BMC Medical Informatics
and Decision Making

RESEARCH Open Access

Using a robust model to detect the association between anthropometric factors and T2DM: machine learning approaches

Nafiseh Hosseini[1,2], Hamid Tanzadehpanah[3,4,5], Amin Mansoori[6], Mostafa Sabzekar[7], Gordon A. Ferns[8], Habibollah Esmaily[9,10*] and Majid Ghayour-Mobarhan[1,4*]

Abstract

Background The aim of this study was to evaluate the potential models to determine the most important anthropometric factors associated with type 2 diabetes mellitus (T2DM).

Method A dataset derived from the Mashhad Stroke and heart atherosclerotic disorders (MASHAD) study comprising 9354 subject aged 65 – 35. 25% (2336 people) of subjects were diabetic and 75% (7018 people) where non-diabetic was used for the analysis of 10 anthropometric factors and age that were measured in all patients. A K-nearest neighbor (KNN) model was used to assess the association between T2DM and selected factors. The model was evaluated using accuracy, sensitivity, specificity, precision and f1-measure parameters. The receiver operating characteristic (ROC) curve and factor importance analysis were also determined. The performance of the KNN model was compared with Artificial neural network (ANN) and support vector machine (SVM) models.

Result After feature selection analysis and assessing multicollinearity, six factors (Mid-arm Circumference (MAC), Waist Circumference (WC), Body Roundness Index (BRI), Body Adiposity Index (BAI), Body Mass Index (BMI), age) were used in the final model. BRI, BAI and MAC factors in males and BMI, BRI, and MAC factors in females were found to have the greatest association with T2DM. The accuracy of the KNN model was approximately 93% for both genders. The best K (number of neighbors) for the model was 4 which had the lowest error rate. The area under the ROC curve (AUC) was 0.985 for men and 0.986 for women. The KNN model achieved the best result of the models explored.

Conclusion The KNN model had a high accuracy (93%) for predicting the association between anthropometric factors and T2DM. Selecting the K parameter (nearest neighbor) has an essential impact on reducing the error rate. Feature selection analysis reduces the dimensions of the KNN model and increases the accuracy of final results.

*Correspondence:
Habibollah Esmaily
esmailyh@mums.ac.ir
Majid Ghayour-Mobarhan
ghayourm@mums.ac.ir

Full list of author information is available at the end of the article

Introduction

One important health concern is the rising prevalence of type 2 diabetes mellitus (T2DM), a condition that is caused by lifestyle choices, including their level of adiposity, and their lack of physical activity, and genetic factors [1]. The complications associated with T2DM include: macrovascular and microvascular disease [1]. Diabetes affects roughly 422 million people globally, according to the most recent statistics provided by the World Health Organization (WHO) [2]. It is predicted that by 2040, the prevalence of T2DM will be 642 million., and is responsible for 1.6 million deaths each year [2]. The prediction of who is likely to develop diabetes is essential. It is feasible to give a better quality of life for patients and society in the future by predicting people who are likely to develop diabetes. This will allow for the reduction of complications, the reduction of the length of hospitalization, the optimization of therapy, and/or the provision of options that are less intrusive [3, 4].

The ability to predict diseases and other health issues is one possible application of artificial intelligence (AI) [5]. Artificial intelligence algorithms can find patterns in the data by evaluating large amounts of patient data, clinical symptoms, and medical information. This can then assist medical professionals in taking the appropriate steps [6]. Machine learning (ML) is a subset of artificial intelligence (AI) that enables computer programs to improve their accuracy in making predictions regarding future events even when they have not been specifically designed to do so [7–9]. In order to provide accurate forecasts of future patient values, machine learning algorithms require historical patient data as input. The utilization of such gathered data can be beneficial for diabetes prediction [7, 10].

Different machine learning classification algorithms such as Naive Bayes (NB), SVM, LR (Linear Regression), Adaboost, RF, KNN (K Nearest Neighbor), DT and NN (Neural Network) have been used to predict diabetes [11–21]. For example, Khanam and Foo et al., (2021) after analyzing some characteristics related to diabetes such as pregnancy, BMI, insulin level, age, blood pressure, skin thickness, glucose, diabetes genealogy function. All the models provided an accuracy of >70% [14]. Chou et al. (2023) used the same factors such as pregnancy, BMI, insulin level, age, diastolic blood pressure, sebum thickness, glucose, and diabetes pedigree function and using different models. Finally, they investigate that RF was the best model [19].

Madhu et al., (2023), used machine learning to predict the probability of a Pima Indians developing diabetes. In that demographic and health records of 768 Pima Indians by LR, DT, random forest, KNN, AdaBoost, NB and XGBoost model analysis was used. The best accuracy (86,61%) was obtained using KNN model [13].

Daanouni, et al. (2019), used four machine learning algorithms (DT, KNN, ANN and deep neural network) to predict T2DM or healthy individuals. These techniques were analyzed on two data sets, one with 2000 instances and the other with 768 cases. Data related to BMI, glucose, blood sugar and pregnancy characteristics were entered into the algorithms. The result of their study showed that KNN has a maximum accuracy of 97.53% and an AUC of 0.96 [18].

As previous studies show the KNN model has a high capability in predicting diabetes. The K-nearest neighbor, or KNN, algorithm is one of the most significant and straightforward approaches to machine learning [22]. Its results are very straightforward and simple to interpret. In this algorithm, a new piece of data is given a classification based on the votes of the majority of its neighbors, and it is then assigned to the class whose data is around it before being categorized. In addition, KNN is able to build highly non-linear and very adaptive decision boundaries for each data point. It can also make complete use of the local information that is available [22].

The anthropometric indicators of multiple individuals who were recruited into the Mashhad Stroke and heart atherosclerotic disorders (MASHAD) study were collected, and KNN was used to examine associations between factors and diabetic persons.

Method

Study population

The participants were included from the baseline of the Mashhad Stroke and heart atherosclerotic disorders (MASHAD) study, Mashhad, north-eastern Iran. 9704 individuals aged 35–65 years participated in this cohort study after cleaning data analysis was performed by 9354 cases. 2336 (25%) cases were diabetic and 7018 (75%) were non-diabetic. T2DM was defined as a fasting blood glucose (FBG) ≥ 126 mg/dl or being treated with available oral hypoglycemic medications or insulin [23]. Age of the subject and anthropometric factors including demispan, Hip Circumference (HC), Mid-arm Circumference (MAC), Waist Circumference (WC), Body Roundness Index (BRI), Body Adiposity Index (BAI), A Body Shape Index (ABSI), Body Mass Index (BMI), Waist-to-height Ratio (WHtR), and Waist-to-hip Ratio (WHR) were measured in all patients.

Statistical analysis

For statistical analyses, SPSS 23 and for classification and modeling SPSS Modeler 18.0 were used. The Kolmogorov–Smirnov (KS) test was used to check the normality of factors. Values are reported as mean ± SD for normally distributed variables (or median and IQR for non-normal distributed variables). Based on the results of the normality test (KS) the parametric tests used to compare diabetic and non-diabetic groups in baseline.

Chi-square test was applied to measure the association between qualitative variables. Also, independent t-test was used for quantitative variables between the two diabetic and non-diabetic groups. All of the analyses were undertaken for males and females, separately.

In addition, KNN as machine learning technique has been used to predict the association between T2DM and anthropometric measurements. To evaluate the models' accuracy, specificity, sensitivity, precision and, F1-measure metrics were used and ROC curves and predictor importance charts were drawn. Also, the KNN model was compared with ANN and SVM models.

K-nearest neighbor (KNN)

KNN is a supervised classifier that classifies cases based on their similarity to each other. It is developed for pattern recognition of data without requiring an exact match to any stored patterns or cases. All cases are points in n-dimensional space. Similar cases are near each other called neighbors. The distance between the two cases is a measure of their dissimilarity. For every new case (holdout), distance from other cases will be computed and placed into the category by the greatest number of nearest neighbors. Distance computation was a Euclidean metric. Normalizing was performed for range input. The factors (predictors) were weighted by importance when computing distances. The models were built by different possible K between 3 and 7 and the best results were selected. All 10 predictors were used in modeling. The minimum change was 0.01 and the seed was set to 12,345 for both gender models [24, 25].

Table 1 Baseline characteristics of male and female

Variables	Diabetes+ (875)	Diabetes- (3080)	*P*-value
Male			
BMI (kg/m^2)	27.52 ± 4.25	26.90 ± 4.26	< 0.001
WC (cm)	97.49 ± 11.35	95.10 ± 10.87	< 0.001
MAC (cm)	30.54 ± 3.51	29.81 ± 4.11	< 0.001
BAI	29.66 ± 4.22	28.68 ± 4.65	< 0.001
BRI	5.19 ± 1.40	4.78 ± 1.45	< 0.001
Age	54.42 ± 6.90	50.63 ± 8.54	< 0.001
Female			
Variables	**Diabetes+ (1461)**	**Diabetes- (3938)**	***P*-value**
BMI (kg/m^2)	29.66 ± 4.50	29.19 ± 4.85	0.001
WC (cm)	101.49 ± 11.58	96.97 ± 11.99	< 0.001
MAC (cm)	30.99 ± 3.62	30.59 ± 3.59	< 0.001
BAI	36.90 ± 5.49	36.95 ± 5.65	0.021
BRI	6.92 ± 1.89	6.15 ± 1.90	< 0.001
Age	54.56 ± 6.74	49.17 ± 8.30	< 0.001

Abbreviations: BMI, Body Mass Index; BAI, Body Adiposity Index; BRI, Body Roundness Index; WC, Waist Circumference; MAC, Mid-arm Circumference

Feature selection

Using feature selection methods enhances model interpretability, reduces training times and overfitting, and increases the accuracy of models in large datasets [24]. Three steps were taken to select the proper features in SPSS Modeler 18.0. At first (Screening) factors with too many missing values (over 70%) or with too little variation to be useful were removed (coefficient variation less than 0.1). Then, inputs were sorted and ranked based on importance (*p*-value), for categorical factors Pearson chi-square was used (Ranking). Finally, features with importance over 0.95% were selected to use in the model (selecting). The 0.95% is an acceptable cutoff point for factor importance. Of course, the threshold can be changed by experts' ideas. Finally, multicollinearity was investigated between factors by Variance inflation factor (VIF). High relation of two or more independent factors leads to a multicollinearity problem [26].

The variance inflation factor for the j^{th} predictor is:

$$\mathrm{VIFj} = \frac{1}{1 - R_j^2}$$

VIF > 10 shows significant relation between factors [26].

Factor importance chart

The factor importance chart showed the relative importance of each factor in making a prediction. The values are relative so, the sum of the values for all factors is 1.0. Predictor importance relates to the importance of each predictor, not the accuracy of it.

Result

The SPSS Modeler 18.0 was used to implement the KNN model. In this version of the software feature selection and KNN classification nodes are available. The software was established on a usual system with 4 Gig RAM and an Intel(R), Corei5-2400 CPU, 3.10 GHz. Also, some statistical analyses such as checking the normality of variables and multicollinearity detection were done by SPSS version 23. The11 factors include BMI, WC, HC, WHR, WHtR, MAC, BAI, ABSI, BRI, Demispan, and age were analyzed. After using feature selection, HC, WHR, ABSI, and Demispan factors had coefficient variation under 0.1 and were removed. Then multicollinearity assessment was done and WHtR with variance inflation factor (VIF) near 132 was removed. So, six factors were used in final model (Table 1). The characteristics of the baseline subjects are shown in Table 1. The *P*-value is reported for comparing two groups using a T-test or chi-square. From 9354 individuals 5399 were women and 3955 were men witch 1461 women and 875 men had T2DM throughout the study.

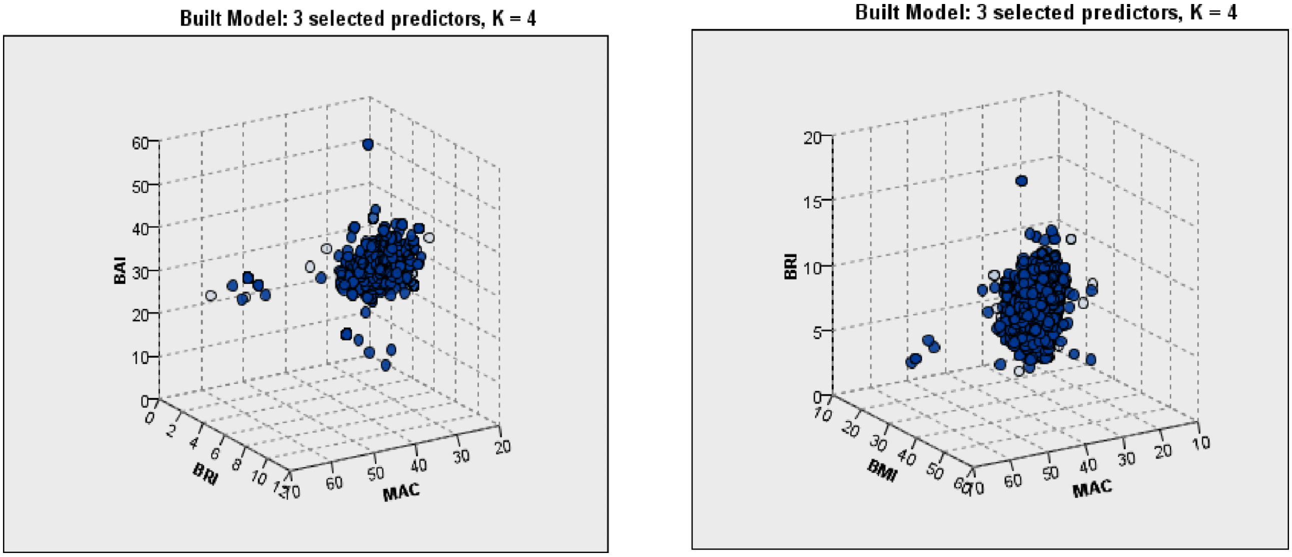

Fig. 1 A lower-directional projection of predictor space for males (**A**) and females (**B**)

After feature selection male and female data were mapped to the n-dimension space separately (Fig. 1). In Fig. 1, each axis represents a factor in the models, and the location of points in the chart shows the values of these factors for cases in the training partitions. The top three factors in each model were selected to show.

The selection of the number of neighbors (k) is important and affects the model performance. In different data sets, the best k would be different. Figure 2 shows the correlation between the number of neighbors (K) and the error rate. The graph shows that the optimum value of K is 4 then the error rate is minimum (Fig. 2). The error rate is calculated as follows:

$$\text{Error rate} = (\text{FP} + \text{FN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Performance metrics of KNN on its optimum value of K are shown in Table 2 for males and females separately. The data were split to the train and test for males and females (80% vs. 20%).

The confusion matrix and area under curve (AUC) are shown in Table 3 for males and females. An AUC > 0.9 shows that detection power of the KNN model is good. As confusion matrix shows FP cases are too low. The ROC curve of diabetic and non-diabetic males and females are shown in Fig. 3. The ROC curve also show TPR vs. FPR is high.

Finally, a predictor importance analysis is done to detect the most important factors for association analysis in this model. The BRI and MAC factors were important for both genders, and BAI for males and BMI for females were the top three important factors (Fig. 4).

Discussion

It is now possible to identify and diagnose many diseases based on clinical parameters and patient history by data mining algorithms [27]. Early detection of T2DM by ML may lead to earlier treatments [28]. In the current study, an acceptable number of patients ($N = 9354$) participated with suitable variables and minimal missing data. The Mashhad stroke and heart atherosclerotic disorder (MASHAD) data set was used include T2DM and healthy individuals [23]. The purpose of this study was to introduce a model using the KNN algorithm to evaluate associations between T2DM and some anthropometric factors including BMI, WC, HC, WHR, MAC, BAI, WHtR, ABSI, BRI, and Demispan. In addition to these indicators, the age of the participants was also included in the study. The advantage of analyzing these indicators compared to clinical and laboratory indicators is that they are non-invasive and their measurement causes less stress to patients. After feature selection, WHR, HC, ABSI and Demispan factors were removed from the model (Table 1). A review of the ML models and feature selection methods used in detecting T2DM shows that using FS can improve the performance of ML models [21]. Also, a VIF was used to detect multicollinearity among remained factors and WHtR factor with high VIF was removed. So, the final model was constructed by six factors. There are studies that show multicollinearity has lower effect on non-linear classifications [29]. Also, Morris et al. showed that the multicollinearity problem has no essential effect on the accuracy of the classification models. However, it would be confounding if the goal of a study is factor importance extraction. To address both goals, the researchers may need to consider their relative importance [30]. Because of the attention of this research

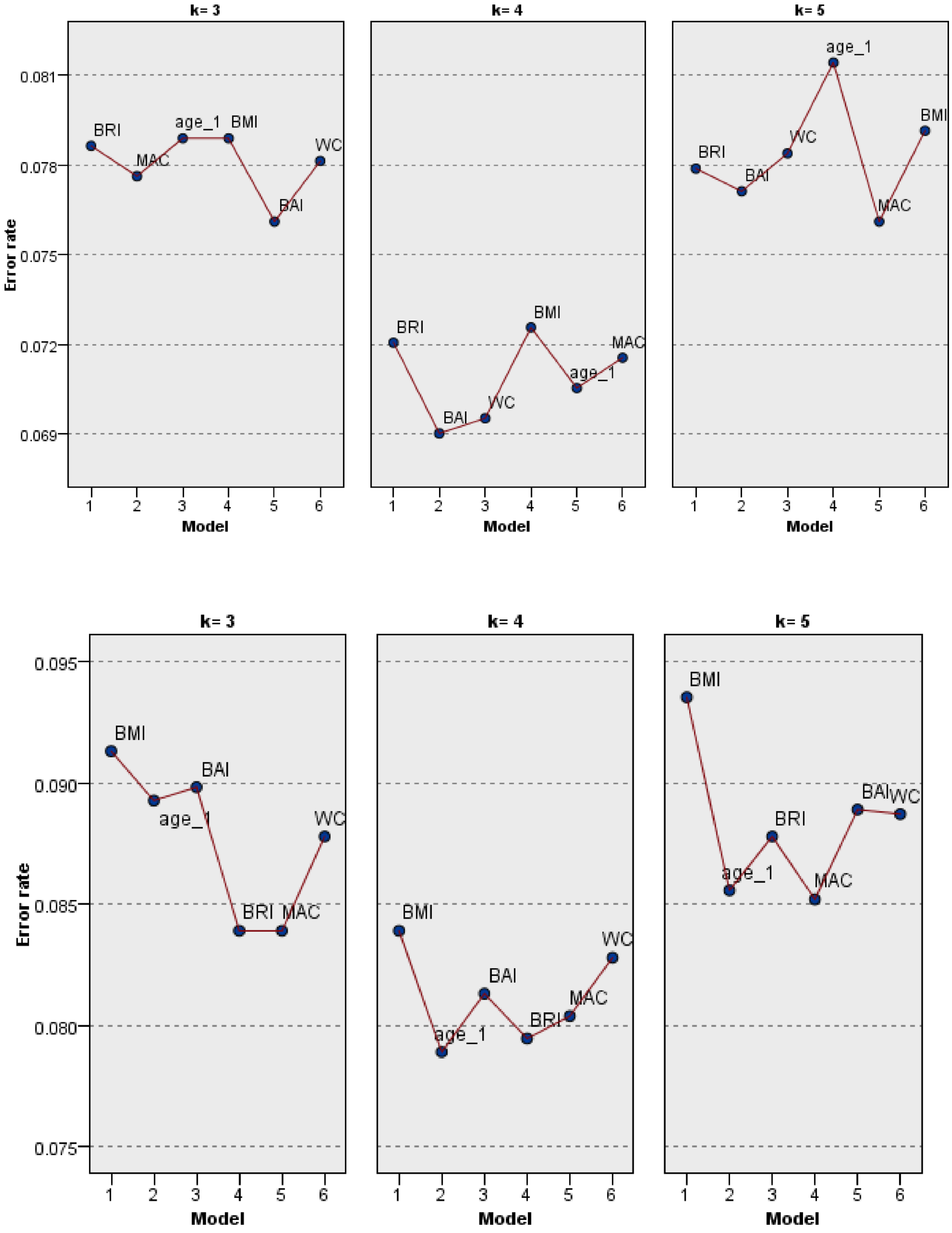

Fig. 2 Predictor selection in different number of neighbors (K) for males (**A**) and females (**B**)

Table 2 Performance parametric of the KNN model (test) in comparison ANN and SVM

	Model	Accuracy (%)	Sensitivity/ Recall(%)	Specificity(%)	Precision(%)	F1-measure
Male	KNN	**93.28**	71.08	99.06	95.16	**81.38**
	SVM	82.46	74.1	95.92	82.55	78.1
	ANN	91.42	69.88	97.34	87.22	77.59
Female	KNN	**93.49**	79.87	98.84	96.41	**87.36**
	SVM	76.95	37.62	92.37	77.64	47.9
	ANN	87.92	80.2	90.94	88.8	78.9

Table 3 Confusion Matrix, AUC and Gini Index of the KNN model for the males and females

Test (1,076)	Females		Test (804)	Males	
	Predict			Predict	
Actual	Diabetic	Non-Diabetic	**Actual**	Diabetic	Non-Diabetic
Diabetic	242	61	Diabetic	118	48
Non-Diabetic	9	764	Non-Diabetic	6	632
	AUC	**Gini**		**AUC**	**Gini**
	0.986	0.973		0.985	0.97

to both goals we used VIF and removed WHtR with high VIF but used WC even had a VIF near 10 because had not essential effects on the factor importance chart.

Our findings showed that some factors have a high association with diabetes. Among the indicators analyzed, BRI and BAI, and MAC for males and BMI, BRI, and MAC for females were the most important indicators. In comparison to the SVM and ANN models, the KNN achieved the best results. Accuracy, f1-measure and precision of the KNN model was much better than the other models. Specificity of all tree models was good but the KNN was the best. Also, the sensitivity of KNN was 71.08% for males and 79.87% for females. The ROC curve and AUC also confirmed the results. By using the logistic regression (LR) technique, Saberi-Karimian et al. demonstrate that among the anthropometric indices, WC and BIA in males and Demispan and WC in females had the highest connection with the probability of developing T2DM [31]. The decision tree method was also utilized in that study, and the findings indicated that WC, followed by HC, and BAI, had the most significant impact on the probability of developing T2DM [31]. They also used the Mashhad stroke and heart atherosclerotic disorder (MASHAD) data set [23], and the DT model accuracy was 77.59% and 79.77% for men and women, respectively [31]. However, using the KNN model in the current investigation indicated greater accuracy, with values of 93.28% and 93.49%, respectively, for men and women.

Several researchers have investigated the association between T2DM and some parameters using the KNN model [15, 25, 32–36]. These studies show that finding the most important factors associated with T2DM depends on the characteristics of the models. Habibi et al. (2015) investigated risk variables for T2DM in 450 diabetics and 450 healthy patients by using DT algorithms. According to their findings, the most significant risk factors for T2DM are older age, a higher body mass index (BMI), a history of diabetes in one's family, and higher systolic blood pressure [37]. In the diabetes prediction model that was built by Wu et al. (2018), the authors developed a KNN and regression model to predict diabetes. According to the model that they created, the most significant risk factors were the age, body mass index (BMI), fasting plasma glucose (FPG), and triglycerides. They demonstrate that the performance of the newly developed model is superior, with accuracy ratings of 81.55% [38]. Khaleel et al., (2023) used LR, NB and KNN model to predict T2DM using Pima Indian Diabetes (PIDD) dataset. In their research, the accuracy of the KNN model was the lowest one (KNN = 69%) [15]. Josh et al., (2021) also used the PIDD dataset to predict T2DM. Glucose, BMI, and age were the most important factors in their research [16]. KNN, DT, SVM, RF, NB, and LR are the machine learning algorithms that were used by Sarwar et al., (2018) in their investigation into predicting diabetes ($n = 768$). Based on the results of their experiments, SVM and KNN provide the maximum accuracy. The accuracy provided by both of these algorithms, at 77%, is the greatest among the other four algorithms that were utilized [39]. A successful model for the diagnosis of T2DM was developed by Barakat and his colleagues (2010) through the process of enhancing and optimizing the technique of SVM. The data set that was used contained information on 4682 patients, each of whom had their gender, BMI, blood pressure, cholesterol, and blood glucose levels recorded. One of the benefits of this study was the substantial number of samples used in the modeling. On the other hand, the most notable drawback of this method is the low number of variables used in the modeling [40]. A SVM model was employed in a study conducted by Purnami et al. (2009) to diagnose diabetes

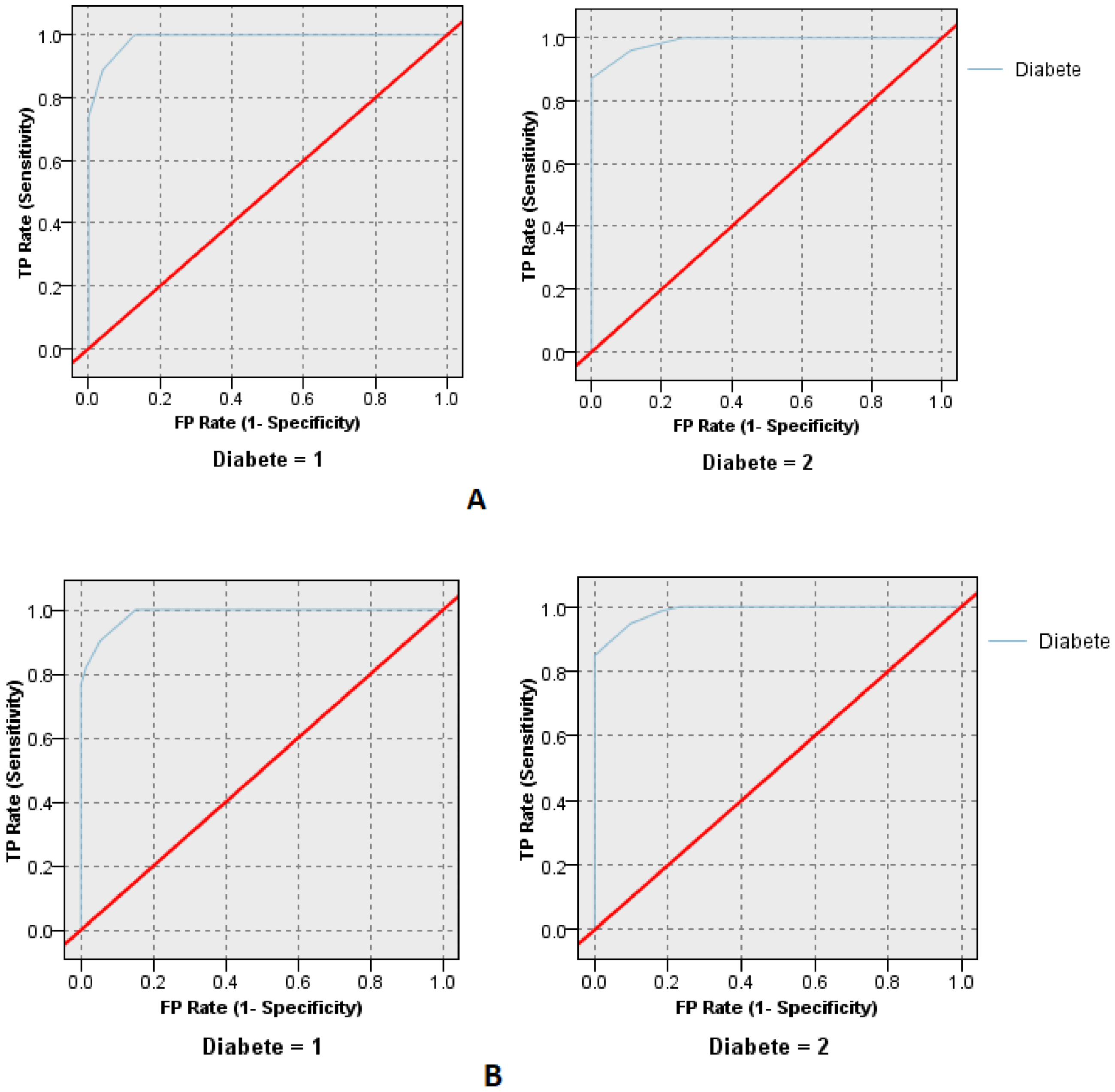

Fig. 3 ROC curve of diabetic (Diabete = 1) and non-diabetic (Diabete = 2) for male (**A**) and female (**B**) cases

in 768 individuals by taking into account eight different variables, including blood pressure and the amount of insulin that was injected [41]. It can be inferred from the findings of this project and other research on data mining-based diabetes association investigation that various algorithms introduce various measurement markers for examination. However, practically all of them agreed that diabetes and BMI are directly related [42–44]. Based on our research BMI in females is important. In this regard, Strings et al. (2023) demonstrated that racial differences also exist in the association between BMI and diabetes [45]. For Latinos and Whites, BMI was related to a 10% increased risk of prediabetes/T2DM compared to having normal HbA1c levels. However, among blacks, the link between BMI and prediabetes/T2DM was noticeably weaker. The contradictory correlation indicates that BMI alone is not a reliable indicator of T2DM [45]. When evaluating the likelihood of acquiring T2DM, reliance on BMI causes bias, especially in blacks. Data mining enables us to take into account more signs as a result. The other indices such as BRI and WC, which had the highest scores among the candidate indicators, were looked at in this study [45]. Bai et al. (2022) studied 69,388 people under 60 to determine how several anthropometric

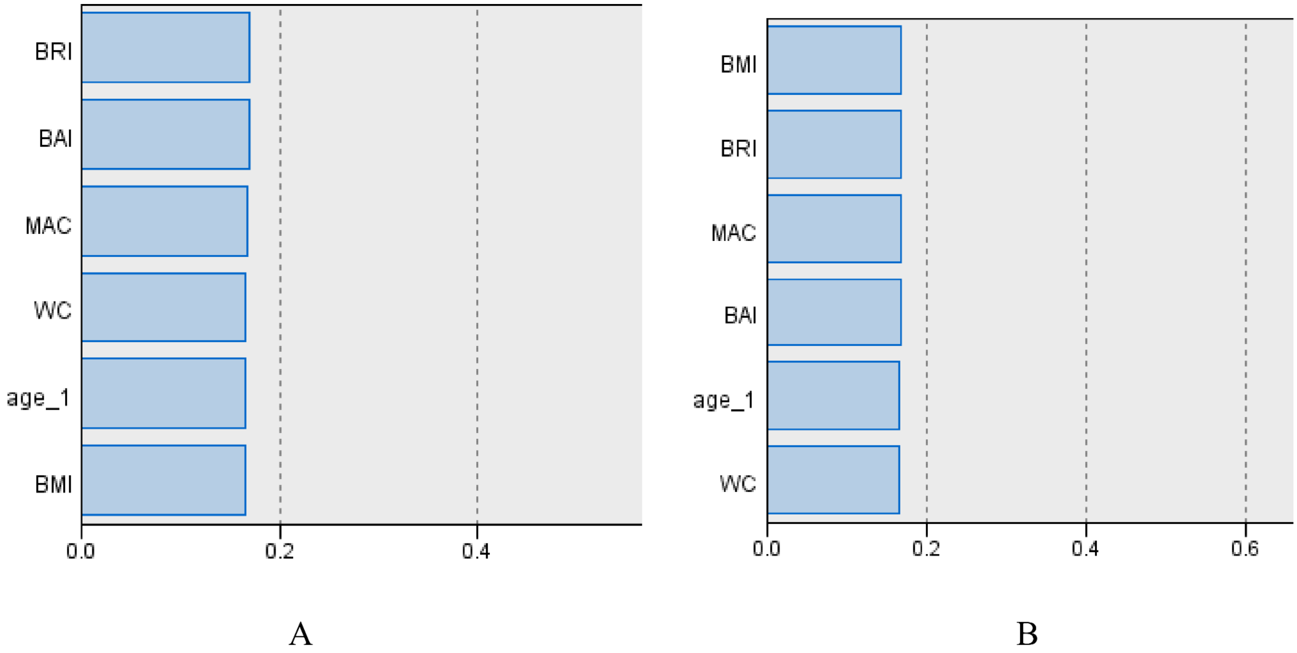

Fig. 4 Predictor Importance for males (**A**) and females (**B**)

variables (BMI, WC, WHtR, BAE, and BRI) related to diabetes [46]. They discovered that a higher risk of T2DM was linked to higher BMI and WC. The sex subgroup analysis produced the same results. In terms of T2DM, there was no additive interaction between BMI and WC. Also, T2DM was favorably correlated with WHtR, BRI, and BAE in both men and women [46]. The results of this study also show that WHtR, BRI, and BAI in males and BMI, MAC, and BRI in females have high association with T2DM. Bhat et al. (2023) used 6 MLs: RF, Multi-Layer Perceptron (MLP), SVM, Gradient Boost (GB), DT, and LR. RF was the most accurate classifier with 98% accuracy. They used K-fold method to enhance classifier results [20].

This current study builds on previous studies. The large number of people studied, the use of anthropometric indicators that can be easily measured and the use of a suitable model with high performance parameters are improvements over previous studies.

As shown in Table 1, the data set was imbalanced. This issue can affect the accuracy of models. Also, in some cases, the results may be biased. This means that even though the classifier achieved high accuracy, the classification is not done well and most of the cases of the lower class are wrongly classified. But in this research, the confusion matrix showed that the FP and FN of the KNN model were low and the ROC curve and high AUC confirmed that. So, the results are acceptable.

Our future work will focus on the integration of other methods and features into the used model for tuning the parameters of models for better accuracy. Jaiswal et al. (2021), conducted a review of machine learning approaches to predict diabetic cases and finding the importance predictors. They mentioned that current research have some limitations and are not tested on different datasets from different countries [17]. External validation is critical for ensuring reliability before clinical deployment [47, 48]. So, the model will be evaluated with another dataset in feature works externally.

Author contributions

Nafiseh Hosseini: conception, data analyzing; Hamid Tanzadehpanah: data analyzing; Amin Mansoori: drafting the article; Mostafa Sabzekar: revising the article; Gordon A. Ferns: revising the article; Habibollah Esmaily: corresponding author and drafting the article; Majid Ghayour-Mobarhan: corresponding author.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All the participants consented to take part in the study by signing written informed consent. The study protocol was reviewed and all methods are approved by the Ethics Committee of Mashhad University of Medical Sciences with approval number IR.MUMS.REC.1386.250. The study was approved on July 2007. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication

Not Applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]International UNESCO Center for Health-Related Basic Sciences and Human Nutrition, Mashhad University of Medical Sciences, Mashhad 99199-91766, Iran
[2]Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[3]Antimicrobial Resistance Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[4]Metabolic Syndrome Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[5]Basic Sciences Research Institute, Mashhad University of Medical Sciences, Mashhad, Iran
[6]Department of Applied Mathematics, School of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran
[7]Department of Computer Engineering, Birjand University of Technology, Birjand, Iran
[8]Brighton and Sussex Medical School, Division of Medical Education, Brighton, UK
[9]Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran
[10]Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran

Received: 30 April 2024 / Accepted: 20 January 2025
Published online: 31 January 2025

References

1. Manaswini TK, Nayak P, Harshitha VS, Barlapudi S, editors. Predictions of Diabetic Mellitus using ML Techniques: A Systematic Overview. 2023 International Conference on Sustainable Computing and Systems S. (ICSCSS); 2023: IEEE.
2. Tumuluru P, Burra LR, Sushanth KK, Vali SN, SaiBaba CHMH, Yellamma P, editors. DPMLT: Diabetes Prediction Using Machine Learning Techniques. 2022 International Conference on Electronics and Renewable Systems (ICEARS); 2022 16–18 March 2022.
3. Oyewole OO, Ale AO, Ogunlana MO, Gurayah T. Burden of disability in type 2 diabetes mellitus and the moderating effects of physical activity. World J Clin Cases. 2023;11(14):3128.
4. Salom Vendrell C, García Tercero E, Moro Hernández JB, Cedeno-Veloz BA. Sarcopenia as a little-recognized comorbidity of type II diabetes Mellitus: a review of the diagnosis and treatment. Nutrients. 2023;15(19):4149.
5. Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. Discover Artif Intell. 2023;3(1):5.
6. Gupta NS, Kumar P. Perspective of artificial intelligence in healthcare data management: a journey towards precision medicine. Comput Biol Med. 2023:107051.
7. Alanazi N, Alruwaili Y, Alazmi A, Alazmi A, Alanazi M, Alruwaili W. A Systematic Review of Machine Learning and Artificial Intelligence for Diabetes Care. J Health Inf Developing Ctries. 2023;17:01.
8. Ghazizadeh H, Shakour N, Ghoflchi S, Mansoori A, Saberi-Karimiam M, Rashidmayvan M, et al. Use of data mining approaches to explore the association between type 2 diabetes mellitus with SARS-CoV-2. BMC Pulm Med. 2023;23(1):1–14.
9. Mansoori A, Farizani Gohari NS, Etemad L, Poudineh M, Ahari RK, Mohammadyari F et al. White blood cell and platelet distribution widths are associated with hypertension: data mining approaches. Hypertens Res. 2023:1–14.
10. Mansoori A, Hosseini ZS, Ahari RK, Poudineh M, Rad ES, Zo MM et al. Development of Data Mining algorithms for identifying the best anthropometric predictors for Cardiovascular Disease: MASHAD Cohort Study. High Blood Press Cardiovasc Prev. 2023:1–11.
11. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes mellitus with machine learning techniques. Front Genet. 2018;9:515.
12. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci. 2020;167:706–16.
13. Madhu B, Aerranagula V, Mahomad R, Ravindernaik V, Madhavi K, Krishna G, editors. Techniques of Machine Learning for the Purpose of Predicting Diabetes Risk in PIMA Indians. E3S Web of Conferences; 2023: EDP Sciences.
14. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. Ict Express. 2021;7(4):432–9.
15. Khaleel FA, Al-Bakry AM. Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings. 2023;80:3200-3.
16. Joshi RD, Dhakal CK. Predicting type 2 diabetes using logistic regression and machine learning approaches. Int J Environ Res Public Health. 2021;18(14):7346.
17. Jaiswal V, Negi A, Pal T. A review on current advances in machine learning based diabetes prediction. Prim Care Diabetes. 2021;15(3):435–43.
18. Daanouni O, Cherradi B, Tmiri A, editors. Predicting diabetes diseases using mixed data and supervised machine learning algorithms. Proceedings of the 4th International Conference on Smart City Applications; 2019.
19. Chou C-Y, Hsu D-Y, Chou C-H. Predicting the onset of diabetes with machine learning methods. J Personalized Med. 2023;13(3):406.
20. Bhat SS, Selvam V, Ansari GA, Ansari MD, Rahman MH. Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district bandipora. Comput Intell Neurosci. 2022;2022(1):2789760.
21. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A. Investigating health-related features and their impact on the prediction of diabetes using machine learning. Appl Sci. 2021;11(3):1173.
22. NirmalaDevi M, Alias Balamurugan SA, Swathi U, editors. An amalgam KNN to predict diabetes mellitus. 2013 IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN); 2013: IEEE.
23. Ghayour-Mobarhan M, Moohebati M, Esmaily H, Ebrahimi M, Parizadeh SMR, Heidari-Bakavoli AR, et al. Mashhad stroke and heart atherosclerotic disorder (MASHAD) study: design, baseline characteristics and 10-year cardiovascular risk estimation. Int J Public Health. 2015;60:561–72.
24. Gupta SC, Goel N, editors. Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT); 2020: IEEE.
25. Prasad BS, Gupta S, Borah N, Dineshkumar R, Lautre HK, Mouleswararao B. Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach. Prev Med. 2023;174:107619.
26. Sundus KI, Hammo BH, Al-Zoubi MB, Al-Omari A. Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset. Inf Med Unlocked. 2022;33:101088.
27. Mansoori A, Hosseini ZS, Ahari RK, Poudineh M, Rad ES, Zo MM, et al. Development of Data Mining algorithms for identifying the best anthropometric predictors for Cardiovascular Disease: MASHAD Cohort Study. High Blood Press Cardiovasc Prev. 2023;30(3):243–53.
28. Rastogi R, Bansal M. Diabetes prediction model using data mining techniques. Measurement: Sens. 2023;25:100605.
29. Suleiman S, Badamsi S. Effect of multicollinearity in predicting diabetes mellitus using statistical neural network. Euro J Adv Eng Technol. 2019;6(6):30–8.
30. Lieberman MG, Morris JD. The precise effect of multicollinearity on classification prediction. Multiple Linear Regres Viewpoints. 2014;40(1):5–10.
31. Saberi-Karimian M, Mansoori A, Bajgiran MM, Hosseini ZS, Kiyoumarsioskouei A, Rad ES, et al. Data mining approaches for type 2 diabetes mellitus prediction using anthropometric measurements. J Clin Lab Anal. 2023;37(1):e24798.
32. Raikwal J, Saxena K. Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. Int J Comput Appl. 2012;50(14).
33. Premamayudu B, Muralikrishna K, Pramodh K. Diabetes prediction using machine learning KNN-Algorithm technique. International Journal of Innovative Science and Research Technology; 2022.
34. Pawlovsky AP, editor. An ensemble based on distances for a kNN method for heart disease diagnosis. 2018 international conference on electronics, information, and communication (ICEIC); 2018: IEEE.
35. Oza A, Bokhare A, editors. Diabetes prediction using logistic regression and K-nearest neighbor. Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 2; 2022: Springer.
36. Mustafa MS, Simpen IW, editors. Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba. SISITI: Seminar Ilmiah Sistem Informasi dan Teknologi Informasi; 2019.
37. Habibi S, Ahmadi M, Alizadeh S. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Global J Health Sci. 2015;7(5):304.
38. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Inf Med Unlocked. 2018;10:100–7.

39. Sarwar MA, Kamal N, Hamid W, Shah MA, editors. Prediction of diabetes using machine learning algorithms in healthcare. 2018 24th international conference on automation and computing (ICAC); 2018: IEEE.
40. Barakat N, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed. 2010;14(4):1114–20.
41. Purnami SW, Embong A, Zain JM, Rahayu S. A new smooth support vector machine and its applications in diabetes disease diagnosis. J Comput Sci. 2009;5(12):1003–8.
42. Howell NA, Booth GL. The weight of place: built environment correlates of obesity and diabetes. Endocr Rev. 2022;43(6):966–83.
43. Aggarwal R, Bibbins-Domingo K, Yeh RW, Song Y, Chiu N, Wadhera RK, et al. Diabetes screening by race and ethnicity in the United States: equivalent body mass index and age thresholds. Ann Intern Med. 2022;175(6):765–73.
44. Zhang S, Liu H, Li N, Dong W, Li W, Wang L, et al. Relationship between gestational body mass index change and the risk of gestational diabetes mellitus: a community-based retrospective study of 41,845 pregnant women. BMC Pregnancy Childbirth. 2022;22(1):336.
45. Strings S, Wells C, Bell C, Tomiyama AJ. The association of body mass index and odds of type 2 diabetes mellitus varies by race/ethnicity. Public Health. 2023;215:27–30.
46. Bai K, Chen X, Song R, Shi W, Shi S. Association of body mass index and waist circumference with type 2 diabetes mellitus in older adults: a cross-sectional study. BMC Geriatr. 2022;22(1):1–10.
47. Ho SY, Phua K, Wong L, Goh WWB. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns. 2020;1(8).
48. Chalkidis G, McPherson JP, Beck A, Newman MG, Guo J-W, Sloss EA, et al. External validation of a machine learning model to predict 6-month mortality for patients with advanced solid tumors. JAMA Netw Open. 2023;6(8):e2327193–e.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.