mg/dL (milligrams per deciliter).[7]

In "Predictive Modelling and Analytics for Diabetes using Machine Learning Approach" by Harleen Kaur and Vinita Kumari, research was performed on the Pima Indians Diabetes Database, which is freely available on Kaggle.[1] This dataset contains 768 medical reports collected by the National Institute of Diabetes and Digestive and Kidney Diseases. In their research, five models were trained on the Pima Indians dataset by using Linear kernel SVM, Radial Basis Function (RBF) SVM, *K*-Nearest Neighbors (KNN), Artificial Neural Network (ANN) and Multifactor Dimensionality Reduction (MDR).[1]

Based on this research, they concluded that Linear kernel SVM was the most accurate machine learning model at detecting diabetes in a patient's medical reports.[1]

3. DATA PRE-PROCESSING

The dataset received from the diabetologist contained 20,578 medical reports and only three columns were considered (HbA1c, fasting blood glucose and postprandial blood glucose), but this data was unrefined and needed to be processed before being used for training a machine learning model to maintain maximum accuracy.



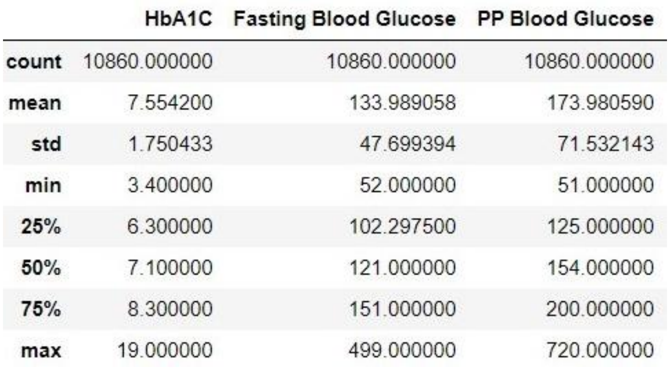
	HbA1C	Fasting Blood Glucose	PP Blood Glucose
count	18734.000000	18432.000000	15614.000000
mean	9.028674	140.310463	141.600836
std	175.136959	1705.541025	1101.279332
min	0.000000	0.000000	0.000000
25%	6.200000	98.000000	78.000000
50%	7.000000	118.000000	136.000000
75%	8.200000	147.525000	181.192500
max	23423.000000	230117.000000	135191.000000

Fig-1: Description of dataset before data pre-processing

The original and unedited dataset contained 20,578 reports which included null values, 0 (zero) values and unrealistic values which could be the result of human error while recording the data.

As shown in Fig-1, the dataset contains a varying number of rows for each parameter which contains data because some rows contain null values for certain columns. To process this data, the rows which contain at least one null value must be removed from the dataset.

Similarly, rows containing zero values and values which lie beyond feasible ranges must be dropped to keep only authentic and accurate data.



	HbA1C	Fasting Blood Glucose	PP Blood Glucose
count	10860.000000	10860.000000	10860.000000
mean	7.554200	133.989058	173.980590
std	1.750433	47.699394	71.532143
min	3.400000	52.000000	51.000000
25%	6.300000	102.297500	125.000000
50%	7.100000	121.000000	154.000000
75%	8.300000	151.000000	200.000000
max	19.000000	499.000000	720.000000

Fig-2: Description of dataset after data pre-processing

As shown in Fig-2, the dataset now contains only 10,860 reports out of the original 20,578 medical reports received from the diabetologist. These reports are accurate, feasible and suitable for training a machine learning model which will detect diabetic status of real-time inputs.

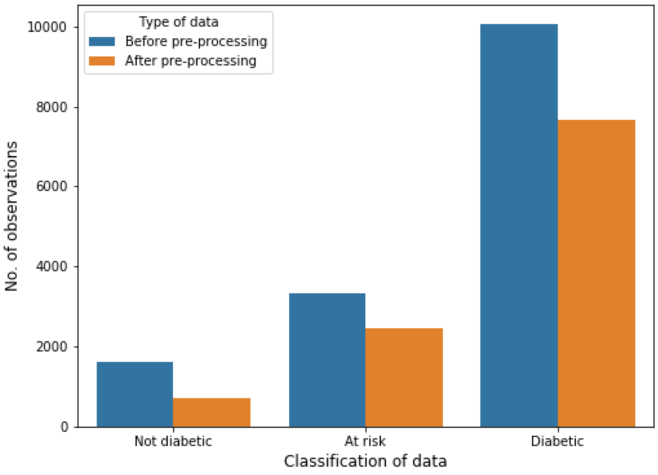


Fig-3: Representation of classification of reports before and after data pre-processing

Fig-3 provides a graphical representation of how many medical reports in the dataset were classified as non-diabetic, at-risk or diabetic before data pre-processing and how many reports were classified as non-diabetic, at-risk and diabetic after data pre-processing.

Fig-3 also illustrates the importance of data cleaning and pre-processing without which, 9,718 invalid and inaccurate reports would have also been implemented in machine learning model training which would have affected the accuracy and performance of the data model.

4. DESIGN AND ARCHITECTURE

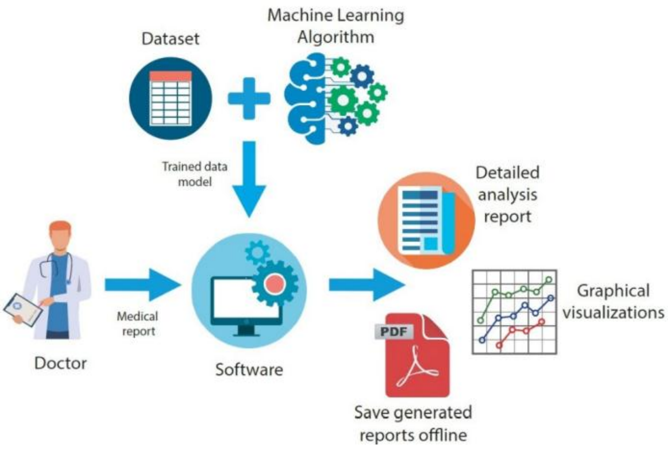


Fig-4: Architecture diagram

In the developed application, the cleaned dataset which was obtained after the data pre-processing phase is used to train a machine learning model which provides maximum accuracy by using the most effective machine learning algorithm. This model was integrated into the application so that it could be used with a front-end interface for accepting medical reports of a patient.

These reports are stored for future use and are passed to the machine learning model which then analyses the values and determines whether the patient is diabetic, at-risk or not diabetic.

The application also generates a risk analysis report based on the medical parameter values. This report contains data visualizations generated using the same patient's previous medical history for easier understanding of a patient's overall medical health. Such visualizations also assist the diabetologist in determining the effectiveness of the patient's medical treatment and adjust it if necessary.

5. IMPLEMENTED TECHNOLOGIES

5.1. Machine Learning

Machine learning is one of the most powerful technologies in the world. The purpose of machine learning is the construction of algorithms that can adapt and learn from their experience.[8] Machine learning tasks are simply classified into three broad categories:

Supervised learning: In supervised learning, the system infers a function from labeled training data.[2]

Unsupervised learning: In unsupervised learning, the learning system tries to infer the structure of unlabeled data.[2]

Reinforcement learning: In reinforcement learning, the system interacts with a dynamic environment.[2]

5.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is used in both classification and regression. In the Support Vector Machines, the data points are represented on the space and are categorized into groups and the points with similar properties fall in the same group.[1] Support Vector Machines implement numerous kernels such as Linear, Radial Basis Function (RBF), Laplace RBF, Sigmoid, Anove RBF, Polynomial and Gaussian to deal with non-linearity and higher dimensions.[9]

5.3. *K*-Nearest Neighbor (KNN)

The *K*-Nearest Neighbors (KNN) is an easy-to-implement and simple supervised machine learning algorithm which can be used for both regression and classification tasks. KNN postulates that similar entities exist in proximity.[10] For every entity, KNN calculates its distance from every other entity, sorts them in ascending order of the distances and classifies the first *k* elements into the class to which the maximum number of the first *k* elements belong to. [10]

5.4. Artificial Neural Network (ANN)

Artificial neural network mimics the functionality of the human brain. It can be seen as a collection of nodes called artificial neurons. All of these nodes can transmit information to one another. The neurons can be represented by some state (0 or 1) and each node may also have some weight assigned to them that defines its strength or importance in the system.[1]

The structure of ANN is divided into layers of multiple nodes; the data travels from the first layer (input layer) and after passing through middle layers (hidden layers) it reaches the output layer. Every layer transforms the data into some relevant information and finally gives the desired output.[1]

5.5. *K*-means Clustering

K-means Clustering Algorithm is an iterative algorithm that splits data into *k* predefined, distinctive, clusters (groups), where each observation in the data belongs to only one group. To split the data into *k* clusters, centroids are selected at random from the entire dataset.[11]

After the centroids have been finalized, the square distance between the remaining data points and the centroids is calculated and the data points are assigned to clusters whose centroid is at a minimum distance from them. Multiple iterations can be added to this process to ensure that the data points are always assigned to the accurate and same clusters.[11]

6. RESULTS AND EVALUATION

Predictive Models	Evaluation Parameters			
	Accuracy	Precision	Recall	F1 Score
SVM (Linear)	0.99	1	1	1
SVM (RBF)	0.81	0.80	0.81	0.80
KNN	0.78	0.77	0.78	0.78
K-means	0.49	0.74	0.50	0.51
ANN	0.93	0.93	0.93	0.93

Table-1: Evaluation of different models trained on the pre-processed dataset

After the dataset was pre-processed, the most accurate machine learning model had to be chosen for integrating into the application which takes user input of medical values and uses the chosen model to analyze the data and detect diabetes in the patient's medical reports. According to the study performed by training and evaluating multiple machine learning algorithms, as shown in Table-1, the model trained by using Linear kernel Support Vector Machine is the most accurate as compared to the rest. As a result, Linear kernel SVM model is implemented in the application to diagnose a patient's records and provide the diabetologist with the required analysis.

7. ADDITIONAL FUNCTIONALITIES

In the designed application, as shown in Fig-4, the machine learning algorithm (Linear SVM) was used to train a machine learning model using the pre-processed dataset and was used at the backend to accept real-time medical reports from the diabetologist.

All medical reports are stored in a database along with a unique identification key for each patient and whenever a new report is entered by the diabetologist, the report is analyzed by the machine learning model and the diabetic status and risk analysis of that report are evaluated.

Using previous medical reports of the same patient, a graph is generated for each medical parameter that illustrates how the values of each medical parameter have changed over time due to the current medical treatment to help the diabetologist understand the effectiveness of the current treatment and adjust the medications and prescribed routine if necessary.

Simultaneously, a comprehensive summary for the current medical report is generated in PDF (Portable Document Format) which contains a medical report, the generated graphs and information about each parameter and this summary is sent to the patient at his registered email address.

8. CONCLUSION

Detection of diabetes at an early stage and its reversal is the fastest and most efficient way to reduce the spread and growth of this disease, which will help to contribute in saving countless lives. Structural and methodical development of an application to detect and analyze the risk level of diabetes is an effective approach to tackle this disease and support the diabetologists in detecting a patient's vulnerability to this disease and finding the optimal treatment for them and machine learning is an immensely effective tool for achieving this goal.

REFERENCES

[1] Kaur, H., Kumari, V., Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach, Applied Computing and Informatics (2018). https://doi.org/10.1016/j.aci.2018.12.004

[2] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research. https://doi.org/10.1016/j.csbj.2016.12.005

[3] Dansinger, M. (2019, December 13). Types of Diabetes Mellitus - Causes of Diabetes. WebMD. https://www.webmd.com/diabetes/guide/types-of-diabetes-mellitus

[4] HbA1c: Hemoglobin A1c (HbA1c) Test for Diabetes, (2018, November 01). Retrieved from: https://www.webmd.com/diabetes/guide/glycated-hemoglobin-test-hba1c

[5] Guide to HbA1c, (n.d.). Retrieved from: https://www.diabetes.co.uk/what-is-hba1c.html

[6] Fasting blood glucose: Fasting Blood Sugar Levels, (2019, January 15). Retrieved from: https://www.diabetes.co.uk/diabetes_care/fasting-blood-sugar-levels.html

[7] Postprandial blood glucose: Postprandial Plasma Glucose Test, (n.d.). Retrieved from: https://www.diabetes.co.uk/diabetes_care/postprandial-plasma-glucose-test.html

[8] Edwards, G. (2018, November 18). Machine Learning | An Introduction. Towards Data Science. https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0

[9] Zoltan, C. (2018, November 13). SVM and Kernel SVM. Towards Data Science. https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200

[10] Harrison, O. (2018, September 11). Machine Learning Basics with the K-Nearest Neighbors Algorithm, Towards Data Science. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[11] Dabbura, I. (2018, September 17). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks, Towards Data Science. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

BIOGRAPHIES

Swapnil Karkhanis
Student,
Department of Information Technology,
P.E.S.'s Modern College of Engineering, Pune,
Maharashtra, India.

Mohd. Wajahatraza Qayyumkhan Pathan
Student,
Department of Information Technology,
P.E.S.'s Modern College of Engineering, Pune,
Maharashtra, India.

Yash Gandhi
Student,
Department of Information Technology,
P.E.S.'s Modern College of Engineering, Pune,
Maharashtra, India.

Omkar Nalawade
Student,
Department of Information Technology,
P.E.S.'s Modern College of Engineering, Pune,
Maharashtra, India.

Sarita D. Deshpande
Head of Department,
Department of Information Technology,
P.E.S.'s Modern College of Engineering, Pune,
Maharashtra, India.