Predicting Future Diabetes Using Machine Learning Models in a High-Risk Pediatric Cohort

Madhumita Sinha (Madhumita.Sinha@nih.gov)
National Institute of Diabetes and Digestive and Kidney Diseases
Elsa Vazquez Arreola
National Institute of Diabetes and Digestive and Kidney Diseases

Article

Keywords:

Posted Date: August 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1891969/v1

Abstract

The acute rise in childhood obesity has been associated with an increased risk of metabolic dysfunction and early-onset diabetes. Metabolic syndrome (MetS), a cluster of risk factors predicting future cardiometabolic disease, is not well established in children. This study leverages data from a longitudinal observational study of diabetes (1965–2007) that enrolled pediatric and adult participants, to create machine learning-based models (ML) to predict future incident diabetes. Anthropometric, clinical, and laboratory data from baseline examinations of 2,049 nondiabetic children were obtained, and 17-metabolic risk predictor variables were included. Multiple classification schemes were explored, and an ensemble model was created by combining the 5-best algorithms with the highest predictive accuracy. The area under the receiver-operating-characteristic curve (AUROC) of the ensemble model was significantly higher than the logistic regression model in predicting future diabetes [0.88 (95% CI: 0.85–0.9) and 0.63 (95% CI: 0.57–0.69), ($p < 0.001$)]. The predictor variable importance of the best performing Random Forest model identified additional risk predictors outside of the pediatric MetS components as recommended by the International Diabetes Federation. This study used consistent and well-structured clinical data to create ML-based models that predicted future incident diabetes in a pediatric cohort with a high degree of accuracy.

Introduction

Childhood obesity has been rising steadily in the United States (US) over the past several decades. According to the most recent data from the National Health and Nutrition Examination Survey (NHANES), the estimated prevalence of obesity is 19.3% among children between 2–19 years of age. [1] The increasing prevalence of obesity has led to a rise in associated comorbidities such as dysglycemia, dyslipidemia, and hypertension, that is often manifested at an early age. [2–4] Recent studies have also reported a contemporaneous increase in both the incidence and prevalence of youth onset type 2 diabetes (T2D) in the US, that is particularly evident among minority populations. [5, 6] Hence identifying metabolic risk factors in childhood is important for prevention of future diabetes.

Metabolic syndrome (MetS) that evolved from Reaven's "syndrome X", [7] comprises a clustering of interrelated clinical risk factors centered around adiposity and insulin resistance, a common etiological pathway for cardiometabolic disease. In adults, MetS as described by the National Cholesterol Education Program (NCEP) Adult Treatment Panel III (ATP III), has been shown to be a good predictor of future cardiometabolic risk. [8, 9] However, in children, a diagnosis of MetS and its long-term association with development of chronic conditions such as diabetes is not well established. In 2007, the International Diabetes Federation (IDF) announced a consensus definition for MetS in children and adolescents. [10] These clinical risk factors include abdominal obesity, impaired fasting glucose (IFG), elevated triglycerides, low high-density lipoprotein cholesterol (HDL-C), and hypertension. In children 10 years of age or older, a MetS diagnosis can be made with central obesity and presence of two or more additional clinical features as mentioned previously.[10] Despite these recommendations, there is hesitancy among

pediatric health care providers to use this definition in the absence of long-term scientific evidence of its predictive accuracy. [11] Instead, the American Academy of Pediatrics emphasizes focus on screening for cardiometabolic risk in children and adolescents based on specific metabolic risk clustering that is obesity driven, rather than defining a syndrome that relies on cut-points and risk measures that are not evaluated in a continuum. [11] In addition, the instability of the MetS diagnosis itself as children transition through the different life-course stages such as childhood, adolescence, and into adulthood has been a cause for concern. [11, 12] In a prior study examining MetS components, we found that only body mass index (BMI) and impaired glucose tolerance were predictors of future diabetes, whereas the other components were not. [13]

The role of machine learning (ML) in medicine is evolving fast because of its ability to analyze highly complex, and nonlinear relationships in large medical data sets to improve prognostic and diagnostic accuracy of disease conditions. [14] The dilemma of identifying metabolic risk measures during childhood that would have future prognostic significance in diabetes prediction forms the basis of our current study. Our objective is to create ML based predictive models using multiple metabolic risk variables that are components of IDF's MetS and to explore additional risk measures obtained during childhood from a longitudinal observational study to predict future incident diabetes. We also compared the predictive performance of an ensemble ML model, a combination of the best performing ML models, with a conventional binomial logistic regression (LR) model.

Methods

Data Source:

Data was obtained from a longitudinal observational study of diabetes and its related conditions that was conducted in an American Indian (AI) community in the southwestern United States (US) over a 43-year study period between 1965–2007 by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), Phoenix Branch. Children 5-years of age or older, and adults were invited to participate in comprehensive research examinations biennially during the study. Data collected in this study included anthropometric measurements, clinical data, and biochemical tests. The current analysis included children and adolescents, who had their first non-diabetic exam with a complete set of relevant clinical and biochemical metabolic risk measures between 5 and < 21 years of age, and at least one follow-up exam before their 55th birthday. There were 3,415 children and adolescents who had data on all baseline parameters. Of them, 75 patients were diagnosed with diabetes prior to or at their baseline research exam and were excluded. Another 1,291 patients did not have a follow up research examination and were not eligible for inclusion since their future diabetes status could not be determined. This yielded a final dataset with 2,049 unique pediatric participants. Among these, data on maternal diabetes status was available in 1,965 patients and fasting insulin levels were available in 1,978 patients.

The primary outcome variable was a diagnosis of diabetes. In this AI population, diabetes is overwhelmingly type 2 diabetes (T2D) and mutations in the Maturity-Onset Diabetes of the Young (MODY) genes do not constitute a significant cause for diabetes in the youth. [15, 16] Definition of diabetes in this study was based on the American Diabetes Association (ADA) standard criteria: FPG ≥ 126 mg/dL (7.0 mmol/l), 2hPG ≥ 200 mg/dL (11.1 mmol/l), HbA_{1c} ≥ 6.5% (48 mmol/mol) or a previous clinical diagnosis.[16] The follow-up in years for diabetes was based on whether a child developed diabetes before 55 years of age. If the child did not develop diabetes before age 55, the follow-up time in years was calculated from the baseline measurement to their last research examination before age 55 years. All laboratory testing in this study was performed at NIDDK Phoenix's Clinical Laboratory Improvement Amendments (CLIA) certified laboratory during the entire study period.

The Institutional Review Board of the National Institutes of Health (NIH) approved the study (Protocol ID: OH76DK0256). Written informed consent was obtained from parents at study initiation and assent was obtained from the children. The study was performed in accordance with relevant guidelines and regulations put forth by the NIH.

Predictors and preprocessing of data:

The metabolic risk predictors were selected based on the IDF's consensus on MetS diagnosis in children and adolescents published in 2007 [10] that included age-sex-adjusted waist circumference percentile, blood pressure (BP), fasting plasma glucose, serum triglycerides, and high-density lipoprotein (HDL) cholesterol. Additional diabetes risk predictor variables were explored in this large pediatric cohort for developing the ML analytical models. These included: *Demographic information*: age; sex; and history of maternal diabetes; *Anthropometric measurements*: height, weight, and body mass index (BMI); *Biochemical tests*: 2-hour plasma glucose following an oral glucose tolerance test with a 75-g oral glucose load (2-hr OGTT); glycated hemoglobin (HbA1c), and serum total cholesterol. Of the 17 predictor variables, only 2 had missing values, fasting insulin in 3.4% and history of maternal diabetes in 4.0% cases. For fasting insulin, missing values were imputed as the means of the non-missing values. Age-sex-adjusted BMI z-scores were computed using the computer program and 2000 CDC growth charts for children between 24 and 239 months of age (cdc.gov/growthcharts/computer_programs.htm accessed October 4, 2016). The "modified z-score" is similar to the usual z-score method (distance from the mean in standard deviation units), however unlike the commonly used unmodified z-score provided by the CDC program, it does not compress the frequency distribution of high z-scores such that very few have values > 3.[17] Diabetes incidence per thousand person-years was calculated using the number of incident cases of diabetes and person-years of follow-up through age 55 years.

Classification modeling:

Classification schemes are commonly used in data mining with the objective of creating a model that predicts the value of a target (or dependent variable) based on the values of several inputs or independent variables. [18, 19] The 17 known metabolic risk variables obtained from the baseline non-diabetic exam in 2,049 children included in the study cohort were used to build ML models. The dataset

was randomly split into training (70%) and testing sets (30%). Based on a review of relevant published information, several classification algorithms were initially applied to the dataset to determine the best classifier to build a suitable predictive model and predict future diabetes. The following classifying schemes were initially evaluated: neural network (NN), k-nearest neighbor (KNN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector machine (SVM), Linear Support Vector machine (LSVM), and Chi-square Automated Interaction Detection (CHAID).

In predictive data mining processes, it has been shown that combining the predictions from multiple independent algorithms often yields more accurate predictions than can be derived from any one method by reducing the generalization error.[20] It is particularly useful when the types of algorithms included in the project are different from each other, such as the ones used in this project. After building and evaluating individual classification algorithms as mentioned previously, the results were combined in an ensemble by stacking the predictions made by the 5 best algorithms selected based on the accuracy with a minimum threshold of AUC of > 0.8 and using the confidence weighting voting method. To avoid overfitting and decreasing variance errors, bootstrapping was used in the training dataset. For comparison of predictive accuracy of traditional regression models to ML based models, a forward stepwise logistic regression model was built to predict incident diabetes using the same 17 metabolic risk variables.

Predictor Importance:

Predictor importance of the RF and LR classification models were computed by calculating the reduction in variance of the target variable attributable to each predictor, via a sensitivity analysis. The software automatically calculates this sensitivity index and generates a graph of the top 10 predictors in terms of decreasing sensitivity index (importance). [21] Predictor importance based on the logistic regression analysis using the Wald statistics were also computed.

Sample size analysis:

Given the large number of observations (> 2,000) available for training and testing datasets, and considering the number of input variables, the sample size was adequate for representative model building. It has been shown that the predictive accuracy for binary outcome variables is higher using modern ML based modelling techniques and approximately 20–50 observations per variable was required to obtain a stable AUC with LR and other ML-based models. [22] Only 8.7% of patients in this cohort developed diabetes during the follow up period, and this imbalance between the classes can affect the performance of some of the classification algorithms. To optimize model performance, the training dataset was balanced by generating new samples in the under-represented minority class using the built-in random minority oversampling feature of the modeling software in a 10:1 ratio. [23]

Statistical analysis:

Demographic and clinical characteristics were summarized as means and standard deviations (SD) or medians and interquartile ranges [IQR: Q3-Q1)] for non-Gaussian variables. We compared categorical

variables using chi-squared tests (or Fisher's exact tests for cell counts < 5). Continuous variables were compared using the unpaired Student's *t* test, and medians using the Mann–Whitney U test. Statistical significance was assessed at $p < 0.05$.

Data mining and machine learning models were built and evaluated by using the IBM SPSS Modeler, version 18.2 (IBM Corp., Armonk, N.Y., USA).

Results

A total of 2,049 children and adolescents had a baseline non-diabetic exam between June 1993 and March 2006 with at least one follow-up exam prior to age 55 years. Mean age of the study cohort was 12.0 years (SD 3.8), and 54.5% were females, all participants were of American Indian race. Over a median follow-up period of 6.3 years (interquartile range [IQR]: 3.6–9.6), there were 178 incident cases of diabetes (12.9 cases/1,000 person years). The baseline metabolic risk input variables of the whole cohort and by the primary outcome of diabetes diagnosis is shown in Table 1.

Table 1
Input metabolic risk predictor variables in child at their baseline (first non-diabetic exam)

METABOLIC RISK VARIABLES (INPUT)				
Child metabolic characteristics	**Total N = 2,049**	**Diabetic at Follow up N = 178**	**Non-diabetic at Follow up N = 1,871**	**P value**
Age, years	12.4 (3.8)	14.1 (4.1)	12.2 (3.8)	< 0.001*
Sex (n, %)	933 (45.5)	70 (39.3)	863 (46.1)	0.08
• Male	1116 (54.5)	108 (60.7)	1008 (53.9)	
• Female				
Weight, kg	60.6 (26.3)	80.3 (27.1)	58.7 (25.4)	< 0.001*
Height, cms	150.6 (16.3)	158.0 (14.2)	150.0 (16.3)	< 0.001*
BMI modified z-scores	1.5 (1.4)	2.5 (1.4)	1.4 (1.4)	< 0.001*
Waist circumference, percentile	87.8 (60.0, 96.1)	95.9 (88.9, 98.2)	86.5 (57.6, 95.6)	< 0.001*
Cholesterol, mg/dl	155.4 (28.8)	162.3 (31.0)	154.7 (28.5)	0.002*
Triglycerides, mg/dL	90.1 (53.9)	123.2 (59.6)	87.0 (52.3)	< 0.001*
HDL-C, mg/dL	44.5 (11.3)	38.7 (8.7)	45.0 (11.3)	< 0.001*
Systolic BP, mm Hg	107.4 (13.8)	113 (13.7)	106.9 (13.7)	< 0.001*
Diastolic BP, mm Hg	59.5 (10.2)	63.4 (10.4)	59.1 (10.1)	< 0.001*
Fasting glucose, mg/dL	87.8 (7.3)	92.7 (9.1)	87.4 (6.9)	< 0.001*
HbA1c, %	5.1 (0.4)	5.4 (0.5)	5.1 (0.4)	< 0.001*
2-hour glucose, mg/dL	101.5 (23.8)	122.5 (30.6)	99.5 (22.1)	< 0.001*

Data are shown as n (%), median (quartile 1, quartile 3) or mean ± SD, *P-value < 0.05 considered significant

METABOLIC RISK VARIABLES (INPUT)				
Fasting insulin, µU/mL	12.2 (6.0, 21.9)	25.0 (14.7, 40.0)	11.8 (6.0, 19.9)	< 0.001*
Albumin creatinine ratio	11.9 (7.0, 21.7)	9.3 (6.5, 15.5)	12.4 (7.1, 21.9)	< 0.001*
	N = 1,965	**N = 174**	**N = 1,791**	
History of maternal diabetes (n, %) • Yes • No	956 (48.7) 1009 (51.3)	128 (73.6) 46 (26.4)	828 (46.3) 963 (53.7)	< 0.001*
Data are shown as n (%), median (quartile 1, quartile 3) or mean ± SD, *P-value < 0.05 considered significant				

There were significant differences in all metabolic risk input variables except sex between children who developed diabetes at a follow-up exam and those who did not.

Predicting Future Diabetes

For prediction of the primary outcome of future diabetes in this pediatric cohort the four machine learning models that significantly outperformed the binomial LR model were RF, CHAID, NN and SVM. Table 2 outlines the predictive performance of the four selected classification algorithms showing their area under the receiver operating characteristic curve (AUROC) and 95% Confidence Intervals (CI). The Random Forest model had the best performance with an AUC of 0.92 in the testing set.

Table 2
Performance of best individual ML algorithms

MODELS	Area Under the Curve	Standard Error	95% Confidence Intervals
Random Forest	0.919	0.021	0.878–0.959
Neural Network	0.817	0.019	0.779–0.855
CHAID	0.825	0.017	0.791–0.859
SVM	0.785	0.022	0.742–0.827

Figure 1. shows the overall performance in predicting diabetes by combining the results of the classification algorithms in an ensemble model in the testing set and compares it to the binomial LR-model.

The AUC of the ensemble model and logistic regression model in predicting diabetes were 0.88 (95% CI: 0.85–0.9) and 0.63 (95% CI: 0.57–0.69). The ensemble model outperformed the LR model significantly

($p < 0.001$) in predicting incident diabetes (Table 3). In comparison to the LR model, the ensemble model correctly identified diabetes in 590 out of 639 cases compared to 459 correctly identified by the LR model.

Table 3
Comparison of predictor performance between the Machine Learning ensemble model and the logistic regression model.

	AUC CI*	Accuracy %	Sensitivity %, CI	Specificity %, CI	Positive Predictive Value %, CI	Negative Predictive Value %, CI
Ensemble Model	0.88 (0.85–0.91)	92.6	58.0 (43.2–71.8)	95.9 (94.0-97.4)	57.5 (46.1–68.1)	96.0 (94.6–97.1)
Logistic Regression Model	0.63 (0.57–0.69)	73.6	83.3 (69.7–92.5)	72.8 (68.9–76.4)	22.5 (19.5–25.9)	97.0 (96.1–98.9)

*CI: Confidence Interval

Predictor importance

The RF predictor variable importance not only captures the impact of each predictor individually but also interactions with other predictor variables. Figure 2 shows the feature importance in the best performing RF model and shows the variables that are most important in predicting diabetes in this study cohort. The 5 most important features were: 2-hour OGTT, fasting insulin, HbA1c, waist circumference percentile (indicating central adiposity), and BMI z-score (indicating general adiposity). In the LR-model the top 5 predictors included 2-hour OGTT, maternal diabetes history, waist circumference percentile (indicating central adiposity), triglycerides and BMI z-score in order of importance.

Discussion

In a recent review article exploring the use of ML across pediatric subspecialties, the authors found that, although ML models have been used for diagnosis of clinical conditions in children, particularly in high income countries with access to structured clinical data through electronic health record systems, there are relatively less scientific publications in disease prognosis. [25] In the current study, we leveraged availability of high-quality data collected over four decades in a large pediatric cohort with later follow-up exams for developing ML models. We validated our predictive models with a separate set of data from patients that were not included in the original training set. To the best of our knowledge this is the first study that explores predictive performance of a ML model in children and adolescents using multiple metabolic risk measures obtained from standardized research examinations specifically designed to predict the outcome of diabetes thereby reducing bias and improving predictive accuracy. The ML model performance was also compared to a traditional binomial regression model. The predictor performance of the five ML models and the ensemble was significantly superior to the LR-model in this study.

In addition to the variables that are components of the IDF defined pediatric MetS, we also explored additional relevant predictors of future diabetes. The random forest ML model was superior in clinical risk prediction compared to all other ML models and the traditional regression model and identified four additional clinical markers (BMI, 2-hour OGTT, HbA1c, and fasting insulin) obtained during childhood that were outside of the MetS components that could add value for future diabetes prediction and incorporated into a clinical decision-making tool. These additional risk clusters identified as variables of importance in our target binary outcome of presence or absence of future diabetes are an important finding.

There are some limitations to our study, the data were obtained from a study population that has a high-risk for obesity and T2D, although our prior study results have been widely validated in other population groups dispelling the notion of lack of generalizability. In addition, to epidemiologic data, addition of “omic” data also add a new dimension towards disease prediction in the arena of personalized medicine. However, the omics data were not available for incorporation into the current model. Lastly the model would need prospective validation in a real-world clinical setting. The strengths of the study are availability of a large amount of structured clinical and laboratory data from research examinations during childhood and subsequent follow up examinations several years later and into adulthood. Another advantage is the availability of high-quality data collected in a consistent manner for four decades in the parent study. Also, since the parent study was designed to study diabetes in this population, it provided a unique opportunity to examine the complex relationship between the risk measures and the outcome of interest without the need for extensive preprocessing. Additionally, all laboratory tests were conducted at the same NIDDK Phoenix laboratory during the entire study period ensuring consistency in the testing processes during different time periods.

In summary, data quality and quantity are key to creating ML models with high predictive accuracy. Our study fulfilled both criteria. Using well-structured clinical data, our ML models exhibited high predictive accuracy. This may contribute towards development of prognostic clinical decision-making tools for pediatric health care providers in youth onset obesity driven metabolic dysfunction and diabetes.

Declarations

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the NIH, the National Institute of Diabetes and Digestive and Kidney Diseases.

Author contributions

M. S. conceptualized the study, provided clinical expertise, assisted with data analysis, and drafting of the manuscript.

E. V. assisted with data preprocessing and statistical analysis, and drafting of the manuscript

All authors reviewed the manuscript

Data availability statement

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request and as per data sharing rules and regulations set forth by the National Institutes of Health Intramural Research Program.

Competing Interests Statement

The authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

1. Fryar CD, Carroll MD, Afful J. Prevalence of overweight, obesity, and severe obesity among children and adolescents aged 2–19 years: United States, 1963–1965 through 2017–2018. NCHS Health E-Stats. 2020.
2. Skinner AC, Perrin EM, Moss LA, Skelton JA. Cardiometabolic Risks and Severity of Obesity in Children and Young Adults. N Engl J Med. 2015 Oct;373(14):1307–17.
3. Tanamas SK, Reddy SP, Chambers MA, Clark EJ, Dunnigan DL, Hanson RL, Nelson RG, Knowler WC, Sinha M. Effect of severe obesity in childhood and adolescence on risk of type 2 diabetes in youth and early adulthood in an American Indian population. Pediatr Diabetes. 2018 Jun;19(4):622–629.
4. Franks PW, Hanson RL, Knowler WC, Sievers ML, Bennett PH, Looker HC. Childhood obesity, other cardiovascular risk factors, and premature death. N Engl J Med. 2010 Feb 11;362(6):485–93.
5. Mayer-Davis EJ, Lawrence JM, Dabelea D, et al. Incidence trends of type 1 and type 2 diabetes among youths, 2002–2012. N Engl J Med 2017;376:1419–1429.
6. Dabelea D, Mayer-Davis EJ, Saydah S, et al. Prevalence of type 1 and type 2 diabetes among children and adolescents from 2001 to 2009. JAMA 2014;311:1778–1786.
7. Reaven, G. M. (1988). Role of insulin resistance in human disease. Diabetes, *37*(12), 1595–1607.
8. Lorenzo C, Okoloise M, Williams K, Stern MP, Haffner SM; San Antonio Heart Study. The metabolic syndrome as predictor of type 2 diabetes: the San Antonio heart study. Diabetes Care. 2003; *26*: 3153–3159.
9. Laaksonen DE, Lakka HM, Niskanen LK, Kaplan GA, Salonen JT, Lakka TA. Metabolic syndrome and development of diabetes mellitus: application and validation of recently suggested definitions of the metabolic syndrome in a prospective cohort study. Am J Epidemiol. 2002; *156*: 1070–1077.
10. Zimmet P, Alberti KG, Kaufman F, Tajima N, Silink M, Arslanian S, Wong G, Bennett P, Shaw J, Caprio S; IDF Consensus Group. The metabolic syndrome in children and adolescents - an IDF consensus report. Pediatr Diabetes. 2007 Oct;8(5):299–306.

11. Barlow SE; Expert Committee. Expert committee recommendations regarding the prevention, assessment, and treatment of child and adolescent overweight and obesity: summary report. Pediatrics. 2007 Dec;120 Suppl 4:S164-92.
12. Sheela N. Magge, Elizabeth Goodman, Sarah C. Armstrong, COMMITTEE ON NUTRITION, SECTION ON ENDOCRINOLOGY, SECTION ON OBESITY. The Metabolic Syndrome in Children and Adolescents: Shifting the Focus to Cardiometabolic Risk Factor Clustering. Pediatrics August 2017; 140 (2): e20171603. 10.1542/peds.2017 – 1603.

Figures

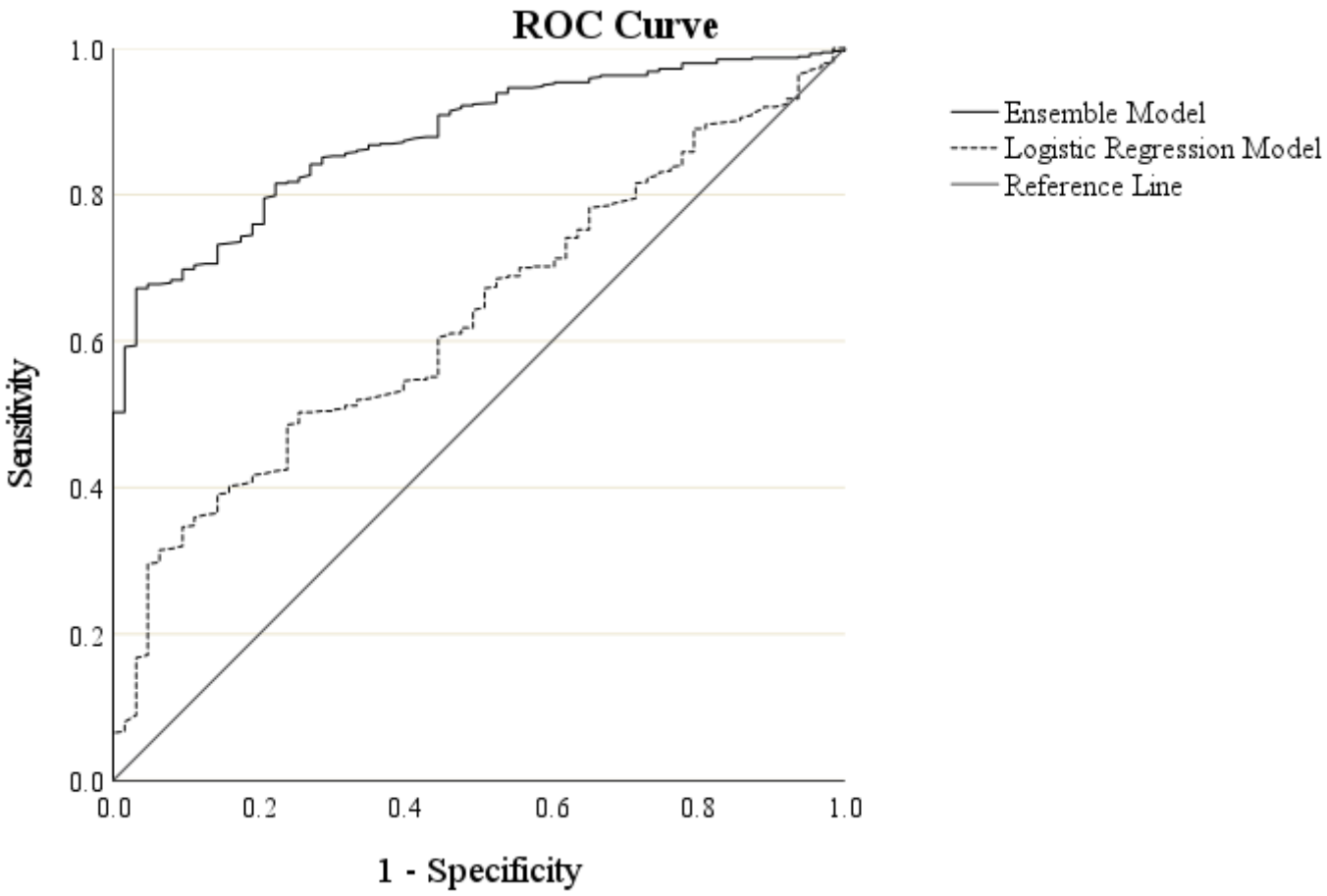

Figure 1

shows the overall performance in predicting diabetes by combining the results of the classification algorithms in an ensemble model in the testing set and compares it to the binomial LR-model.

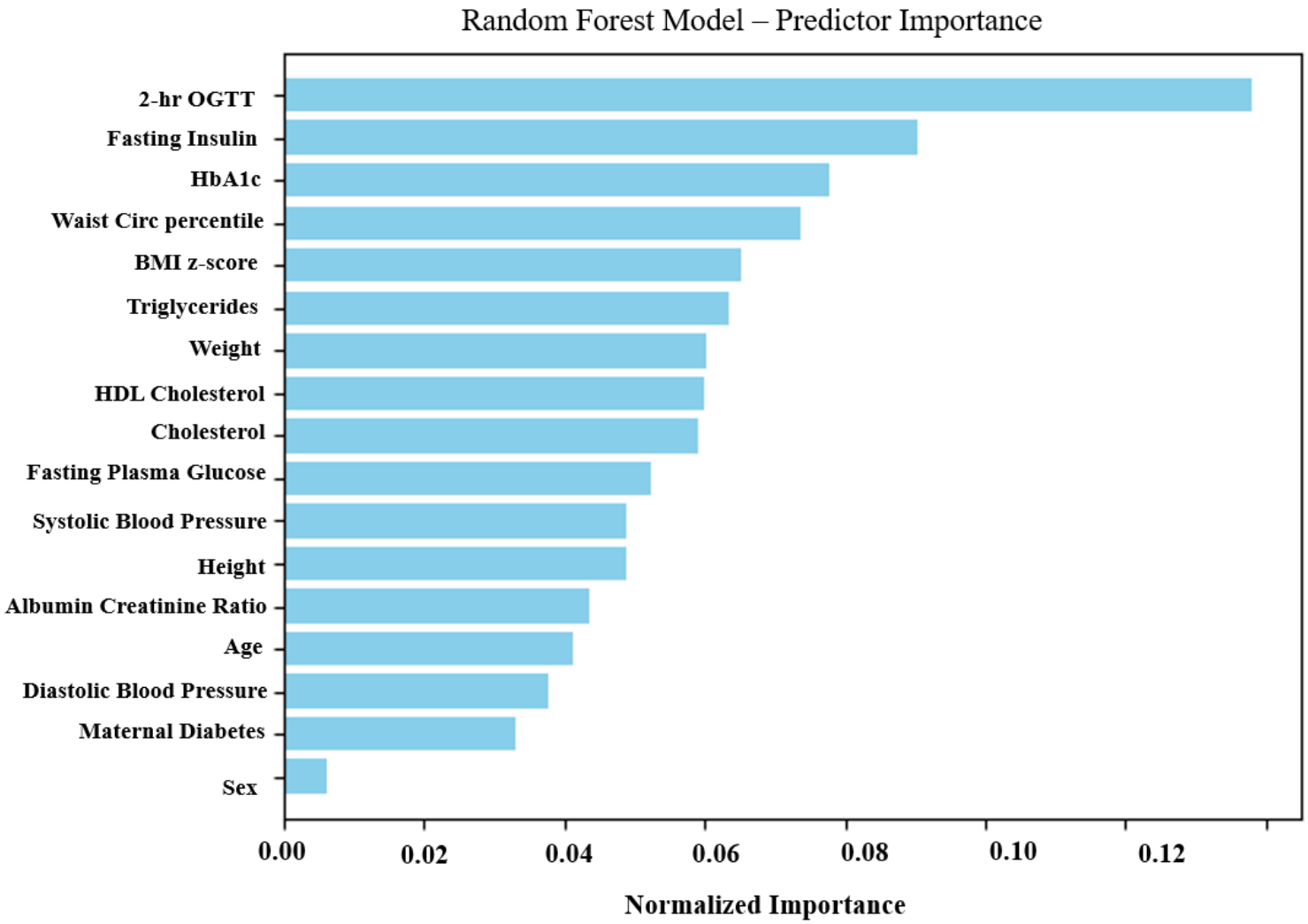

Figure 2

shows the feature importance in the best performing RF model and shows the variables that are most important in predicting diabetes in this study cohort.