Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

Predictive Analysis of Diabetes Miletus Using Machine Learning

Sneha Shetty R[1], Ashwitha A[2], SarithaM[3]

[1]*Assistant Professor, Dept of ISE, VVCE, Mysore, India(AffiliatedtoVTU)*
[2]*Assistant Professor, Dept of ISE, MSRamaiah Institute Of Technology,Bangalore, India(Affiliatedto VTU)*
[3]*Assistant Professor, Dept of CSE, SDMIT, Ujire, India(AffiliatedtoVTU)*
sneharshetty17@gmail.com[1],ashwitha.a.1990@gmail.com[2],saritha.shetty85@gmail.com[3]

Abstract
Diabetes is a chronic disease or metabolic disease group where someone suffers from an elevated amount of blood glucose inside the body which is inadequate to generate insulin, or where insulin is not adequately reacted to the cells of the body.The constant diabetes hyperglycaemia is linked to long-term damage, breakage, and breakdown of the different organs, especially the kidneys, eyes, heart, veins, and nerves.This research aims to use imperative attributes, construct a forecasting algorithm with machine learning, and discovery the best assorted to provide the closest result compared to clinical results.The suggested approach is designed to rely on predictive analysis to identify the characteristics for early diagnosis of Diabetes Miletus.The selection of optimal features from datasets is often extended to increase the precision of classification.

1.INTRODUCTION

This project is used to predict diabetes based on daily monitoring of food consumption. A system that gives an early alert may be helpful to reduce the workload of a physician. The patient will also access simple diabetes facts and recommendations on diabetes that will be controlled in the future. The main aim of the project is to predict diabetes and prevent it from being diabetic [1].

Health conditions are maybe one of the critical issues affecting our community's well-being directly.Diabetes Mellitus disorders are one of the main health issues faced by community residents.This project aims to build systems for reducing the time between the doctor and the patient to decide whether they suffer from diabetes through symptomatic selection1[2].

1.1. Problem Description

Diabetes is known to be unique of the greatestlethalin additionlong-lasting disorders that induce plasma sugar to grow.Several risks will be induced if diabetes remains unidentified and untreated. The tedious identification procedure leads to the person with diabetes approaching medicalhelp andaccessing the doctor.However, increasing the use of machine learning methods helps to resolve this important challenge.This research aims to design a model that can predict with full precision the probability of diabetes in patients.Consequently, in this experiment, three algorithms of machine learning classified, including Naive Bayes, KNN, and SVM are used for early diabetes detection dependent on food consumption.

2. LITERATURE REVIEW

Munam Ali Shah, Wajeeha Hamid, Nasir Kamal, and Muhammad Azeem Sarwar, “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare” [3]. This article addresses predictive analysis in healthcare, and this study includes six different algorithms of machine learning.A dataset of the medical background is collected for experimental purposes and various diverse machine learning algorithms are employedwith the data source. The efficiencyas well asexactness of the applied techniques were reviewedand also concomitant.

DeeptiSisodia,Dilip Singh Sisodia, “Prediction of Diabetes using ClassificationAlgorithms” [4]. This paper evaluates the efficiency of all three algorithms for different measurements such as accuracy, precision, recall as well as F-measure.Accuracy is calculated against instances appropriately and inappropriately classified.The results achieved indicate that Naive Bayes exceeds other algorithms with higher accuracy of 76.30%.These findings are verified using properly and systematically with ROC (“Receiver Operating Characteristic”) curves.

Prof. Pramila M. Chawan, Tejas N. Joshi, “Diabetes Prediction Using Machine Learning Techniques” [5]. The project aims to predict diabetes through three various supervised machine learning techniques, such as ANN, Logistic regression, and SVM. The goal of this project is also to suggest an efficient method for early

diabetes prediction.The method will also help researchers to construct a precise and efficient mechanism for improved decision-making regarding the disorder condition.

Parthajeet Ghosh, Debpriyo Paul, and Debadri Dutta, "Analysing Feature Importance for Diabetes Prediction using Machine Learning" [6]. In this article, we can learn about the essential elements for the occurrence of diabetes.We would also focus on the most important characteristics to determine whether an individual will be able to become diabetic in the future.We may infer that Random Forest is the best optimal predictor for diabetes, providing approximately 84 percent accuracy.One should maintain lower glucose levels and adopt a healthy diet with increasing age to avoid diabetes.

Sushant Ramesh, H. Balaji, N.Ch. S.N. Iyengar1 and Ronnie D. Caytiles, "Optimal Predictive analytics of Pima Diabetics using Deep Learning" [7]. Using python, the deep neural network is coded which helps to get numerical results on the diabetic severity and risk factor in the data collection. Finally, the introduction of this model on the type1 diabetes mellitus, Pima Indians diabetes, and the rough set theory model is being performed in a comparative analysis.The contrast reveals that the deep learning models are certainly more accurate than the rough set theory model.

MinyechilAlehegn, and Rahul Joshi, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach" [8].J48, Random Forest, Naive Bayes, and KNN,are the most widely employed predictive algorithms here.The single algorithm offered less precision than ensemble one.The decision tree was highly accurate in most of the tests.Java and Weka are the tools in this hybrid study for predicting diabetes data.

RonakSumbaly, S. Jeyalatha, and AiswaryaIyer, "Diagnosis of Diabetes using Classification Mining Techniques" [9]. This paper would seek strategies for the diagnosis of the disease by analyzing the correlations contained in the data by using Naive Bayes as well as Decision Tree algorithms in classification analysis.The efficiency of the suggested model is seen by experimental findings.For the diabetes diagnosis issue, the efficiency of the methods has been examined.

Aishwarya. R, Gayathri. P and N. Jaisankar, "A Method for Classification Using Machine Learning Technique for Diabetes" [10]. SVM ("Support Vector Machine") is one of the promising techniques in machine learning. SVM has been used for system classification. The classification system was given by SVM Upshot.In the pre-processing phase, the precision of the proposed method is strong as compared to previous work performed without pre-processing.In diabetes classification, pre-processing has a crucial role.

Dr.Naeem Khan, and Uswa Ali Zia, "Predicting Diabetes in Medical Datasets Using Machine Learning Techniques" [11]. In this research, a medical review of bioinformatics for the prediction of diabetes was carried out.The WEKA program was used to diagnose diabetes as a mining tool.In this analysis we plan to use the bootstrapping resample procedure to increase precision, then use and compare the results to check their efficiency with kNN ("k Nearest Neighbors"), Decision Trees, and Naive Bayes. Accuracy may be increased by optimizing data efficiency, algorithms, or even tuning the algorithms.

DharmaiahDeverapalli, PanigrahiSrikanth "A Critical Study of Classification Algorithms Using Diabetes Diagnosis"[12].The classification algorithm examines the Byes algorithm and Rule-based algorithm and Decision tree in this article. To evaluate their classification algorithms, common classifying algorithms were examined and performance metrics can be used to determine reliable outcomes in the Pima dataset classifying diabetes of pregnant patients.

There are different solutions available to find associations between the signs, diseases, and medicines, but these algorithms have drawbacks of their own; continuous arguments, high computational time, and numerous iterations, etc. Naive Bayes overcomes several constraints and can be used in real-time on a massive dataset.Some concerns must still be tackled, including governance issues such as ownership, protection, and privacy.By solving the above constraint, the analysis will contribute to faster growth.

We must first formulate or describe the problem before we try to resolve the issue.The issue you are intending to solve is necessary to describe exactly.Problem formulation is the act of a problem, determining the cause of the problem and, identifying the solution.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

3. PROBLEM STATEMENT

People these days have difficulty searching for a doctor or undertake certain medical tests to keep their body-conscious of their well-being as the workload rises, contributing to a loss of time.That's why we design a mobile diabetes monitoring method.This application will further minimize doctor-patient time.Health conditions in our country in particular diseases associated with blood disorders are growing increasingly these days.

There are different forms of blood disorder illness which include leukemia, anemia, diabetes, hemophilia, blood cholesterol, cancer, HIV/AIDS, etc. around the globe, about 400 million people are affected by Diabetes Mellitus. This chronic condition impacts hundreds of thousands of individuals.These systems have been designed to recognize their health conditions.
The objectives of the proposed project are as follows:

- The objective is to increase awareness about the importance of diabetes as a global public health issue.
- Appealing for diabetes prevention as well as control in underprivileged communities.
- To diagnose people with diabetes at an early stage based on food consumption.
- The significance of lifestyle in diagnosing patients in diabetes treatment and preventing complexities, in particular fitness, and diet.

Serious steps must be taken to eliminate the effects of diabetes at an early stage and also helps to minimize the number of patients who are suffering from diabetes disorder. Apart from that, once anyone suspects now that they are suffering from diabetes, they should concentrate on avoiding problems, like blindness, the renal disease that needs dialysis, amputation, or even death. Moreover, to avoid the occurrence of diabetes it is necessary to have a balanced diet.

4. REQUIREMENTS AND METHODOLOGY

4.1 Hardware Requirements

The Requirements for hardware for the proposed project are shown in the table below.

Table 4.1: Hardware Requirements

Sl. No	Hardware / Equipment	Specification
1	RAM	5GB (Minimum)
2	Processor	I5 processor or above
3	Windows	7 or above

4.2Software Requirements

The software requirements for the proposed project are depicted in the table below:

Table 4.2: Software requirements

Sl. No	Software	Specification
1	Operating system	Windows 10
2	Server	Jupyter Notebook

4.3Methodology Used

It is the methodical conceptual investigation of the proceduresfunctionaltowards the area of learning. It encompasses the speculativeinvestigation of the organisation of procedures and ideologiesconnectedthrough a division of knowledge. It displays the stream of the examinationshown in raising a model.

4.3.1Steps followed for implementing the Proposed Project

Step 1: Input
Load the file/dataset into the tool for pre-process step.
Step 2: Data Pre-processing
Select the classifier, in that choose algorithms.
Step 3: Segmentation

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

In test options, click percentage split options and also give the percentage split option and also give the percentage for dividing the data set into training set and test set Example: Training set =80% and Test set=20%.

Step 4: Output

After completing all the above steps, click start. The results are displayed in the screen. It shows the correctly classified instances, incorrectly classified instances, prediction on test set, total number of instances and confusion matrix.

4.3.2Working Procedure

Step 1: The training set used is the dataset which is a pre-loaded dataset on the tool. An existing training dataset can be accessed through the tool. Various limitations were placed while choosing the instances from a much larger dataset.

Step 2: The device is utilized for information pre-handling. As a first step the dataset must be pre-processed as the data obtained might be incomplete, noisy and dirty. The data might lack attributes, have errors and outliers and could be inconsistent.

Step 3: The process of automatically selecting only those features of the given data that contributes the greatest to the estimateflexible or outcome in which we are involved is known as Feature Selection.

Step 4: This step applies the various supervised classification algorithms namely, Naïve Bayes, and Support Vector Machine (SVM), on the training dataset that is obtained after following the first to third steps.

Step 5: This is the final step of comparing and analysing the accuracy measures and performance on the training dataset by the machine learning algorithms mentioned above.

5. SYSTEM DESIGN

5.1System Design

System design is a one important phase in software or system development. System design can be defined as method of defining different modules required for software or system to fulfil all requirements.

5.1.1Architecture of proposed system

System design shows the general design of the proposed diabetes prediction model, which is shown in figure shown below.

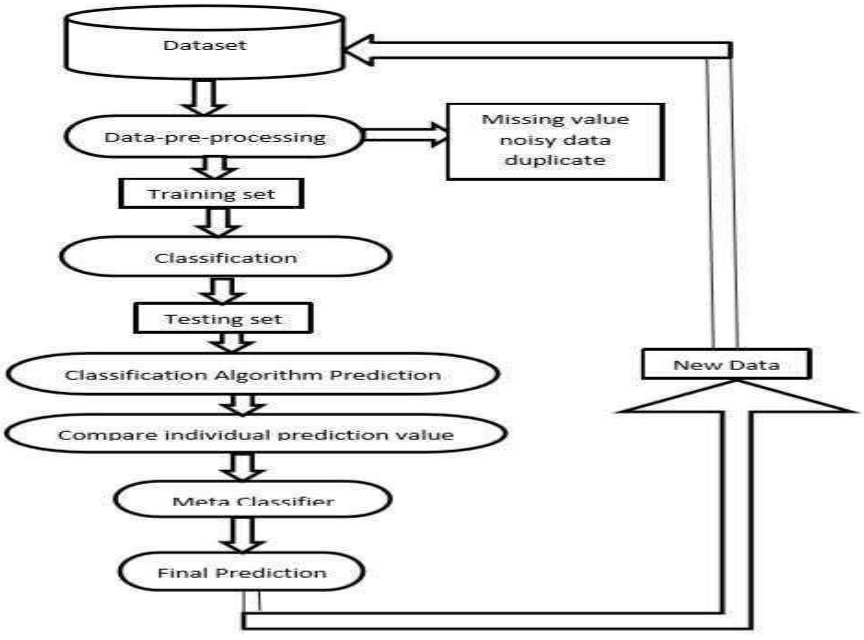

Figure 5.1.1: System Design

5.1.1.1Phases in Diabetes Prediction

The diabetes prediction organisation is skilledby means of supervised learning approach in which it takes dataset of diverse food consumption. The organisationcomprises the training and testing step trailed by pre-

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

processing data, classification algorithm prediction, comparing individual prediction values, and Meta classifier. Phases for implementing the Proposed Project are:

1. ML model is created based on height, weight and age.
2. JavaScript backend with a HTML and CSS in the front end to test input data and to see the performance of the model
3. Daily Calorie/ Sugar recommender for people based on their age, weight, height and giving a management of their daily calorie/ sugar needs based on the food that they eat

5.1.1.2Sequence Diagram

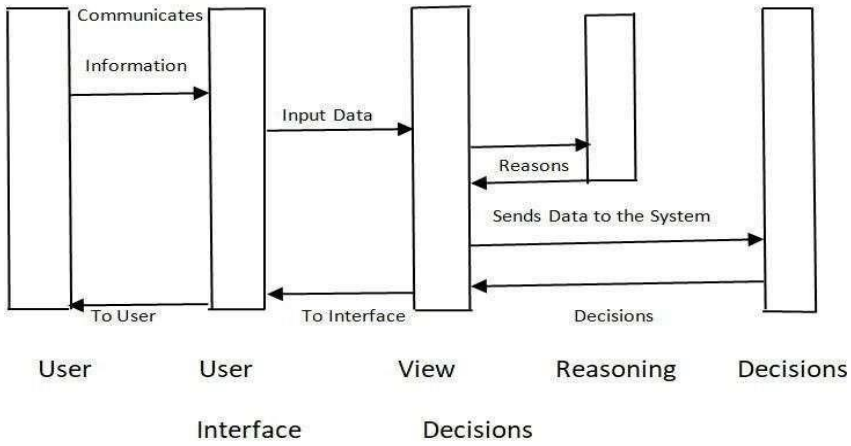

Figure 5.1.1.2: Sequence diagram

6. IMPLEMENTATION

6.1Pseudocode

This is a line by line procedure that can be translated into the programming language.

```
recursion(ts_map, S)
{
  m ←the length of ts_map //m is the number of the elements in ts_map.
  S̄ ←S/m        // the mean length of time series
  cutNumber←0// the number of the sets of time series not to be compressed in this round

  mark ←false
  do for tranverse ts_map
     do ts ← one set of time series
     do if the length of ts ≤ S̄
        then do if mark = false
          then do mark ←true
        do cutNumber ←cutNumber+ the length of ts
          Remove this ts from ts _map;
  if mark = false
     then do return S;
  else
      do S1 ←S- cutNumber;
         return recursion(ts_map, S1);
}
```

6.2System Flowchart

The flowchart for the given system is depicted in the following figure5.1.1.3 shown below:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

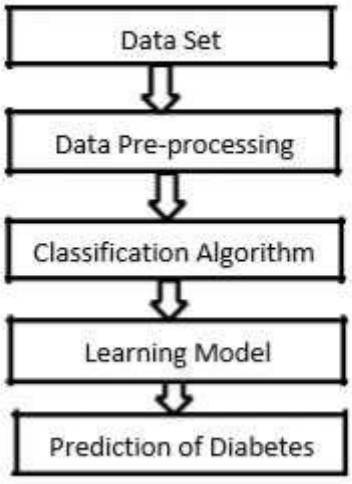

Figure 6.2: Flowchart of System

7. SYSTEM TESTING, RESULTS AND DISCUSSION

7.1System Testing

The goal of testing is to determine bugs that arise in the real time. Testing is the procedure of trying to discover every possibleerror or weakness in an operational invention. Itdelivers a method to patterned the functionality of mechanisms, sub-assemblies, musters ina completedartefact. This is the procedure of training software having the purpose of ensuring that the software system meets its necessities and prospects and does not fail in an intolerablemeans. Software testing is the procedure of exanimating whether the establishedorganisation is employedbestowing to the uniquepurposes and necessities [13]. This procedurebeginswhen the application is produced and the documents and connected data structures remainintended. Software testing staysimportantmeant formodifyingfaults.

7.1.1Unit Testing

Unit Testing is a software testing type that comprises individual testing unit of the application to check whether it works by itself and not dependent on any other units. Thegoal is to authenticate all the single units work properly.

System will get input from user through user interface. After getting input the system will process dataset, then the output of pre-process module will give as input for segmentation process, after segmentation the system will do feature extraction process, the output of feature extraction process is fed as input for classification and meta classifier where classification of dataset using Classification Algorithm will happen and finally accurate result will be given as output in the editable form through UI for user [14].

7.1.2Component Testing

Component testing is called as one of the type of software testing, where testing is conducted on every single component separately without the need of communicating with other components.Organisation will get input from user through user interface. Forecastingprecision is the keyassessmentconsideration in the present work. Precision is calculated by means of the subsequent equation.

Accuracy is called as the overall success rate of an algorithm [15]. Accuracy=(TP+TN)/(P+N) (1)

All the forecasted true positive and true negative divided by all positive and negative. True Positive(TP), True Negative(TN), False Negative(FN) and False positive (FP) are identified.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

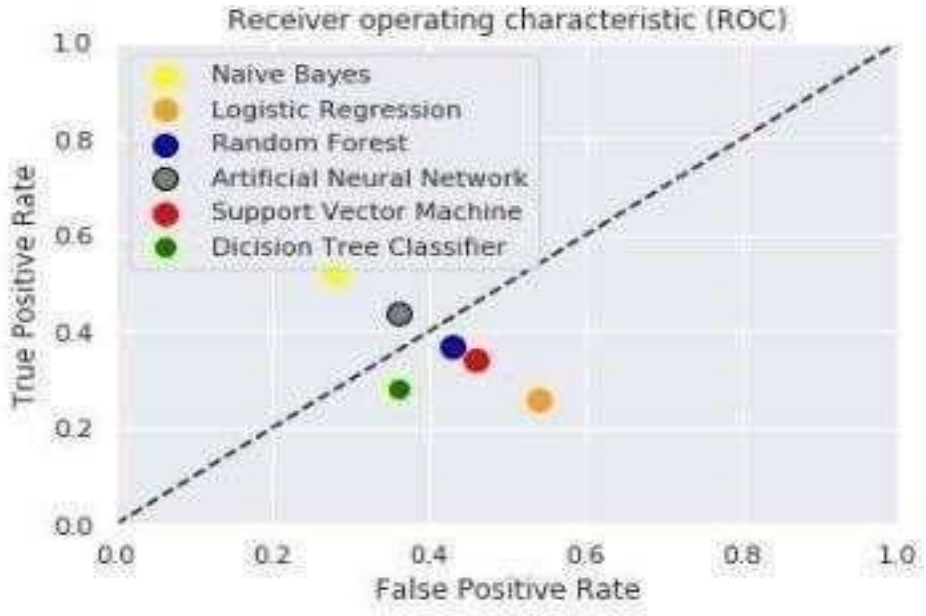

Figure 7.1.2: Comparison graph

7.1.3Integrated Testing

Integrated testing is well-defined as a software testing type, testing is conducted after the components are integrated with the other components.

In this testing we have combined the loss functions used during finding out the performance of each machine learning model by giving it variety of data sets with distinct ranging values. We have also tested the endpoints for the web app and written unit tests specific to the web app.

7.2Result Analysis

The aim of this project is to detect the calories consumed based on their food record. In early stage system was trained using random algorithms based on the dataset by various classification algorithms. Data set as partitioned into training and testing. Dataset consists of different samples which are selected randomly from the research analysis. Based on our personal information number of calories intake per day is recommended. We can add and delete the food that is already taken and interested to intake. This information calculates the total calories and represents the result in the form of pie chart. The most ideal algorithm is SVM and Random forest with Accuracy of 82% and 85% respectively.

Figure 7.2.1: Home Page

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

This is home page where the project application is started. This application is not user specific. Anybody can use this application and check the result. Since the data is not stored there is no need for the user to login.

Figure 7.2.2: Input Data

Since the system is trained with more number of dataset the accuracy will be more. In this page we are entering the information for further analysis. In this application the height and weight is entered in terms of centimetre and kilograms respectively. The procedure used to calculate the total calories recommended for a day depends on height and weight only.

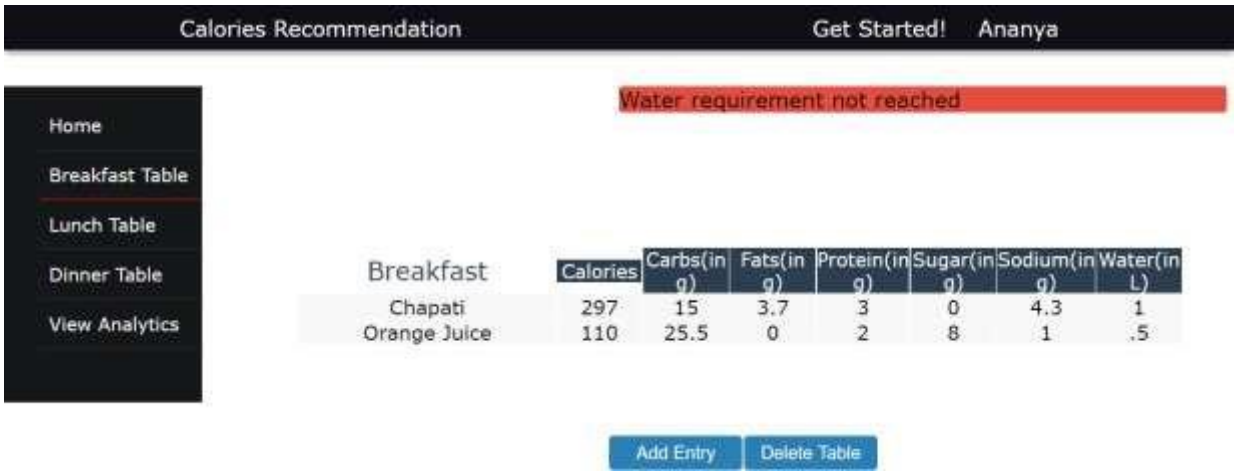

Figure 7.2.3: Menu Table

We need to create a table separately for breakfast lunch and dinner based on the recommended number of calories mentioned in view analytics. We need to satisfy all the attributes such as calories, carbs, fats, proteins, sugar, sodium and water individually. We need to enter the food and remaining attributes manually.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

Figure 7.2.4: Result

This is the result of above example. As we can see the recommended number of calories per day is 1750.52. We can add and delete food. And the result is displayed in the form of pie chart.

7.3Summary

Communication is the key factor of living in this world. Diabetes is an increasing disease,mainlybecause of the kind of nourishment we are having these days and the conflicting eating regimen and schedule that we take after. Diabetes is fundamentally caused because of obesity or high glucose level, and so forth. So our project aims on controlling diabetes in terms of food and also creates awareness about it. Likewise, we will determine whether the person intake calories are safe or not.

8. CONCLUSION

Machine learning has the potential to revolutionize the risk of diabetes by utilizing sophisticated processing techniques by providing vast volumes of risk datasets for diabetes epidemiologic and biological. In the early stages, diabetes detection is the secret to cure it. This study outlined a machine-learning method to detect levels of diabetes. The method may also aid researchers in developing a precise and reliable approach to help physicians make a smarter judgment regarding the status of the disease. This project focus on developing an application that detects diabetes in its early stage based on food consumption and prevents from being diabetic.

8.1 Scope for Future Work

The dataset analysed in this study was based on some main food. This study can be conducted with a larger dataset sample in rural and urban community settings in multiple states across India. This research study has only targeted on avoiding diabetes based on food and is applicable to all the users.

Various other key features in the medical records can also be analysed. It will be interesting to perform a more exhaustive exploration of additional features in the dataset and study their relevance.

Living with diabetes is challenging and distressful. Diabetic patient's condition cannot be understood only from consumption of food chart. There is a need to collect and analyse both subjective and objective patient information I order to fully understand. Subjective data can be captured by interviewing patients or by conducting surveys which will enrich the depth of patient information. The conversation between doctor and patient can also be collected and analyzed which could help to extract important feature

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9667 - 9676
Received 05 March 2021; Accepted 01 April 2021.

References

[1] Bădescu, S. V., et al. "The association between diabetes mellitus and depression." Journal of medicine and life 9.2 (2016): 120.

[2] Moradi-Lakeh, Maziar, et al. "Diabetes mellitus and chronic kidney disease in the Eastern Mediterranean region." (2017).

[3] Sarwar, Muhammad Azeem, et al. "Prediction of diabetes using machine learning Algorithms in healthcare." *2018 24th International Conference on Automation and Computing (ICAC)*. IEEE, 2018

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

[5] Joshi, Tejas N., and P. P. M. Chawan. "Diabetes prediction using machine learning techniques." *Ijera* 8.1 (2018): 9-13.

[6] Dutta, Debadri, Debpriyo Paul, and Parthajeet Ghosh. "Analysing feature importances for diabetes prediction using machine learning." 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2018.

[7] Ramesh, Sushant, et al. "Optimal predictive analytics of pima diabetics using deep learning." *International Journal of Database Theory and Application* 10.9 (2017): 47-62.

[8] Joshi, Rahul, and MinyechilAlehegn. "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach." *International Research Journal of Engineering and Technology* 4.10 (2017): 426-435.

[9] Iyer, Aiswarya, S. Jeyalatha, and RonakSumbaly. "Diagnosis of diabetes using classification mining techniques." *arXiv preprint arXiv:1502.03774* (2015).

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

[11] Zia, U. Ali, and Naeem Khan. "Predicting diabetes in medical datasets using machine learning techniques." International Journal of Scientific & Engineering Research Volume 8.5 (2017).

[12] Srikanth, Panigrahi, and DharmaiahDeverapalli. "A critical study of classification algorithms using diabetes diagnosis." 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE, 2016.

[13] E. Naresh and S. K. Kalaskar," A Novel Testing Methodology to Improve the quality of testing a GUI application," MSR journal of Engineering and technology research, 1(1), 2013, pp. 41-46

[14] E, Naresh, and B.P.VijayKumar, "Innovative Approaches in Pair Programming to Enhance the Quality of software Development," International Journal of Information Communication Technologies and Human Development,10(2),2018, pp.42-53.ion in fashion markets," Systems Science & Control Engineering, 3(1), 2017,pp. 154-161.

[15] Black, Cameron J., NicosMakris, and Ian D. Aiken. "Component testing, seismic evaluation and characterization of buckling-restrained braces." *Journal of Structural Engineering* 130.6 (2004): 880-894.