Prediction of Type 2 Diabetes mellitus using soft computing.

Poonam Punjabi
SMS Medical College, Jaipur, Rajasthan, India
Anuradha Yadav
SMS Medical College, Jaipur, Rajasthan, India
Manisha Sankhla
SMS Medical College, Jaipur, Rajasthan, India
Mamta (mamtakulhari213@gmail.com)
Mahatma Gandhi Medical College and Hospital, Jaipur, Rajasthan, India
Sandeep Mathur
SMS Medical College, Jaipur, Rajasthan, India
Harsh S Dave
SBKS Medical Institute & Research Centre, Vadodara, Gujarat, India
Vaishnavi Patel
University of Perpetual Help System Dalta, Las Pinas, Philippines
Tushar Chavhan
Indian Institute of Technology, Patna, India
Manisha
Banasthali Vidhyapith, Jaipur, Rajasthan, India
Amit Tak
RVRS Medical College, Bhilwara, Rajasthan, India

Research Article

Keywords: biceps skinfold thickness, diabetes mellitus, machine learning, prediction models, soft computing

Posted Date: May 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1639270/v1

Abstract

Background

Type 2 Diabetes Mellitus (DM) is another pandemic of 21 century, and its control is of immense importance. Researchers developed many predictor models using soft computing techniques. The present study developed a prediction model for Type 2 DM using machine learning classifiers. The analysis excludes plasma glucose concentration and insulin concentration as predictors to explore relationships with other predictors.

Methods

This cross-sectional study enrolled 108 participants aged 25 to 67 years from SMS Medical College, Jaipur (Rajasthan, India), after approval from the ethics committee. The study developed a prediction model using machine learning techniques. The classifiers used in the application include decision trees, support vector machines, K-nearest neighbors, and ensemble learning classifiers. A total of 25 predictors were collected and underwent feature reduction. The response levels include diabetes mellitus, prediabetes, and no diabetes mellitus. The models were run using three predictors and a response variable. The prediction model with the best accuracy and area under the receiver operator characteristic curve was selected.

Results

The features that vary among the three groups include age, WHR, biceps skinfold thickness, total lipids, phospholipids, triglycerides, total cholesterol, LDL, VLDL, and serum creatinine, and family history of DM. After feature reduction, the age, biceps skinfold thickness, and serum creatinine were run on the Classification learner application to predict the diabetic category. The best model was subspace discriminant with accuracy, sensitivity, specificity, and AUC under the ROC curve was 62.4%, 74%, 94%, and 0.70, respectively.

Conclusion

The present study concludes that age, biceps skinfold thickness, and serum creatinine combination have higher specificity in predicting type 2 DM. The study emphasized the selection of appropriate predictors along with newer machine learning algorithms.

Introduction

International Diabetes Federation reported that 537 million adults between 20 and 79 years living with diabetes mellitus globally, and the number may rise to 643 million by 2030 and 783 million by 2045. The

disease is responsible for 6.7 million deaths in 2021, leading to at least 966 billion dollars in health expenditure. (1) As per the Center for Disease Control and Prevention, 37.3 million people in the United States have diabetes constituting 11.3% US population, and 96 million people aged 18 years and above have prediabetes constituting 38% population. (2) Diabetes is a chronic disease, and its complications such as blindness, kidney failure, heart disease, diabetic foot (gangrene), and stroke are a severe threat to the health. COVID-19 cases suffering from Diabetes mellitus showed higher mortality rates that outnumbered deaths due to other cardiovascular diseases and cancer. (3) American Diabetes Association's Professional Practice Committee defined population health as "the health outcomes of a group of individuals, including the distribution of health outcomes." Clinical practice recommendations are tools that improve population health; however, personalized care results in optimal outcomes. (4) Therefore, forecasting diabetic cases at early stages has a vital role in preventing diabetes. Several prediction methods are available based on statistical or machine learning methods. (5)(6)(7)(8)

We propose a method for predicting and classifying diabetes mellitus using machine learning algorithms. In addition, readily available features, including demographic variables, anthropometric measurements, and information gathered from history taking, were used to learn the model.

Material And Methods

The participants of this hospital-based cross-sectional study came from SMS Medical College, Jaipur (Rajasthan, India), after approval from the Institutional Ethics Committee. A total of 108 subjects aged 25 to 67 years were enrolled. The study aimed to develop a prediction model to classify subjects into various diabetic categories using machine learning algorithms. Data on socio-demographic characteristics, information on physical examination, and laboratory test data were collected. The inclusion criterion includes patients above 25 years of age having type 2 DM, prediabetes, or found healthy. The patients suffering from chronic diseases were excluded. The response variable was the diabetic category. The relevant predictor variables for prediction were selected through feature reduction. The chosen features were used to train machine learning classifiers. The flowchart shows the overview of the protocol. (Fig. 1)

Features and response variable

Researchers collected 25 features, including age, anthropometric characteristics such as body mass index (BMI in kg/square meter), waist-hip ratio (WHR), biceps skinfold thickness, triceps skinfold thickness, supra-iliac skinfold thickness, and subscapular skinfold thickness. The lipid profile includes serum concentration of total lipids, phospholipids, low-density lipoproteins, high-density lipoproteins, low-density lipoproteins, total cholesterol, and triglycerides. The renal function tests include serum creatinine, urinary albumin concentration, urinary creatinine, and mean arterial pressure. The liver function test includes SGOT and SGPT. The qualitative features include gender, family history of diabetes, smoking habits, alcohol consumption, and tobacco use. As per the American Diabetes Association, based on glycated hemoglobin (HbA1c in %) the diabetic status (response variable) had three levels. The three

levels were – diabetes mellitus, having HbA1c $\geq 6.5\%$, prediabetes, having HbA1c between 5.7 and 6.5%, and no diabetes mellitus, with HbA1c $\leq$ 5.7%. (9)

Feature Reduction

The authors aimed to reduce the feature space to three based on the sample size. The general rule is to have at least ten subjects for each diabetic category. (10) The features that differ significantly among the three diabetic groups were used, otherwise rejected. Further, the selected features were run on machine learning algorithms individually, and the three features with the highest accuracy and AUC were selected.

Classification and Training

The three selected features were used for training various machine learning classifiers, including K-Nearest-Neighbours, Support Vector Machines, Decision Trees, and Ensemble-based Learning. The Classification learner application was used for training. (MATLAB 2019a)(11) Classifiers with the highest accuracy and performance metrics were chosen for prediction.

Statistical Analysis

The descriptive statistics for quantitative data were expressed in terms of means and standard deviation. The qualitative data were expressed in proportions. After an appropriate assumption check, quantitative features were compared among the three groups using One-way ANOVA or Kruskal Wallis test. The independence of qualitative features and the three groups were tested with the chi-squared test. The models' performance was assessed using accuracy, area (AUC) under receiver operator characteristic (ROC) curve, sensitivity, and specificity. The significance level was considered at 5%. The JASP version 0.16.1.0 software was used for statistical analysis. (12)

Results

The continuous variables were tested for normality. The features that vary among three groups include age [W = 18.027; $p < 0.001$], WHR [W = 18.495; $p < 0.001$], biceps skin fold thickness [W = 7.233; $p = 0.027$], total lipids [W = 12.058; $p < 0.002$], phospholipids [W = 14.308; $p < 0.001$], triglycerides [W = 16.564; $p < 0.001$], total cholesterol [W = 9.381; $p < 0.009$], LDL [W = 12.503; $p < 0.002$], VLDL [W = 18.547; $p < 0.001$] and serum creatinine [W = 7.599; $p < 0.001$] (Table 1). The total lipid was chosen as a feature due to its high correlation with other parameters of lipid profile. (Figure 2 and Figure 3) Among the categorical variables, the history of diabetes mellitus showed relationship with the three diabetic groups. (Table 2 and Figure 4) Subsequently, the features including age, WHR, biceps skin fold thickness, total lipids, serum creatinine and history of diabetes mellitus were ran on Classification learner app individually as predictors. The three features with highest accuracy and AUC were age, serum creatinine and biceps skin fold thickness. (Table 3) Finally, the Classification learner app was used to train models using the three chosen features. The best model was subspace discriminant with accuracy, sensitivity, specificity and AUC under the ROC curve was 62.4%, 74%, 94% and 0.70 respectively.(Table 4 and Figure 5)

Discussion

Diabetes mellitus (DM) is a set of disorders with hyperglycemia in common. Based on pathogenesis, DM is further classified. Type 1 DM is characterized by insulin deficiency. In contrast, type 2 DM is a heterogeneous collection of disorders characterized by variable degrees of insulin resistance, impaired insulin secretion, and excessive hepatic glucose production. Type 2 DM is an ongoing pandemic and is among the critical diseases. (13) Many risk factors are associated with diabetes, including age, obesity, lack of physical activity, family history of diabetes, fat-rich diet, and high blood pressure. Every three years, screening is recommended for individuals over 45 years and younger people having risk factors or whose body mass index is 25 kg/m^2. (14) The role of insulin in glucose homeostasis has been well established. (15) Several studies used plasma glucose concentration and insulin concentration as predictors, which are used to define diabetes. (16)The present study classifies the diabetic category based on glycated hemoglobin and used other demographic, clinical, and laboratory parameters as predictors but excluded plasma glucose and insulin concentration. In contrast to binary classification in most studies, the present study divided subjects into three response categories – diabetic, prediabetic, and non-diabetic. The subspace discriminant algorithm best-classified diabetics with specificity and AUC of 94% and 0.70, respectively.

Zheng et al. extracted features of 300 patients from Electronic Health Repository ranging from 2012 to 2014. They applied machine learning models, including k-Nearest-Neighbors, Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression, to predict type 2 diabetes mellitus. The learning methods achieved high identification performances (~0.98 in average AUC) compared to the state-of-the-art algorithm (0.71 in AUC). (17) Maniruzzaman et al. conducted a study on 768 subjects (268 diabetic and 500 controls) to classify them into the diabetic and non-diabetic categories. Researchers quoted that due to non-linearity, non-normality, and inherent correlation structure in the medical data, the Gaussian process (GP)-based classification technique uses linear, polynomial, and radial basis kernel. This model's accuracy, sensitivity, and specificity were 81.97%, 91.79%, and 63.33%, respectively. Compared to naïve Bayes, linear and quadratic discriminant analysis models, the GP-based model showed better performance. (18) Although the accuracy and sensitivity of the GP-based model are higher, the study has lower specificity. In a study using the Pima Indian dataset from the UCI repository involving 768 females at least 21 years of age, Mercaldo et al. train classifiers with eight feature vectors, including the number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skinfold thickness (mm), 2-Hour serum insulin (mu U/ml), body mass index (in kg per sqm), diabetes pedigree function, and age (in years). Six machine learning classification algorithms were used, including J48, Multilayer Perceptron (a deep learning algorithm), Hoeffding Tree, JRip, BayesNet, and Random Forest. They evaluated performance metrics using Precision, Recall, F-measure, and ROC Area. They found the best precision value of 0.770 and a recall equal to 0.775 using the Hoeffding Tree algorithm to predict the onset of type 2 diabetes mellitus within five years in Pima Indian women. (19) Zhang et al. tested the ability of machine learning algorithms to predict the risk of type 2 diabetes mellitus (T2DM) in a rural Chinese population. The

authors focused on 36,652 eligible participants from the Henan Rural Cohort Study. Six machine learning classifiers were used, including logistic regression, classification and regression tree, artificial neural networks, support vector machine, random forest, and gradient boosting machine. Among the top-10 variables across all methods were a sweet flavor, urine glucose, age, heart rate, creatinine, waist circumference, uric acid, pulse pressure, insulin, and hypertension. The study includes new important risk factors such as urinary indicators and sweet flavor. The GBM model performed best with an AUC of 0.872 and 0.817 with and without laboratory data. (20) Using the National Health and Nutrition Examination Survey (NHANES) dataset, machine learning models, including logistic regression, support vector machines, random forest, and gradient boosting were evaluated on their classification performance. The models were then combined to develop a weighted ensemble model capable of leveraging the performance of the disparate models to improve detection accuracy. The information gained from tree-based models was used to identify the key variables within the patient data that contributed to the detection of at-risk patients in each disease class by the data-learned models. In diabetes classification (based on 123 variables), the eXtreme Gradient Boost (XGBoost) model achieved an area under ROC (AU-ROC) score of 86.2% (without laboratory data) and 95.7% (with laboratory data). The ensemble model had a top AU-ROC score of 73.7% (without laboratory data) for prediabetic patients, and XGBoost performed the best at 84.4% for laboratory-based data. The top five predictors in diabetes patients were waist size, age, self-reported weight, leg length, and sodium intake. (21) Kandhasamy et al. compared the performance of various machine learning algorithms, including J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines. Researchers concluded that the J48 decision tree classifier achieved higher accuracy of 73.82% than other classifiers. (22) Abbas et al. used the San Antonio Heart Study data to develop a type-2 diabetes prediction model using support vector machines with 10-fold cross-validation. The results showed 84.1% accuracy with a recall rate of 81.1% averaged over 100 iterations. (23)

Conclusion

The development of predictor models for type 2 diabetes mellitus has a crucial role in preventing and controlling the disease. However, selecting relevant predictors is as essential as selecting the best machine learning classifier. The present study concludes that age, biceps skinfold thickness, and serum creatinine combination have higher specificity in predicting type 2 DM.

Limitations of the study: Researchers chose three predictors to train the classifiers based on the sample size in each category. However, a study with larger sample size is required to increase performance metrics in the future.

References

1. IDF Diabetes Atlas | Tenth Edition [Internet]. [cited 2022 Apr 22]. Available from: https://diabetesatlas.org/

2. National Diabetes Statistics Report | Diabetes | CDC [Internet]. [cited 2022 Apr 22]. Available from: https://www.cdc.gov/diabetes/data/statistics-report/index.html
3. Bhandari S, Shaktawat AS, Tak A, Shukla J, Gupta J, Patel B, et al. Evaluating interactions between hyperglycemia and clotting factors in patients suffering with SARS-CoV-2 infection. Clin Diabetol [Internet]. 2021 [cited 2022 Apr 22];10(1):114–22. Available from: https://journals.viamedica.pl/clinical_diabetology/article/view/DK.a2021.0022
4. Diabetes Mellitus | Harrison's Manual of Medicine, 19e | AccessMedicine | McGraw Hill Medical [Internet]. [cited 2022 Apr 22]. Available from: https://accessmedicine.mhmedical.com/content.aspx?bookid=1820&sectionid=127559730
5. Pangaribuan JJ, Suharjito. Diagnosis of diabetes mellitus using extreme learning machine. In: 2014 International Conference on Information Technology Systems and Innovation, ICITSI 2014 - Proceedings. 2014.
6. Bhandari S, Shaktawat AS, Tak A, Patel B, Shukla J, Singhal S, et al. Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. Ibnosina J Med Biomed Sci [Internet]. 2020 [cited 2022 Apr 13];12(2):123. Available from: http://www.ijmbs.org/article.asp?issn=1947-489X;year=2020;volume=12;issue=2;spage=123;epage=129;aulast=Bhandari
7. Bhandari S, Singh Shaktawat A, Tak A, Patel B, Gupta J, Gupta K, et al. Independent Role of CT Chest Scan in COVID-19 Prognosis: Evidence From the Machine Learning Classification (1) (2) (3) (4) (5) (6) (7). Scr Med. 2021;52(4):273–81.
8. Tak A, Dia S, Dia M, Wehner TC. Indian COVID-19 Dynamics: Prediction Using Autoregressive Integrated Moving Average Modelling ARTICLE INFO (1) (2). Scr Med [Internet]. 2021 [cited 2022 Apr 13];52(1):6–14. Available from: https://github.com/CSSEGISand-
9. Association AD. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care [Internet]. 2010 Jan [cited 2022 Apr 23];33(Suppl 1):S62. Available from: /pmc/articles/PMC2797383/
10. Abhaya, Indrayan Malhotra R. Medical Biostatistics. fourth. Florida, USA: CRC Press, Taylor & Francis Group,; 2018. 472–73 p.
11. MATLAB - MathWorks - MATLAB & Simulink [Internet]. [cited 2022 Apr 3]. Available from: https://in.mathworks.com/products/matlab.html
12. JASP Team. JASP (Version 0.16.1)[Computer software] [Internet]. 2022. Available from: https://jasp-stats.org/
13. Ginter E, Simko V. Type 2 diabetes mellitus, pandemic in 21st century. Adv Exp Med Biol [Internet]. 2012 Aug 1 [cited 2022 Apr 24];771:42–50. Available from: https://pubmed.ncbi.nlm.nih.gov/23393670/
14. Diabetes Mellitus | Harrison's Manual of Medicine, 19e | AccessMedicine | McGraw Hill Medical [Internet]. [cited 2022 Apr 24]. Available from: https://accessmedicine.mhmedical.com/content.aspx?bookid=1820&sectionid=127559730

15. Guyton JR, Foster RO, Soeldner JS, Tan MH, Kahn CB, Koncz L, et al. A model of glucose-insulin homeostasis in man that incorporates the heterogeneous fast pool theory of pancreatic insulin release. Diabetes [Internet]. 1978 [cited 2022 Apr 24];27(10):1027–42. Available from: https://pubmed.ncbi.nlm.nih.gov/700259/
16. Mujumdar A, Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. In: Procedia Computer Science. 2019.
17. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;
18. Maniruzzaman M, Kumar N, Menhazul Abedin M, Shaykhul Islam M, Suri HS, El-Baz AS, et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput Methods Programs Biomed. 2017;
19. Mercaldo F, Nardone V, Santone A. Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Comput Sci. 2017;
20. Zhang L, Wang Y, Niu M, Wang C, Wang Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. Sci Rep. 2020;
21. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019;
22. Kandhasamy JP, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. In: Procedia Computer Science. 2015.
23. Abbas H, Alic L, Rios M, Abdul-Ghani M, Qaraqe K. Predicting diabetes in healthy population through machine learning. In: Proceedings - IEEE Symposium on Computer-Based Medical Systems. 2019.

Tables

Table 1. Comparison of quantitative features among the three diabetic groups using One-way ANOVA or nonparametric Kruskal Wallis test based on the distribution in the three groups.

Parameters	Distribution	Test statistic	*p*
Age	Non-normal	18.027	< .001
WHR	Non-normal	18.495	< .001
Biceps skinfold thickness	Non-normal	7.233	0.027
Triceps skinfold thickness	Non-normal	4.297	0.117
Subscapular skinfold thickness	Non-normal	1.814	skinfold
Total lipids	Non-normal	12.058	0.002
Phospholipids	Non-normal	14.308	< .001
Triglycerides	Non-normal	16.564	< .001
Total Cholesterol	Non-normal	9.381	0.009
LDL	Non-normal	12.503	0.002
VLDL	Non-normal	18.547	< .001
MAP	Non-normal	4.784	0.091
SGOT	Non-normal	1.766	0.413
SGPT	Non-normal	5.448	0.066
BMI	Normal	2.623	0.077
Supra-iliac skinfold thickness	Normal	0.19	0.827
HDL	Normal	2.8	0.065
S Creatinine	Normal	7.599	< .001

Table 2. shows the association between categorical features and the three diabetic groups using the c^2 test.

Contingency table	c^2 value	Degree of freedom	*p*
Gender Diabetic Groups	1.534	2	0.464
Tobacco Diabetic Groups	3.795	2	0.15
Smoking Diabetic Groups	3.209	2	0.201
Alcohol Diabetic Groups	3.26	2	0.196
Urinary albumin Diabetic Groups	10.213	6	0.116
Urinary creatinine Diabetic Groups	5.187	10	0.878
Family history of DM Diabetic Groups	10.312	2	0.006

Table 3. shows the performance metrics of machine learning classifiers trained with individual predictors in the prediction of DM class.

Variable	Classifier	AUC	Accuracy
Age	Coarse Gaussian SVM	0.66	52.8%
WHR	Subspace Discriminant	0.62	50.9%
Biceps	Medium Gaussian SVM	0.64	50%
History of DM	Boosted Trees	0.64	49.1%
Total lipids	Coarse Tree	0.59	58.8%
S. Creatinine	Naïve Bayes	0.65	57.1%

Table 4 Shows performance metrics of the trained classifier ‘Subspace Discriminant’ using age, serum creatinine, and biceps skinfold thickness as predictors in the prediction of DM class.

Performance metrics	Value
Accuracy	62.4%
AUC	.70
Sensitivity	74%
Specificity	94%

Declarations

Competing Interest – There is no competing interest among authors.

Figures

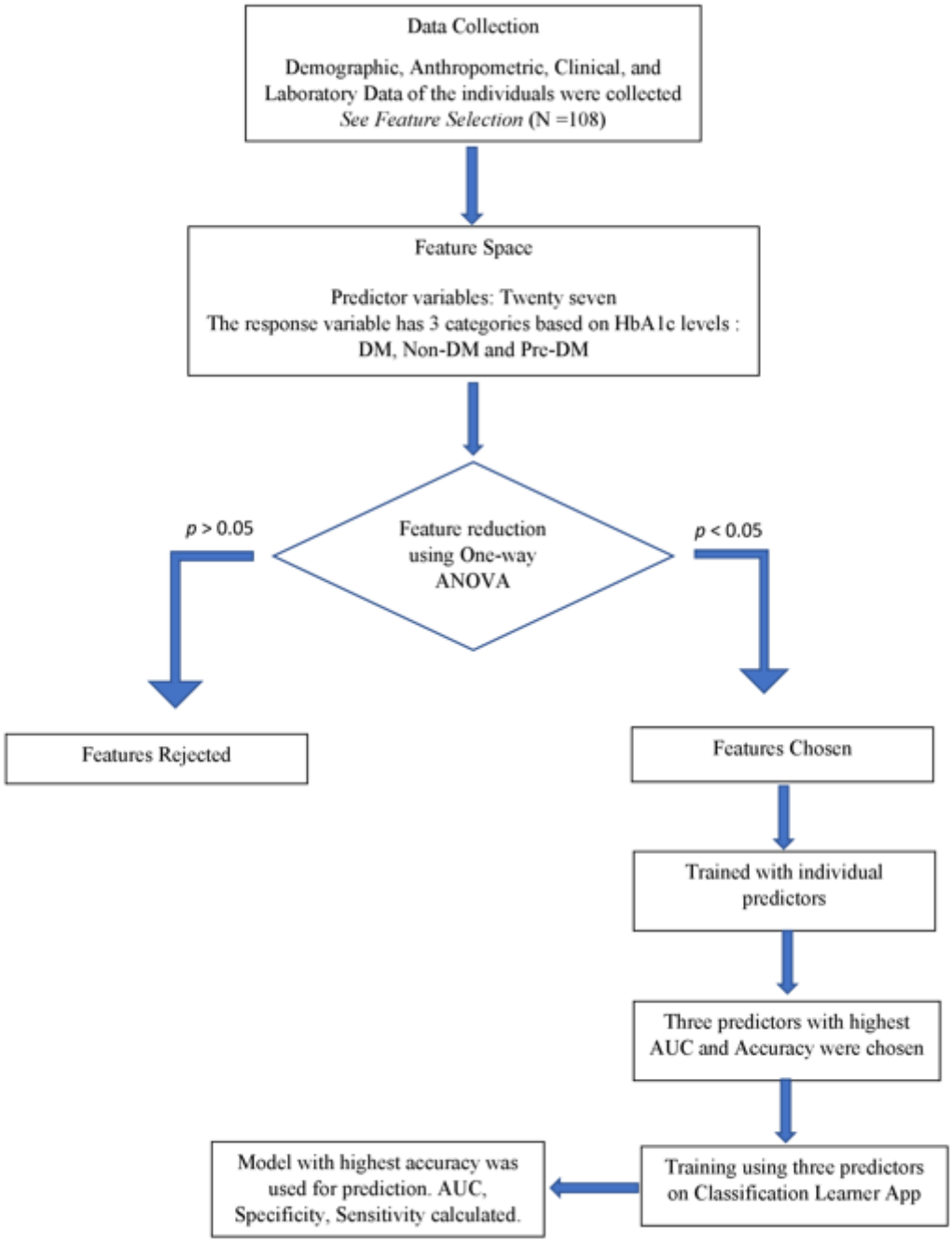

Figure 1

Shows classification of diabetic groups into diabetes, prediabetes, and no diabetes, based on glycated hemoglobin (%)

Figure 2

Shows heatmap depicting high correlation among various parameters of lipid profile.

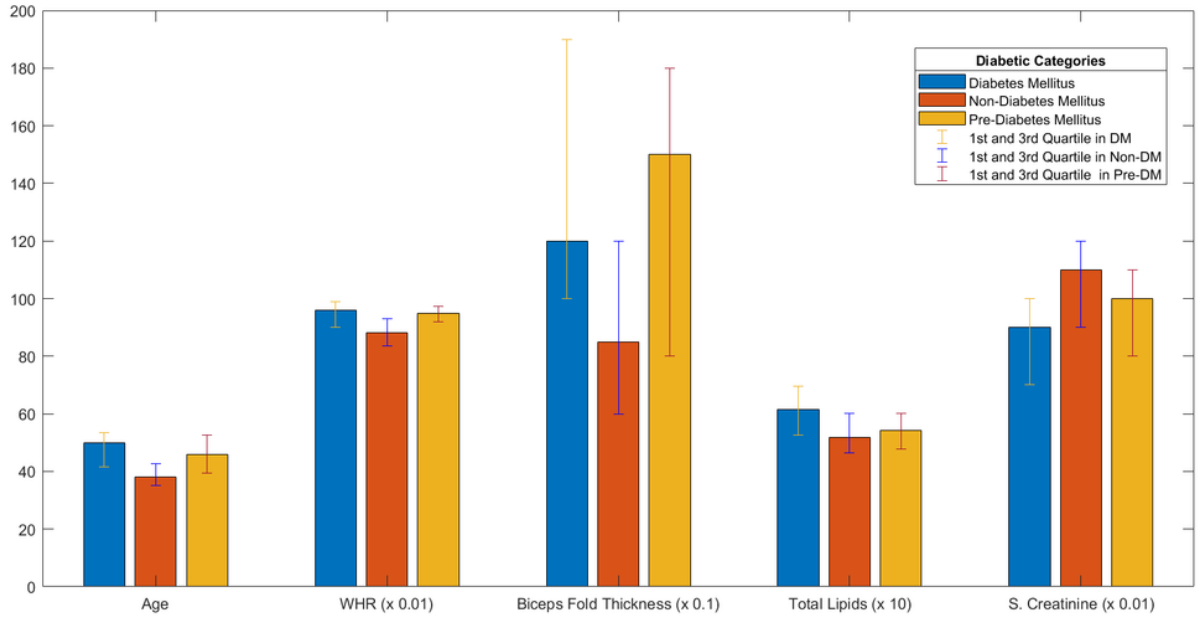

Figure 3

Shows distribution of age, WHR, biceps skinfold thickness, total lipids, and serum creatinine across the three diabetic groups.

Figure 4

Shows distribution of family history of diabetes across the three diabetic groups.

Figure 5

Shows area under the receiver operator characteristic curve for the subspace discriminant classifier when age, biceps skinfold thickness, and serum creatinine were used to predict DM class.