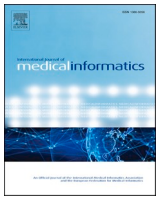
International Journal of Medical Informatics 179 (2023) 105228

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model

Yitayeh Belsti [a,e], Lisa Moran [a], Lan Du [d], Aya Mousa [a], Kushan De Silva [c], Joanne Enticott [a,*], Helena Teede [a,b,*]

[a] Monash Centre for Health Research and Implementation (MCHRI), Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia
[b] Monash Health, Melbourne, Australia
[c] Department of Radiation Sciences, Faculty of Medicine, Umeå University, Sweden
[d] Monash University, Faculty of Information Technology
[e] University of Gondar, College of Medicine and Health Science, Ethiopia

ARTICLE INFO

Keywords:
Machine learning
Predictive model
Prognosis
Gestational diabetes mellitus

ABSTRACT

Background: Early identification of pregnant women at high risk of developing gestational diabetes (GDM) is desirable as effective lifestyle interventions are available to prevent GDM and to reduce associated adverse outcomes. Personalised probability of developing GDM during pregnancy can be determined using a risk prediction model. These models extend from traditional statistics to machine learning methods; however, accuracy remains sub-optimal.

Objective: We aimed to compare multiple machine learning algorithms to develop GDM risk prediction models, then to determine the optimal model for predicting GDM.

Methods: A supervised machine learning predictive analysis was performed on data from routine antenatal care at a large health service network from January 2016 to June 2021. Predictor set 1 were sourced from the existing, internationally validated Monash GDM model: GDM history, body mass index, ethnicity, age, family history of diabetes, and past poor obstetric history. New models with different predictors were developed, considering statistical principles with inclusion of more robust continuous and derivative variables. A randomly selected 80% dataset was used for model development, with 20% for validation. Performance measures, including calibration and discrimination metrics, were assessed. Decision curve analysis was performed.

Results: Upon internal validation, the machine learning and logistic regression model's area under the curve (AUC) ranged from 71% to 93% across the different algorithms, with the best being the CatBoost Classifier (CBC). Based on the default cut-off point of 0.32, the performance of CBC on predictor set 4 was: Accuracy (85%), Precision (90%), Recall (78%), F1-score (84%), Sensitivity (81%), Specificity (90%), positive predictive value (92%), negative predictive value (78%), and Brier Score (0.39).

Conclusions: In this study, machine learning approaches achieved the best predictive performance over traditional statistical methods, increasing from 75 to 93%. The CatBoost classifier method achieved the best with the model including continuous variables.

1. Background

Gestational diabetes mellitus (GDM) is a condition in which the body is unable to utilize insulin effectively, leading to insulin resistance and glucose intolerance [1]. It is defined as any level of impaired glucose tolerance that appears or is first detected during pregnancy [2] and is increasing globally, with up to 1 in 6 pregnancies now affected [3–7]. This is predominantly due to increasing risk factors including rising obesity, gestational weight gain and advancing maternal age [8–11]. GDM increases the risk of adverse outcomes, including stillbirth,

* Corresponding authors at: Monash Centre for Health Research and Implementation (MCHRI), Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia (H. Teede).
E-mail addresses: joanne.enticott@monash.edu (J. Enticott), helena.teede@monash.edu (H. Teede).

https://doi.org/10.1016/j.ijmedinf.2023.105228
Received 7 April 2023; Received in revised form 1 September 2023; Accepted 19 September 2023
Available online 21 September 2023

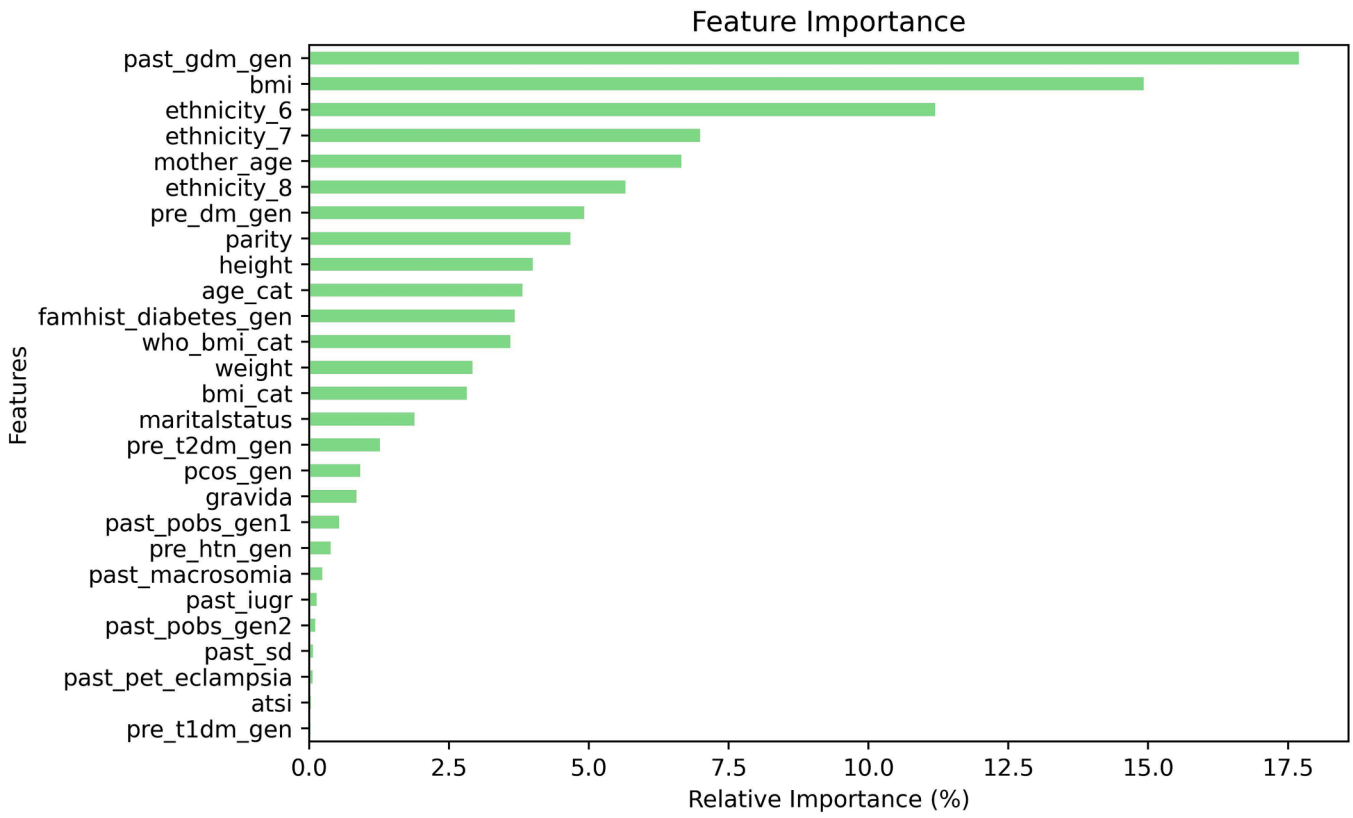


Fig. 1. Feature importance analysis result by CatBoost classifier algorithm. **Footnote for** Fig. 1**:** X-axis: Relative importance of the variable, Y-axis: list of all variables, past_gdm_gen: past history of gestational diabetes mellitus, bmi: body mass index, ethnicity_6: ethnicity (six categories), ethnicity_7: ethnicity (seven categories), mother_age: maternal age(continuous), ethnicity_8: ethnicity(eight categories), pre_dm_gen: previous diabetes (categorical), age_cat: age(six categories): famhist_diabetes_gen: family history of diabetes, who_bmi_cat: body mass index(world health categorization based categorization), bmi_cat: body mass index (categorical), pre_t2dm_gen: previous history of type 2 diabetes, pcos_gen: polycystic ovarian syndrome(categorical), past_pobs_gen1: past poor obstetric history (categorical),

premature birth, neonatal morbidity, and even long-term implications [1,12,13]. Early identification of pregnant women with a high risk of developing GDM is desirable, as prevention is highly effective and can be implemented early to reduce both GDM and associated adverse maternal and neonatal complications [14].

For each woman, the probability of developing GDM may be determined using personalized health data, using a clinical risk prediction model [15,16]. Existing GDM risk prediction models have applied inputs including demographic, anthropometric, clinical and laboratory data. They have been developed using traditional regression analysis or, more recently, using machine learning (ML) methods [17,18]. Due to their predictive performance, parsimonious models derived from easily accessible data, have been recommended and used in clinical practice [19,20]. One example is the Monash GDM risk prediction model by Teede et al. [21] which used routine health data and applied logistic regression methods. Six easily accessible predictors: age, body mass index (BMI) at booking, history of GDM, family history of diabetes, previous poor obstetric outcomes, and ethnicity, achieved fair discriminative ability. The predictive power of this model was externally validated in different population groups internationally [22–24] and temporally in the same population group [25]. Implementation was recommended and occurred in practice, during the coronavirus disease (COVID-19) pandemic [26]. However, performance was still suboptimal.

ML is a common terminology describing diverse, flexible, novel, and complex techniques that enhance the performance of a computer system on specific task by developing algorithms and statistical models [27,28]. The developed algorithms can automatically learn and make predictions or decisions based on patterns and inferences in data, without being explicitly programmed to do so. The popularity of ML techniques in clinical prediction models is increasing, with increasing availability of data and with the potential for increased predictive performance. Due to its ability to handle complex interactions and nonlinearities among input variables, ML has been reported to outperform traditional regression models in some studies [28–31], but this has not been shown consistently [32]. The value of ML using limited variables in GDM remains unclear, in comparison to logistic regression [33].

Hence, we aimed to explore various new ML techniques alongside traditional logistic regression in a large dataset from ethnically diverse pregnant women. First, we created four groupings of predictors (predictor set 1 to 4), as based on the literature it is unclear on the optimal predictor set. Predictor set 1 was the same categorical input variables as in the internationally validated, Monash GDM model [21]. Predictor set 2 was the same as the first set except age and BMI were included as continuous instead of categorical input variables. Predictor set 3 was the same as the second set with the addition of another input variable (parity). Predictor set 4 was the same as the third set except BMI was replaced with its derivative variables (height and weight). Finally, we investigated each predictor set using 11 ML techniques and traditional logistic regression to identify the model with the optimal predictive performance.

2. Methods

2.1. Study population and data sources

Routinely collected health data from n = 48,502 singleton pregnant women at Monash Health maternity hospitals from January 2016 to

Table 1
Input predictors and their types across four predictor sets.

	Age	BMI	Hx GDM	Family Hx DM	Poor obstetric Hx	Ethnicity	Parity	Weight	Height
Model 1	Y [cat]	Y [cat]	Y[d]	Y[d]	Y [d]	Y [cat]	-	-	-
Model 2	Y	Y	Y[d]	Y[d]	Y [d]	Y [cat]	-	-	-
Model 3	Y	Y	Y[d]	Y[d]	Y [d]	Y [cat]	Y	-	-
Model 4	Y	-	Y[d]	Y[d]	Y [d]	Y [cat]	Y	Y	Y

Table 2
Baseline characteristics of study participants.

Variable	Category	Frequency	Percent
Age (year)	<=24	6176	12 70
	25–29	13,806	28 50
	30–34	17,607	36 30
	35–39	8830	18 20
	>=40	2082	4 30
BMI	13–19 9	4169	8 60
	20–24 9	19,144	39 50
	25–26 9	6632	13 70
	27–29 9	7286	15 00
	30–34 9	6276	12 90
	35+	4749	9 80
Parity	0	19,357	39 90
	>=1	29,145	60 10
Past GDM	No	44,770	92 30
	Yes	3732	7 70
Family history of DM	No	29,435	60 70
	Yes	19,067	39 30
Past history of poor obstetric outcomes	No	43,329	89 30
	Yes	5173	10 70
GDM	No	38,159	78 70
	Yes	10,343	21 30
Ethnicity	Caucasian	21,256	43 80
	Oceania (not white-Australian or white-New Zealander)	1852	3 80
	Middle-Eastern, North African, or Sub-Saharan African	2967	6 10
	Southern and Central Asian	14,549	30 00
	South-East and North-East Asian	7778	16 00
	Other	100	0 20

Abbreviations: BMI: Body Mass Index, GDM: Gestational Diabetes Mellitus, DM: Diabetes Mellitus.

June 2021 were used. As one of the largest public health networks in Australia, Monash Health provides services to over two million people across their lifespan, includes 8 hospitals and delivers over 12,000 births per year. This health service operates within Australia's universal freely accessible public healthcare system and cares for women from over 78 countries with a large immigrant population and 65% of mothers born overseas.

2.2. Outcome and predictors

This study built on the existing Monash GDM prediction model developed on an earlier dataset from the same health network [21]. This robustly developed model provided a baseline set of predictors to build on and this was designated as predictor set 1. The primary outcome of this study was diagnosis of GDM, which was defined by the International Association of Diabetes and Pregnancy Study Groups (IADPSG) diagnostic criteria using the 75 g OGTT single measurement of plasma glucose concentration (0, 1, or 2 h of ≥ 5 1, ≥10 1, ≥8 5 mmol/L, respectively) at 24 to 28 gestational weeks [34–36].

Predictor set 1 has six categorical input variables: Binary outcomes (Yes/No) included history of GDM; family history of diabetes; and history of poor obstetric outcome(s). A history of poor obstetric outcome was indicated with any one of the following: a history of macrosomia, shoulder dystocia, pre-eclampsia or eclampsia. Age was categorized into five categories (<25, 25–29, 30–34, 35–39, ≥40 years); BMI had six categories (<20 0, 20 0–24 9, 25 0–26 9, 27 0–29 9, 30 0–34 9, ≥35 0 kg/m^2); and eight ethnicity categories. In the original model [21] the ethnicity categories were: Anglo-Australian, Polynesian, Mainland SE Asian, Maritime SE Asian, Chinese Asian, Southern Asian, African, and other. However, later during updating and temporal validation of the Monash GDM model [25], the ethnicity category was collapsed into six 'categories': Caucasian, Southern and Central Asian, South East and North East Asian, North African, Middle Eastern or Sub-Saharan African, Oceanian not Australian, and other. In eight categories ethnicity was assigned based on country of birth, but to reflect international ethnicity categories and align with the Australian Standard Classification of Cultural and Ethnic Groups [37] later ethnicity was assigned based on self-report (if missing inferred from preferred language and country of birth).

Predictor set 2 was the same as the first set except age and BMI were included as continuous instead of categorical input variables. This was initially examined because of the literature that suggests that categorising continuous variables will often result in loss of information and poorer predictive performance.

Predictor set 3 was the same as the second set with the addition of other input variable (parity) identified by a feature importance analysis (Fig. 1).

Predictor set 4 was the same as the third set except BMI was replaced with its derivative variables (height and weight) again because of the literature suggested that predictive performance improves using derivative instead of combined variables.

The four predicator sets are summarised in Table 1.

2.3. Data processing and missing values

All data processing and analysis were conducted using Python programming language. To assess the multicollinearity between predictors, Pearson correlation coefficients were calculated. Since no features had correlation coefficients greater than 0.75, all predictors were included in the model development. Missingness in the dataset was as follows - weight: n = 202 (0 41%); height: n = 196 (0 40%), BMI: n = 246 (0 50%) and maternal age: n = 1 (0 002%). Although the proportion of missing data was small, we performed multiple imputations to retain the statistical power instead of conducting complete case analyses.

2.4. Normalization and standardization

Since the scales used for each variable differ, regardless of its predictive power, the variable having the highest magnitude will dominate in the modelling. To eliminate the effect of having different dimensions or dimensional units on the results of our modelling, we performed data standardization. We ensured comparability between predictors by setting them on the same scale. To achieve this, predictors were standardized using power transformer and standard scaler techniques to ensure normal distribution and centering around zero with a variance around zero.

2.5. Imbalanced data handling

In binary classification tasks, data imbalance is a relatively common challenge [38]. In this study, we aimed to classify those who developed

Table 3
A brief summary of the ML algorithms used for model development and validation.

Logistic regression (logit model): *a parametric statistical model which mainly estimates the probability of an event based on predictors. It is mainly used for predictive analytics and classification. Mathematically, the probability that the outcome will be developed or not will be derived by multiplying each predictor by the corresponding numeric parameter and summing up all results via a logit/logistic function.*[67]
K-Nearest Neighbors (KNN): *This is a non-parametric, "non-generalizing learning" or "lazy learning" algorithm that predicts the correct class of the data by calculating the distance between the training points and the test data. To enhance the numerical stability of the model, predictors should be transformed by scaling.*[68]
Gaussian Naïve Bayes (GNB): *A powerful and fast probabilistic supervised ML algorithm based on the Bayes theorem, which assumes independence among features.*[69]
Support vector machine classifier (SVC): *A supervised ML algorithm based on kernel tricks that work by finding the line with the maximum margin, which separates different classes lying on either side.*[70] *It is memory intensive and slow, especially in extensive data and the presence of noise, but it is effective in high dimensional spaces.*
A multi-layer perceptron (MLP): *This is type of feedforward artificial neural network (ANN), that consists of multiple layers of interconnected nodes, called neurons. Contains an input layer (visible layer) where the input signal is received, hidden layers where signal processing is conducted, and an output layer where the prediction or decision about the input is made.*[71]
Decision Tree classifier (DTC): *This is a non-parametric supervised learning method that starts with a root node that grows into tree-like structure branches, which is simple to understand and easy for decision-making. A decision tree represents test feature at each internal node, test outcome at each branch, and class label at the leaf node (terminal node).*[72]
Random forest classifier (RFC): *Parallel ensemble learning method that creates a set of decision trees from a random subset of features, with or without bootstrapping, from which it selects for final prediction. This is a bagging-based algorithm.*[73]
ExtraTrees Classifier (ETC) (also called Extreme randomized tree): *Ensemble tree-based ML algorithm which works by aggregating the predictive results of numerous decision trees to output a prediction. The model prediction is by majority voting of decision trees. Contrary to the random forest, it uses original data, reduces bias, and variance over or underfitting is less likely.*[74]
AdaBoost Classifier (Adaptive Boosting, discrete AdaBoost): *a sequential ensemble modeling technique that changes several weak learners (decision trees) to strong learners by continuously rectifying the error in the preceding Model until the error is minimized or correctly predicted. The output of the other learning algorithms ('weak learners') is combined into a weighted sum representing the boosted classifier.*[75,76]
Gradient Boosting Classifier (GBC): *An ensemble boosting technique working in a stage-wise fashion on weak decision tree models to produce highly effective prediction models, which usually perform better than random forests.*[77]
CatBoost Classifier (Categorical Boosting; CBC): *A member of Gradient Boosted Decision Trees ensemble ML techniques that uses gradient boosting on decision trees.*[78,79] *This is a powerful algorithm that can handle categorical data effectively (without one-hot encoding) and having inbuilt innovative missing values handling system.*
XGBoost classifier (eXtreme Gradient Boosting; XGB): *Highly efficient, flexible, and accurate ensemble learning algorithm that uses a decision tree as a base learner. About ten times faster than traditional gradient boosting techniques, this algorithm has an excellent inbuilt split-finding method to improve trees and an inbuilt regularization method to reduce overfitting. This is a boosting-based algorithm outperforms other ML algorithms in tabular data and is considered the most evolved mother of all tree-based algorithms, mainly for tabular data.*[73,75,80]

GDM (positive class) against those who remained GDM-free (negative class). However, the ratio between the two (10341(GDM) / 38161(GDM free)) was skewed, demonstrating substantial class imbalance. To account for class imbalance, we used an up-sampling method called Synthetic Minority Over-Sampling Technique (SMOTE) [39] producing synthetic samples from the minority class.

2.6. Model development

Using the four predictor sets and 11 ML techniques alongside traditional logistic regression, we develop 48 models. Table 3 describes the 11 ML techniques and logistic regression. We used 80% of the available large sample size (development dataset) to create the models. To maximize performance without overfitting, hyperparameters were selected through tuning using the Grid Search method.

Applying the recent guidance in calculating the sample size required for developing a clinical prediction model outlined in the BMJ 2020 paper [40] shows that with: a baseline prevalence of GDM of 0.21 (as in our sample); six to nine predictor variables as in our models; then a sample size of over n = 30,000 was sufficiently large to target a mean absolute error of 0.004 between observed and true outcome probabilities.

2.7. Internal validation

Two internal validations were conducted. (1) The remaining 20% of the data set (validation dataset) were used to internally validate the performance of the model. (2) Five-fold cross-validation was carried out on both the development and validation datasets to evaluate the performance of the models.

2.8. Predictive performance and optimal model

From the 48 models developed, we determined the optimal model as the one with the best predictive performance by examining these model discrimination and calibration performance measures: Area Under the Curve (AUC), recall, precision, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-score, and Brier score. The Brier score measures the average squared difference between the predicted probability of GDM and the actual outcome (GDM (1) or No GDM (0)) across the participants. The perfect Brier score (0) indicates that there is a perfect match between predicted probabilities and actual outcome.

We also plot the calibration curve for each model, which consists of a diagonal line representing perfect calibration and a curve representing the observed relationship between predicted and actual probabilities. A well-calibrated model will have a curve that closely follows the diagonal line, indicating that the predicted probabilities are accurate and reliable.

Precision-Recall (PR) curve is a plot of precision-recall pairs for different classification thresholds, and we also examine this performance output measure for each model.

To check for overfitting, all performance metrics were evaluated on both the development and validation datasets.

Finally, clinical utility of the optimal model(s) are examined using decision curve analysis [41]. In decision curve analysis the net benefit of the prediction models were compared against the two default policies "treat all" and "treat none". Net benefit = (TP- wFP)/N [42]; where TP stands for the number of true positive, FP stands for the number of false positive w is a weight equal to odds of threshold, and N for total sample size, and "treat none" is a horizontal line with net benefit of zero.

3. Results

3.1. Baseline data

A total of 48,502 singleton pregnancies captured in Monash Health network of maternity hospitals from January 2016 to June 2021. The incidence of GDM was 21.3%. There was a higher incidence of GDM among women of older age, with a family history of DM, a history of GDM, overweight BMI, and with previous poor obstetric outcomes (Table 2). The feature selection model identified the top predictors of GDM, as indicated by the feature importance plot (Fig. 1). These predictors included a history of GDM, BMI, ethnicity, age, parity, height, family history of diabetes, weight, marital status, and poor obstetric history. The additional recognised factors were integrated into the new predictor sets (2–4) tested here.

3.2. Predictive performance of the models

Discrimination and calibration metrics for each model is presented in Table 4. In terms of discrimination, models developed using eXtreme Gradient Boosting (XGB) and CatBoost Classifier (CBC) were the best performing (AUC: predictor set 1 (79%) and (79%), predictor set 2 (91%) and (91%), predictor set 3 (92%) and (92%), predictor set 4

Table 4
5-fold cross validated performance measures of each algorithms across the four predictor sets.

Algorithm	Metrics	Model 1		Model 2		Model 3		Model 4	
		Training	Testing	Training	Testing	Training	Testing	Training	Testing
Logistic regression	Accuracy	0 64	0 64	0 68	0 67	0 68	0 68	0 68	0 68
	AUC	0 69(0 68-0 70)	0 69(0 68,0 70)	0 74(0 73,0 74)	0 74(0 73,0 75)	0 75(0 74,0 75)	0 75(0 74,0 76)	0 75(0 74,0 75)	0 75(0 74,0 76)
	Precision	0 65	0 65	0 68	0 68	0 68	0 69	0 69	0 69
	Recall	0 59	0 61	0 65	0 66	0 65	0 66	0 65	0 64
	F1-score	0 62	0 63	0 67	0 67	0 67	0 67	0 67	0 67
	Sensitivity	0 63	0 63	0 67	0 67	0 67	0 67	0 67	0 66
	Specificity	0 65	0 65	0 68	0 68	0 68	0 68	0 69	0 70
	PPV	0 68	0 68	0 70	0 69	0 70	0 70	0 70	0 71
	NPV	0 59	0 60	0 65	0 65	0 65	0 65	0 65	0 64
	Brier Score	0 27	0 30	0 29	0 29	0 29	0 30	0 29	0 29
K Neighbors Classifier	Accuracy	0 60	0 64	0 70	0 75	0 74	0 77	0 80	0 80
	AUC	0 69(0 68,0 70)	0 73(0 72,0 74)	0 78(0 77,0 79)	0 81(0 80,0 82)	0 81(0 80,0 81)	0 82(0 81,0 83)	0 86(0 86,0 87)	0 87(0 86,0 88)
	Precision	0 82	0 83	0 81	0 79	0 79	0 76	0 75	0 75
	Recall	0 25	0 35	0 53	0 68	0 66	0 78	0 90	0 91
	F1-score	0 38	0 49	0 64	0 73	0 72	0 77	0 82	0 82
	Sensitivity	0 56	0 59	0 65	0 72	0 71	0 78	0 87	0 88
	Specificity	0 82	0 83	0 81	0 79	0 79	0 76	0 75	0 75
	PPV	0 95	0 93	0 87	0 82	0 82	0 76	0 70	0 69
	NPV	0 24	0 34	0 52	0 67	0 66	0 78	0 90	0 91
	Brier Score	0 14	0 13	0 32	0 20	0 40	0 23	0 50	0 30
Support vector classifier	Accuracy	0 67	0 67	0 67	0 67	0 68	0 68	0 69	0 68
	AUC	0 72(0 71,0 73)	0 73(0 72,0 74)	0 73(0 72,0 73)	0 73(0 72,0 74)	0 74(0 73,0 75)	0 74(0 73,0 75)	0 75(0 74,0 75)	0 75(0 74,0 76)
	Precision	0 68	0 68	0 68	0 68	0 69	0 68	0 69	0 68
	Recall	0 65	0 66	0 65	0 66	0 67	0 66	0 68	0 67
	F1-score	0 66	0 67	0 67	0 67	0 68	0 67	0 69	0 67
	Sensitivity	0 66	0 67	0 66	0 67	0 68	0 67	0 68	0 67
	Specificity	0 67	0 68	0 68	0 68	0 69	0 68	0 69	0 68
	PPV	0 69	0 68	0 69	0 69	0 70	0 69	0 70	0 69
	NPV	0 65	0 66	0 65	0 65	0 66	0 66	0 68	0 66
	Brier Score	0 25	0 29	0 29	0 29	0 29	0 29	0 29	0 29
gaussian Naive Bayes	Accuracy	0 62	0 63	0 60	0 64	0 60	0 64	0 59	0 64
	AUC	0 71(0 70,0 72)	0 72(0 71,0 73)	0 71(0 71,0 72)	0 71(0 70,0 72)	0 71(0 71,0 72)	0 71(0 70,0 72)	0 71(0 70,0 72)	0 71(0 70,0 72)
	Precision	0 77	0 75	0 76	0 69	0 76	0 69	0 76	0 69
	Recall	0 34	0 38	0 28	0 52	0 28	0 51	0 27	0 52
	F1-score	0 47	0 51	0 41	0 59	0 41	0 59	0 40	0 59
	Sensitivity	0 58	0 58	0 56	0 61	0 56	0 61	0 56	0 61
	Specificity	0 77	0 75	0 76	0 69	0 76	0 69	0 76	0 69
	PPV	0 90	0 87	0 91	0 77	0 91	0 77	0 91	0 76
	NPV	0 34	0 38	0 28	0 52	0 28	0 51	0 27	0 52
	Brier Score	0 21	0 21	0 20	0 20	0 20	0 20	0 20	0 20
Decision tree classifier	Accuracy	0 68	0 70	0 75	0 79	0 79	0 79	0 78	0 76
	AUC	0 75(0 74,0 76)	0 78(0 77,0 79)	0 82(0 82,0 83)	0 82(0 82,0 83)	0 84(0 83,0 84)	0 82(0 81,0 83)	0 79(0 79,0 80)	84 28
	Precision	0 68	0 71	0 78	0 83	0 82	0 82	0 79	0 78
	Recall	0 67	0 67	0 71	0 72	0 73	0 74	0 75	0 74
	F1-score	0 68	0 69	0 74	0 77	0 78	0 78	0 77	0 76
	Sensitivity	0 68	0 68	0 73	0 75	0 76	0 77	0 76	0 75
	Specificity	0 68	0 71	0 78	0 83	0 82	0 82	0 79	0 77
	PPV	0 69	0 73	0 80	0 85	0 84	0 84	0 80	0 78
	NPV	0 67	0 67	0 71	0 72	0 73	0 74	0 75	0 74
	Brier Score	0 30	0 31	0 39	0 39	0 42	0 41	0 49	0 43
Random Forest Classifier	Accuracy	0 68	0 70	0 76	0 79	0 80	0 81	0 84	0 84
	AUC	0 75(0 74,0 76)	0 78(0 77,0 79)	0 84(0 84,0 85)	0 87(0 86,0 88)	0 88(0 87,0 88)	0 89(0 89,0 90)	0 92(0 91,0 92)	0 92(0 91,0 92)
	Precision	0 68	0 70	0 77	0 81	0 81	0 82	0 84	0 84
	Recall	0 68	0 69	0 74	0 76	0 78	0 80	0 84	0 84
	F1-score	0 68	0 70	0 75	0 78	0 79	0 81	0 84	0 84
	Sensitivity	0 68	0 69	0 75	0 77	0 78	0 81	0 83	0 84
	Specificity	0 68	0 70	0 76	0 81	0 80	0 82	0 84	0 84
	PPV	0 68	0 70	0 77	0 82	0 81	0 83	0 84	0 84
	NPV	0 68	0 69	0 74	0 76	0 78	0 80	0 83	0 84
	Brier Score	0 30	0 30	0 36	0 31	0 38	0 29	0 41	0 25
AdaBoost Classifier	Accuracy	0 68	0 70	0 74	0 80	0 78	0 82	0 81	0 82
	AUC	0 75(0 74,0 75)	0 78(0 77,0 79)	0 82(0 82,0 83)	0 89(0 88,0 89)	0 87(0 87,0 88)	0 91(0 90,0 91)	0 90(0 89,0 90)	0 91(0 90,0 91)
	Precision	0 69	0 72	0 75	0 83	0 81	0 87	0 86	0 88
	Recall	0 64	0 67	0 71	0 74	0 74	0 76	0 74	0 75
	F1-score	0 66	0 69	0 73	0 79	0 77	0 81	0 80	0 81
	Sensitivity	0 66	0 69	0 73	0 77	0 76	0 78	0 77	0 78

(*continued on next page*)

Table 4 (*continued*)

Algorithm	Metrics	Model 1		Model 2		Model 3		Model 4	
		Training	Testing	Training	Testing	Training	Testing	Training	Testing
	Specificity	0 69	0 71	0 75	0 83	0 81	0 87	0 86	0 88
	PPV	0 71	0 73	0 76	0 85	0 82	0 89	0 88	0 89
	NPV	0 64	0 67	0 71	0 74	0 74	0 76	0 74	0 75
	Brier Score	0 25	0 25	0 25	0 25	0 25	0 25	0 25	0 25
Gradient Boosting Classifier	Accuracy	0 68	0 71	0 77	0 81	0 81	0 82	0 84	0 77
	AUC	0 75(0 74,0 76)	0 79(0 78,0 80)	0 84(0 83,0 84)	0 86(0 85,0 87)	0 87(0 86,0 87)	0 87(0 87,0 88)	0 91(0 90,0 91)	0 92(0 91,0 93)
	Precision	0 69	0 72	0 79	0 84	0 83	0 85	0 85	0 83
	Recall	0 69	0 68	0 74	0 76	0 77	0 79	0 83	0 68
	F1-score	0 68	0 70	0 76	0 80	0 80	0 82	0 84	0 75
	Sensitivity	0 68	0 69	0 75	0 78	0 79	0 80	0 83	0 84
	Specificity	0 68	0 71	0 79	0 84	0 83	0 84	0 85	0 84
	PPV	0 69	0 73	0 80	0 85	0 84	0 85	0 85	0 84
	NPV	0 68	0 68	0 74	0 76	0 77	0 79	0 83	0 84
	Brier Score	0 30	0 32	0 37	0 39	0 43	0 37	0 50	0 40
XGBClassifier	Accuracy	0 68	0 71	0 77	0 82	0 81	0 84	0 84	0 85
	AUC	0 76(0 75,0 76)	0 79(0 78,0 80)	0 86(0 86,0 87)	0 91(0 90,0 91)	0 90(0 90,0 91)	0 92(0 91,0 93)	0 92(0 92,0 93)	0 92(0 92,0 93)
	Precision	0 68	0 71	0 79	0 87	0 86	0 90	0 92	0 91
	Recall	0 68	0 69	0 73	0 76	0 74	0 77	0 75	0 77
	F1-score	0 68	0 70	0 76	0 81	0 80	0 83	0 83	0 83
	Sensitivity	0 68	0 70	0 74	0 78	0 77	0 80	0 79	0 80
	Specificity	0 68	0 71	0 79	0 87	0 86	0 90	0 92	0 91
	PPV	0 69	0 72	0 80	0 89	0 88	0 91	0 93	0 92
	NPV	0 68	0 69	0 72	0 75	0 74	0 76	0 75	0 77
	Brier Score	0 30	0 33	0 39	0 39	0 43	0 40	0 49	0 39
ExtraTrees Classifier	Accuracy	0 68	0 70	0 71	0 74	0 72	0 74	0 72	0 75
	AUC	0 75(0 74,0 76)	0 78(0 77,0 79)	0 79(0 79,0 80)	0 83(0 82,0 84)	0 79(0 79,0 80)	0 83(0 82,0 83)	0 79(0 79,0 80)	0 83(0 82,0 84)
	Precision	0 68	0 71	0 72	0 75	0 73	0 75	0 73	0 75
	Recall	0 68	0 69	0 70	0 73	0 69	0 73	0 69	0 74
	F1-score	0 68	0 70	0 71	0 74	0 71	0 74	0 71	0 74
	Sensitivity	0 68	0 70	0 71	0 74	0 70	0 74	0 71	0 74
	Specificity	0 68	0 70	0 72	0 75	0 73	0 75	0 73	0 75
	PPV	0 68	0 71	0 72	0 75	0 74	0 75	0 74	0 75
	NPV	0 68	0 69	0 70	0 73	0 69	0 73	0 69	0 74
	Brier Score	0 30	0 30	0 30	0 28	0 29	0 28	0 29	0 27
CatBoost classifier	Accuracy	0 68	0 71	0 76	0 82	0 81	0 84	0 84	0 85
	AUC	0 76(0 75,0 76)	0 79(0 78,0 80)	0 86(0 85,0 86)	0 91(0 90,0 91)	0 90(0 89,0 90)	0 92(0 92,0 93)	0 92(0 92,0 93)	0 93(0 02,0 93)
	Precision	0 69	0 72	0 79	0 87	0 86	0 89	0 92	0 90
	Recall	0 68	0 69	0 71	0 76	0 74	0 78	0 75	0 78
	F1-score	0 68	0 70	0 75	0 81	0 79	0 83	0 82	0 84
	Sensitivity	0 68	0 70	0 74	0 79	0 77	0 80	0 79	0 81
	Specificity	0 68	0 71	0 78	0 87	0 85	0 89	0 91	0 90
	PPV	0 69	0 72	0 81	0 89	0 87	0 91	0 93	0 92
	NPV	0 68	0 69	0 71	0 76	0 74	0 78	0 75	0 78
	Brier Score	0 30	0 32	0 33	0 40	0 36	0 40	0 37	0 39
Multi-layer Perceptron Classifier	Accuracy	0 68	0 68	0 73	0 64	0 72	0 75	0 75	0 78
	AUC	0 74(0 74,0 75)	0 76(0 75,0 77)	0 77(0 76,0 78	0 79(0 78,0 80)	0 80(0 79,0 80)	0 83(0 82,0 83)	0 82(0 81,0 82)	0 86(0 85,0 87)
	Precision	0 68	0 69	0 72	0 65	0 71	0 74	0 73	0 76
	Recall	0 66	0 67	0 75	0 59	0 75	0 79	0 78	0 82
	F1-score	0 67	0 68	0 74	0 62	0 73	0 76	0 76	0 79
	Sensitivity	0 67	0 68	0 74	0 63	0 73	0 77	0 77	0 80
	Specificity	0 68	0 68	0 72	0 65	0 71	0 73	0 73	0 75
	PPV	0 69	0 69	0 72	0 68	0 70	0 71	0 71	0 73
	NPV	0 65	0 67	0 75	0 59	0 75	0 79	0 78	0 82
	Brier Score	0 31	0 31	0 32	0 28	0 36	0 29	0 37	0 27

Abbreviations: PPV: positive predictive value, NPV: negative predictive value, AUC: area under curve, **Model 1**: Age (categorical), BMI at booking (categorical), history of GDM, family history of diabetes, previous poor obstetric outcomes, and ethnicity; **Model 2**: Age (continuous), BMI at booking (continuous), history of GDM, family history of diabetes, previous poor obstetric outcomes, and ethnicity; **Model 3**: Age (continuous), BMI at booking (continuous), history of GDM, family history of diabetes, previous poor obstetric outcomes, ethnicity, and parity; **Model 4**: Age (continuous), Weight at booking (continuous), height at booking (continuous), history of GDM, family history of diabetes, previous poor obstetric outcomes, ethnicity, and parity

(92%) and (93%), respectively), and also had higher accuracy, sensitivity, specificity, NPV, and PPV. In addition, models created by applying ensemble methods including Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier outperformed other models. The least performing ML technique across all models was Gaussian Naive Bayes Classifier (GNB) (AUC: predictor set 1 (72%), predictor set 2 (71%), predictor set 3 (71%), predictor set 4 (71%)). GNB, Support Vector Machine, and logistic regression were with the least performing techniques across all predictor sets. In general, performance increased from predictor set 1 to 4, indicating predictor set 4 as the best-performing set of predictors. ML techniques performed better than logistic regression. Fig. 2 shows the comparative visualization of ROC curve each model. Predictor sets 3 and 4 had higher AUCs than predictor sets 1 and 2.

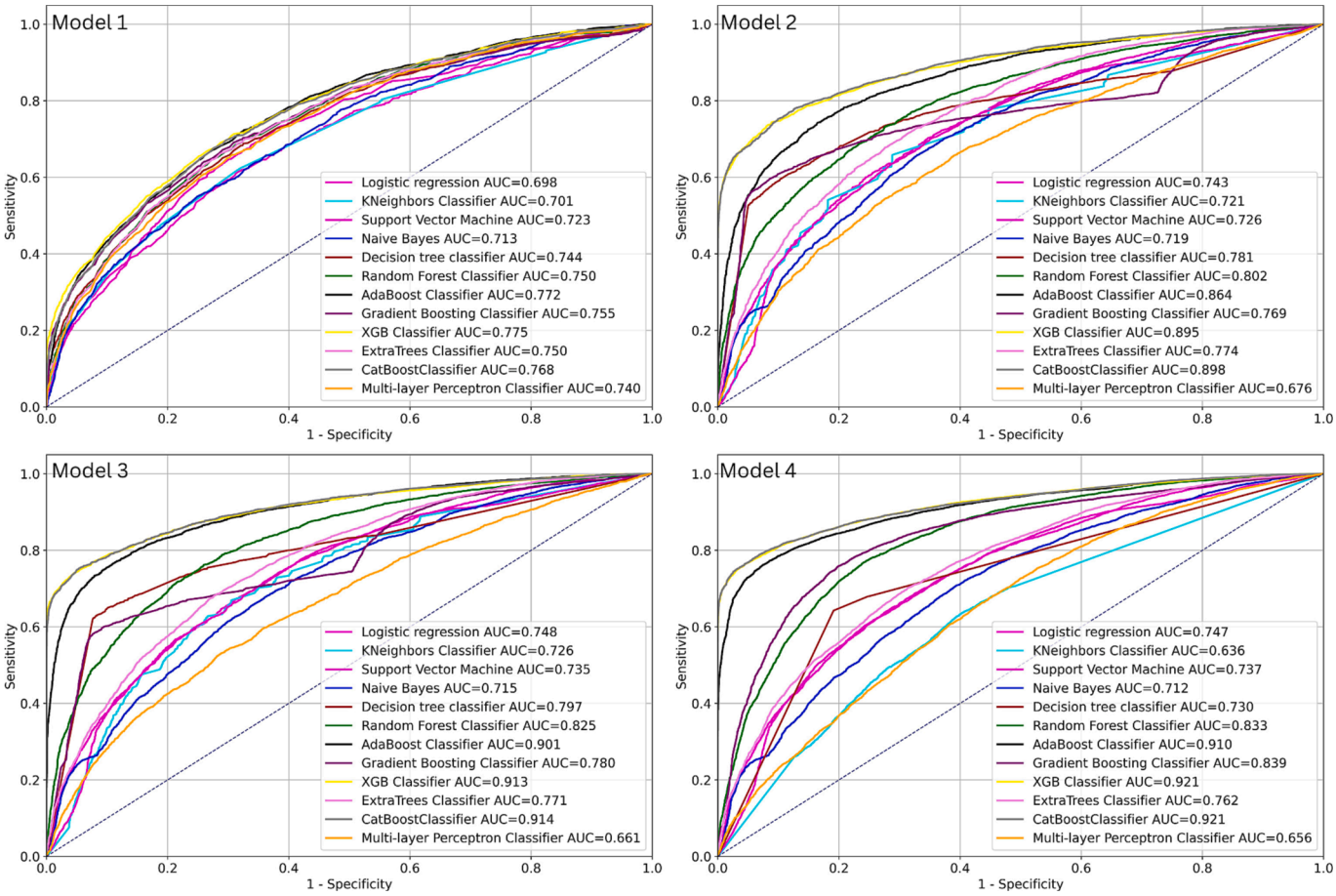


Fig. 2. Area under curve of receiver operative characteristics of 11 ML algorithms and logistic regression across four models Footnotes for Fig. 2: AUC: Area under curve, XGB: Extreme Gradient Boosting.

The Brier score showed that the calibration of most models ranged from 0.13 to 0.40, where 1.0 means the worst calibration (Table 4). In addition to the Brier score, the calibration curve plots are presented in Fig. 3. The calibration plot of two optimal algorithms (CBC and XGB) closely follows the diagonal line in three models indicating that they are well calibrated.

The performance metrics for almost all models in the development and validation datasets show minimal differences, suggesting that the models are not overfitting and are capable of generalizing to new, unseen data. Table 4 represents the discrimination performance of all four models. The Precision-Recall (PR) curve analysis plot is presented in Fig. 4 and shows that Area Under the Precision-Recall Curve of the two optimal algorithms (CBC and XGB) is excellent.

3.3. Clinical utility of the optimal models

The decision curve analysis [42] was performed using all predictor sets 4 for the two optimal ML techniques identified, namely XGB and CBC (Fig. 5) and (Fig. 6). The results indicate that stratifying pregnant women using a model developed using either of these ML techniques provides more benefits compared to managing all or managing none strategies, particularly over the threshold probabilities shown on the decision curve analysis graph. For example, for the model created suing predictor set 4 and CBC GDM prediction model outperforms either of the two decision strategies across the following approximated threshold probabilities (12.5, 96.0).

4. Discussion

We explored multiple ML-based techniques across a range of predictor sets (48 models) and compared their predictive performance to classical regression models, using a large dataset of routinely collected data from an ethnically diverse population. Building on validated traditional statistical models, we have demonstrated that overall, ML techniques achieved the best predictive performance. Applying these ML techniques across various predictor sets, informed by feature importance analysis, the best predictive performance was achieved by ML boosting algorithms (CBC and XGB).

ML-based prediction approaches for GDM are frequently published, but they often suffer from methodological limitations that compromise their quality and reliability. Some of these are developed using a small sample size, contrary to the requirements of ML techniques. In addition, there are issues with inappropriate inclusion and exclusion criteria, a lack of discrimination and calibration assessment reports, and most importantly, a failure to evaluate the clinical utility of the resultant prediction models [17,43]. Considering the limitation of available ML based GDM prediction models, here the best prediction technique was determined and its performance is assessed comprehensively utilizing multiple ML techniques, across various combinations of predictors selected by available evidence, clinical judgment, and ML-based feature importance analysis. The net benefit of the optimal model was also assessed by decision curve analysis across various probability thresholds. To our knowledge, this approach of comparing multiple models including ML derived models in this way is novel.

Here, we have shown that boosting algorithms, tree-based algorithms, and neural network classification-based ML techniques led to

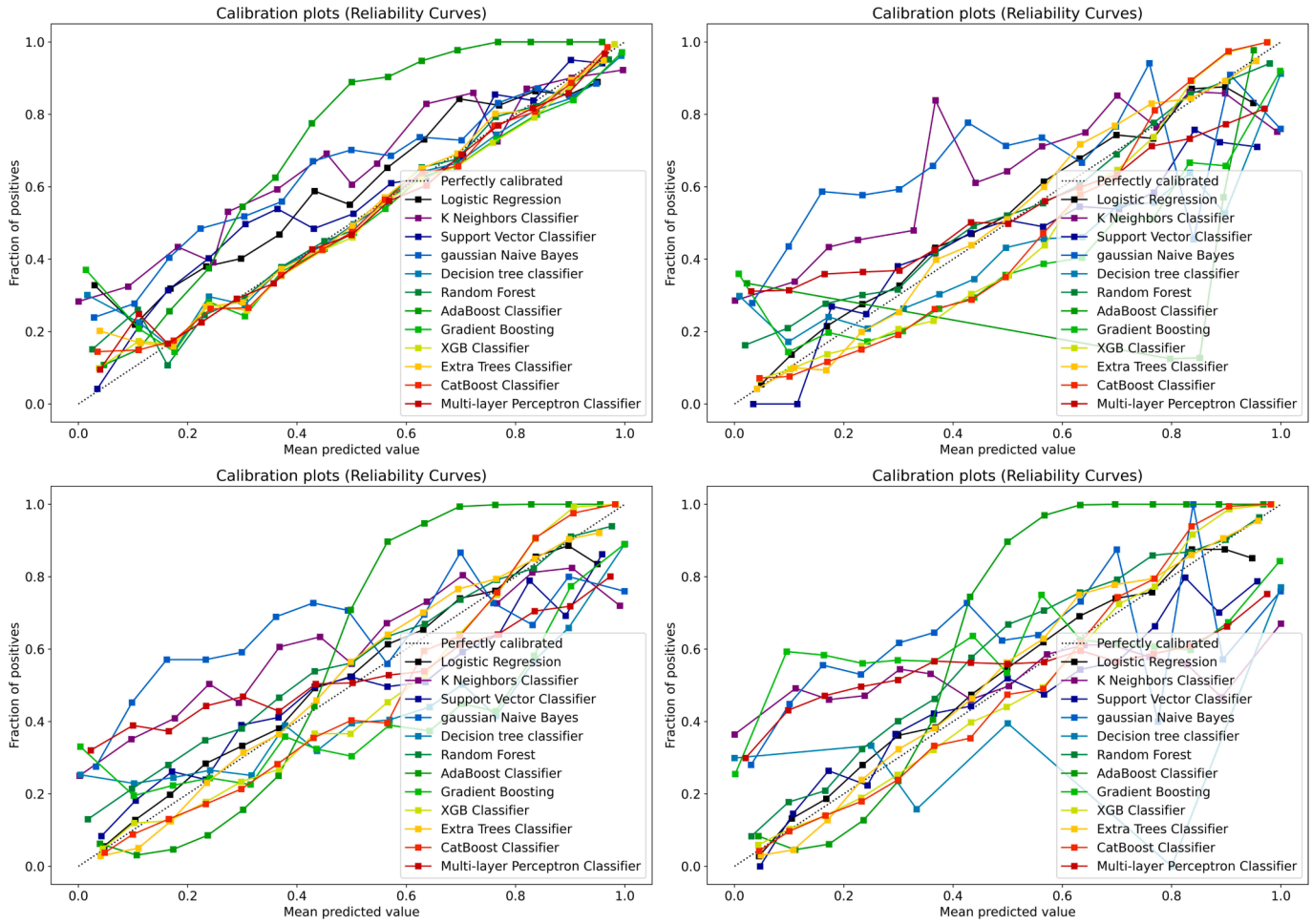


Fig. 3. Calibration curve of 11 Machine Learning algorithms and logistic regression across four models Footnotes for Fig. 3: XGB: Extreme Gradient Boosting.

better discrimination performance in predicting GDM. Boosting ML techniques led to the best accuracy. Other ensemble techniques also led to good predictive abilities, due to the capability to learn from complex and non-linear associations of predictors in a real-world dataset [44]. The predictive performance of GNB ML techniques was the poorest. This could be attributed to its strong assumption of predictor independence, which is often challenging to meet with real-world data [45]. The transportability of the developed optimal techniques and model now needs to be assessed in different geographical populations.

In terms of predictor sets, although categorizing continuous variables are common in clinical practice, this practice is not recommended statistically, due to the limitation on predictive power [42,46]. However, the added predictive value of using continuous variables has not been commonly assessed for GDM prediction. Building on the original Monash GDM Model (with categorical variables; predictor set 1), predictor set 2 included BMI and age as continuous variables, improving the predictive performance of almost all generated models. Feature importance and correlation analyses, identified parity and was added into predictor set 3, which further improved predictive performance.

Although using BMI as a derived variable in predictive models can offer simplicity, normalization, and some clinical relevance, it can also result in loss of information, lack of accuracy, and bias [47,48]. Feature importance and correlation analyses, identified both height and weight as important predictors. Additionally, instead of BMI both weight and height were also identified as important predictors; therefore, a fourth predictor set was developed by including height and weight as predictors instead of BMI, which has also been done in other GDM prediction model studies [49–52]. Utilizing these source derivative variables as continuous variables, further improved prediction model performance. Of four developed predictor sets, the fourth was identified as the optimal set with dramatic enhancement in predictive performances.

Decision curve analysis showed the optimal model identified a threshold probability over which the prediction models are recommended. For transportability purpose, it is vital to test the applicability of the developed optimal models in different healthcare setup by other researchers. After external validation, it can help to stratify high risk women allowing for early intervention deterring possible obstetric complications. By identifying women at high risk of developing GDM, healthcare resources can also be targeted to those who are most likely to benefit from early intervention. With additional external validation, the Monash GDM ML Model will be incorporated into our online risk prediction tools for future use.

4.1. Practical implications

Currently prediction tools are used in clinical practice including for preeclampsia and GDM. In GDM they are used clinically to identify who to target to have both glucose tolerance tests and effective prevention strategies to the highest risk groups [26]. The developed GDM risk prediction models can help health care professionals to identify a woman who are at high risk of developing GDM in early pregnancy allowing for timely interventions and initiate prevention measures. Women who identified as high risk can receive more frequent monitoring ensuring timely management if GDM develop. Furthermore, early risk assessment can enable personalized targeted nutritional and

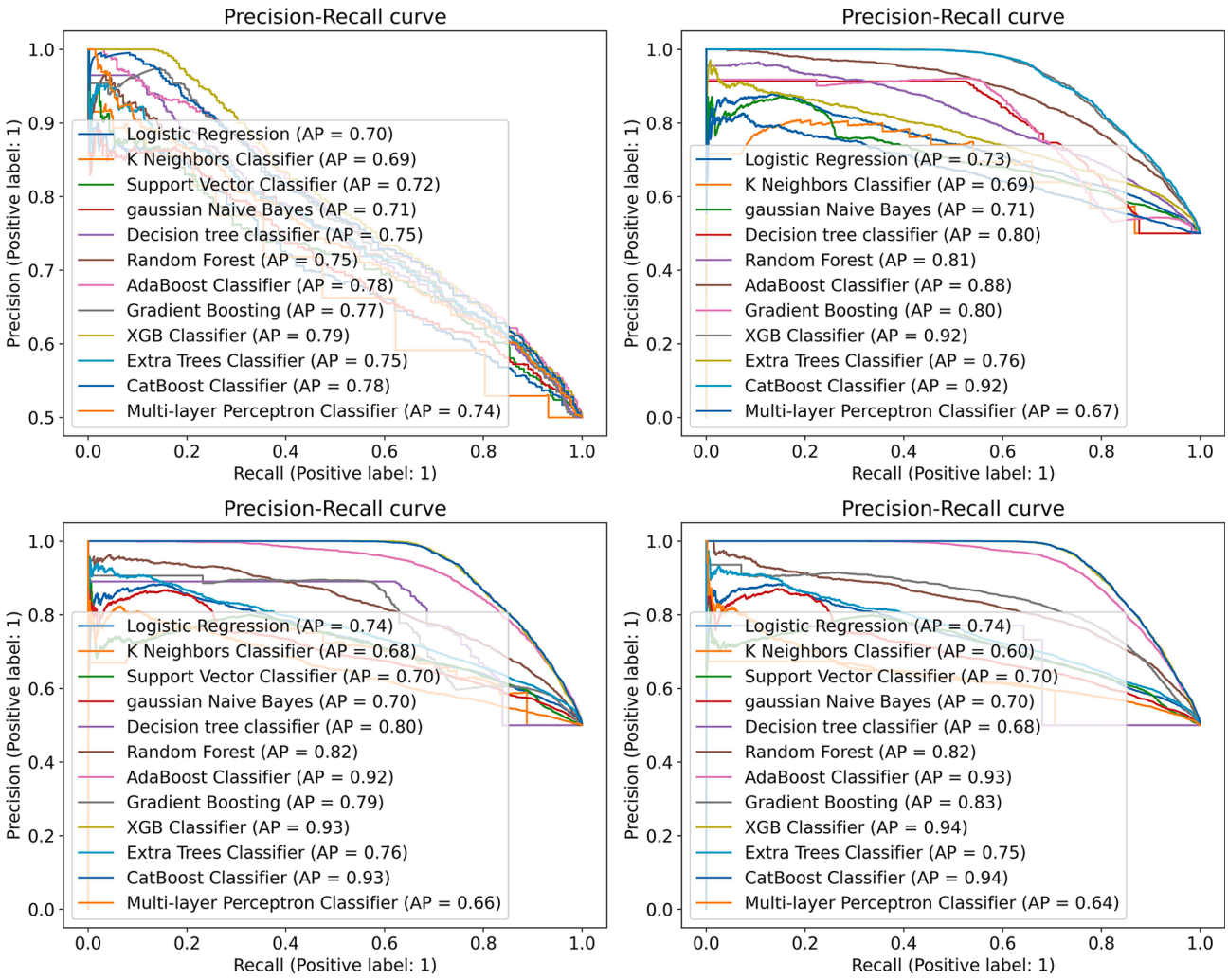


Fig. 4. Precision-recall-curve of 11 Machine Learning algorithms and logistic regression across four models Footnotes for Fig. 4: XGB: Extreme Gradient Boosting, AP: Average Precision,

lifestyle counselling which can help to prevent or delay the onset of GDM. Health care system can more efficiently allocate resources such as specialized clinics or counselling sessions to those who are identified as higher risk by the prediction models. Understanding their risks allows pregnant women to make better informed decision regarding their health, diet and physical activity. It will also enhance also research opportunities by aiding in stratifying patients for clinical studies. Overall, timely prediction and intervention can lead to better health outcomes for mother and baby.

In the context of validating and implementing clinical prediction models, our team will continue to co-developed online digital tools with clinicians and stakeholders, using codesign, as we have done previously (https://www.personalgdm.com/). Acceptability, feasibility, and refinement occur in codesign cycles, to ensure accurate communication and perception of risk such as in the Personal GDM tool (https://www.personalgdm.com/) with more coming online shortly.

4.2. Strength and limitations

This is the first study that fitted GDM prediction models and sought to identify the optimal predictive performance model by examining different modelling methods using 4 predictor sets; 11 ML techniques, and; traditional logistic regression. A seemingly limitation was that the data is from one country with universally accessible healthcare, however, it is actually a strength as the dataset was from one of the most diverse multicultural populations worldwide with about 60% born overseas and large representations from over diverse ethnic categories [53]. This model is now being tested in international datasets, including from low- and middle-income countries to provide external validation. The lack of access to other innovative markers including laboratory tests, may have limited the Models generated, however the overall predictive performance generated here far exceeds past models and is rated as excellent.

4.3. Conclusions

This important and novel study has demonstrated that ML techniques perform better than traditional statistical regression models, in the prediction of risk for GDM, a common condition with major opportunities for targeted prevention. Here, we have developed a simple and accurate risk prediction model, building on prior internationally validated models, optimising included variables and exploring novel ML techniques. We have updated the original internationally validated Monash GDM Risk Prediction Model[21] generating a highly significant improvement in predictive performance as highlighted by the increase in AUC from 0.70 to 0.93, and better accuracy and clinical utility, whilst retaining simple and accessible predictor variables (predictor set 4). This Monash Machine Learning GDM Risk Prediction Model (predictor set 4 with XGB or CBC) is now being validated in international datasets and will ultimately power accessible online risk prediction tools to enable

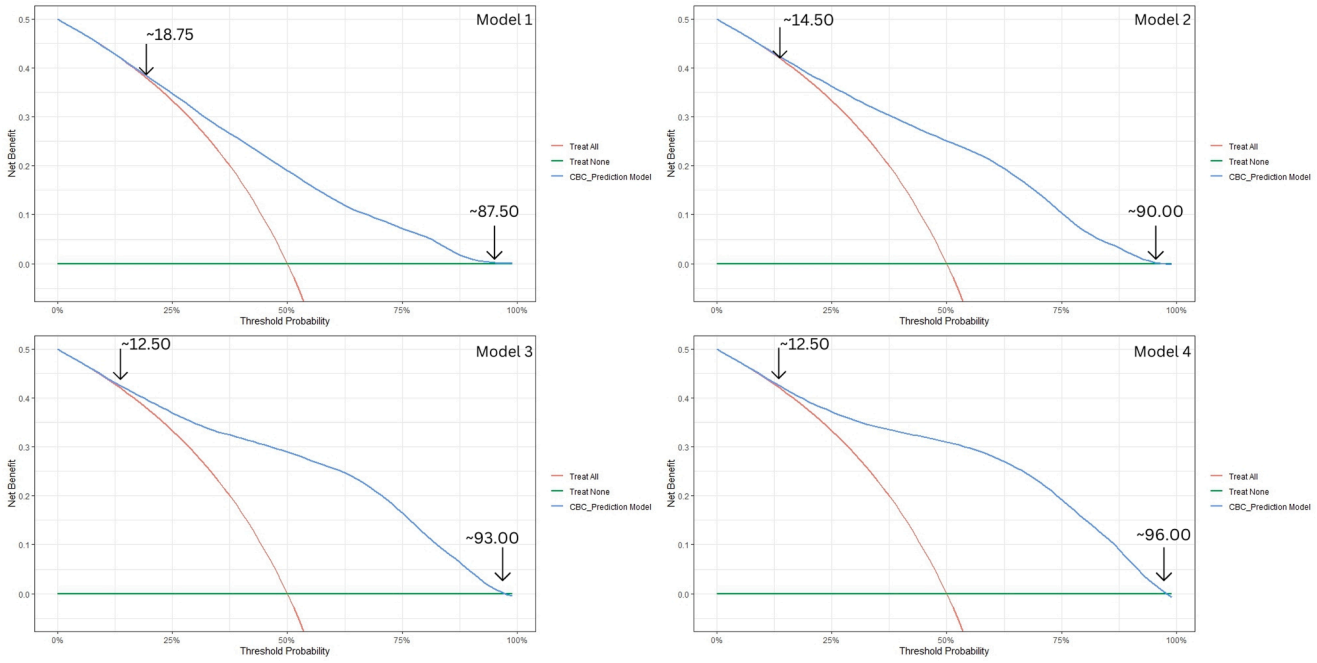


Fig. 5. Decision curve analysis of CatBoost classifier model.

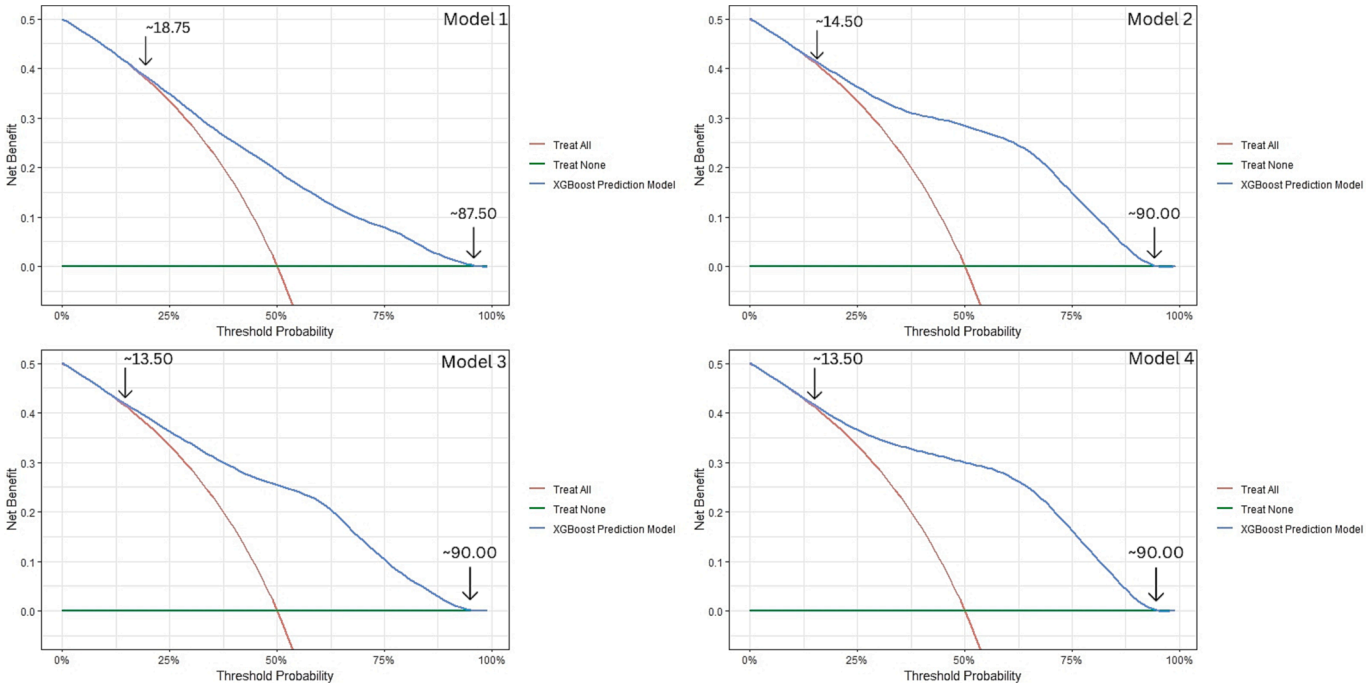


Fig. 6. Decision curve analysis of XGBoost classifier model.

implementation into clinical care.

5. Declarations

5.1. Author contributions

Yitayeh Mengistu, Joanne Enticott, Helena Teede, and Kushan De Silva were involved in the acquisition of data, conception, design, or planning of the study; and drafting of the manuscript. All the remaining authors were involved in the interpretation of the results, critically reviewing or revising the manuscript for important intellectual content.

5.2. Funding

AM and HT are supported by fellowships from the National Health and Medical Research Council (NHMRC) of Australia.

5.3. Availability of data and materials

All data and material access requests can be forwarded to the corresponding author email address.

5.4. Ethics approval

Ethical permission was obtained from Monash health. The research has been conducted in adherence with the Code of Ethics of the World Medical Association, also known as the Declaration of Helsinki.

6. Summary table

6.1. Key findings

Question: Can a machine learning model perform better than traditional logistic regression to accurately predict the onset of Gestational Diabetes Mellitus (GDM)?

Findings: Building on validated traditional statistical models, we have demonstrated that overall, ML methods achieved the best predictive performance.

Meaning: The Monash GDM Machine Learning Model, created using routinely available data, accurately predicted GDM with substantially improved accuracy. Prediction of those at risk can facilitate targeted prevention of this common condition.

CRediT authorship contribution statement

Yitayeh Belsti: Conceptualization, Formal analysis, Data curation, Methodology, Software, Visualization, Writing – original draft. **Lisa Moran:** Supervision, Writing – review & editing. **Lan Du:** Supervision, Writing – review & editing. **Aya Mousa:** Supervision, Writing – review & editing. **Kushan de Silva:** Conceptualization, Formal analysis, Data curation, Methodology, Software, Writing – original draft. **Joanne Enticott:** Conceptualization, Formal analysis, Data curation, Methodology, Software, Supervision, Writing – original draft. **Helena Teede:** Conceptualization, Formal analysis, Data curation, Methodology, Software, Supervision, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105228.

References

[1] Gestational diabetes mellitus (GDM) [internet]; 2019 [cited Nov 29 2022]. Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/gestational-diabetes.

[2] B.E. Metzger, D.R. Coustan, Organizing Committee. Summary and recommendations of the fourth international workshop-conference on gestational diabetes mellitus, Diabetes Care 1 (21) (1998 Aug) B161.

[3] H. Wang, N. Li, T. Chivese, M. Werfalli, H. Sun, L. Yuen, et al., IDF diabetes Atlas: estimation of global and regional gestational diabetes mellitus prevalence for 2021 by international association of diabetes in pregnancy study Group's criteria, Diabetes Res. Clin. Pract. 183 (2022 Jan), 109050, https://doi.org/10.1016/j.diabres.2021.109050, PMID 34883186.

[4] K. Ogurtsova, J.D. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N.H. Cho, et al., IDF Diabetes Atlas: global estimates for the prevalence of diabetes for 2015 and 2040, Diabetes Res. Clin. Pract. 1 (128) (2017 Jun) 40–50, https://doi.org/10.1016/j.diabres.2017.03.024, PMID 28437734.

[5] N.H. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A. W. Ohlrogge, et al., IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045, Diabetes Res. Clin. Pract. 1 (138) (2018 Apr) 271–281, https://doi.org/10.1016/j.diabres.2018.02.023, PMID 29496507.

[6] D.R. Whiting, L. Guariguata, C. Weil, J. Shaw, IDF Diabetes Atlas: global estimates of the prevalence of diabetes for 2011 and 2030, Diabetes Res. Clin. Pract. 94 (3) (2011 Dec 1) 311–321, https://doi.org/10.1016/j.diabres.2011.10.029, PMID 22079683.

[7] American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. Diabetes care. 2018 Jan 1;41 (Supplement_1):S13-27.

[8] H.D. McIntyre, P. Catalano, C. Zhang, G. Desoye, E.R. Mathiesen, P. Damm, Gestational diabetes mellitus, Nat. Rev. Dis. Primers 5 (1) (2019 Jul 11) 47, https://doi.org/10.1038/s41572-019-0098-8, PMID 31296866.

[9] Zhang C, Ning Y. Effect of dietary and lifestyle factors on the risk of gestational diabetes: review of epidemiologic evidence. The American journal of clinical nutrition. 2011 Dec 1;94(suppl_6):1975S-9S.

[10] C.G. Solomon, W.C. Willett, V.J. Carey, J. Rich-Edwards, D.J. Hunter, G.A. Colditz, M.J. Stampfer, F.E. Speizer, D. Spiegelman, J.E. Manson, A prospective study of pregravid determinants of gestational diabetes mellitus, J. Am. Med. Assoc. 278 (13) (1997 Oct 1) 1078–1083.

[11] R.F. Goldstein, S.K. Abell, S. Ranasinha, M. Misso, J.A. Boyle, M.H. Black, et al., Association of gestational weight gain with maternal and infant outcomes: A systematic review and meta-analysis, J. Am. Med. Assoc. 317 (21) (2017 Jun 6) 2207–2225, https://doi.org/10.1001/jama.2017.3635, PMID 28586887.

[12] A. Bener, N.M. Saleh, A. Al-Hamaq, Prevalence of gestational diabetes and associated maternal and neonatal complications in a fast-developing community: global comparisons, Int. J. Womens Health 7 (3) (2011 Nov) 367–373, https://doi.org/10.2147/IJWH.S26094, PMID 22140323.

[13] E.C. Johns, F.C. Denison, J.E. Norman, R.M. Reynolds, Gestational diabetes mellitus: mechanisms, treatment, and complications, Trends Endocrinol Metab 29 (11) (2018 Nov 1) 743–754, https://doi.org/10.1016/j.tem.2018.09.004, PMID 30297319.

[14] H.J. Teede, C. Bailey, L.J. Moran, M. Bahri Khomami, J. Enticott, S. Ranasinha, et al., Association of antenatal diet and physical activity-based interventions with gestational weight gain and pregnancy outcomes: A systematic review and meta-analysis, JAMA Intern. Med. 182 (2) (2022 Feb 1) 106–114, https://doi.org/10.1001/jamainternmed.2021.6373, PMID 34928300.

[15] K.G.M. Moons, P. Royston, Y. Vergouwe, D.E. Grobbee, D.G. Altman, Prognosis and prognostic research: what, why, and how? BMJ 23 (338) (2009 Feb), b375 https://doi.org/10.1136/bmj.b375, PMID 19237405.

[16] R.D. Riley, D. van der Windt, P. Croft, K.G. Moons, Prognosis research in healthcare: concepts, methods, and impact, Oxford University Press, 2019.

[17] D. Mennickent, A. Rodríguez, M. Farías-Jofré, J. Araya, E. Guzmán-Gutiérrez, Machine learning-based models for gestational diabetes mellitus prediction before 24–28 weeks of pregnancy: a review, Artif. Intell. Med. 132 (2022 Oct), 102378, https://doi.org/10.1016/j.artmed.2022.102378, PMID 36207076.

[18] K.A. Lamain, C.A. Naaktgeboren, A. Franx, K.G.M. Moons, M.P.H. Koster, Prediction models for the risk of gestational diabetes: a systematic review, Diagn Progn Res. (2017), https://doi.org/10.1186/s41512-016-0005-7. Feb 8;1(1):3.

[19] M.Z.I. Chowdhury, T.C. Turin, Precision health through prediction modelling: factors to consider before implementing a prediction model in clinical practice, J. Prim. Health Care 12 (1) (2020 Mar) 3–9, https://doi.org/10.1071/HC19087, PMID 32223844.

[20] N.J. Stone, J.G. Robinson, A.H. Lichtenstein, M.C.N. Bairey, C.B. Blum, R.H. Eckel, et al., ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults, J. Am. Coll. Cardiol. (2013;2014(Jul); 63(25_Part_B):2889–934.).

[21] H.J. Teede, C.L. Harrison, W.T. Teh, E. Paul, C.A. Allan, Gestational diabetes: development of an early risk prediction tool to facilitate opportunities for prevention, Aust. N. Z. J. Obstet. Gynaecol. 51 (6) (2011) 499–504, https://doi.org/10.1111/j.1479-828X.2011.01356.x, PMID 21951203.

[22] S. Thériault, J.C. Forest, J. Massé, Y. Giguère, Validation of early risk-prediction models for gestational diabetes based on clinical characteristics, Diabetes Res. Clin. Pract. 103 (3) (2014) 419–425, https://doi.org/10.1016/j.diabres.2013.12.009, PMID 24447804.

[23] F. van Hoorn, M.P.H. Koster, C.A. Naaktgeboren, F. Groenendaal, A. Kwee, M. Lamain-de Ruiter, et al., Prognostic models versus single risk factor approach in first-trimester selective screening for gestational diabetes mellitus: a prospective population-based multicentre cohort study, BJOG 128 (4) (2021) 645–654, https://doi.org/10.1111/1471-0528.16446, PMID 32757408.

[24] M.L. de Lamain-de Ruiter, A. Kwee, C.A. Naaktgeboren, I. de Groot, I.M. Evers, F. Groenendaal, et al., External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study, BMJ 30 (354) (2016 Aug), i4338, https://doi.org/10.1136/bmj.i4338, PMID 27576867.

[25] S.D. Cooray, K. De Silva, J. Enticott, S. Dawadi, J.A. Boyle, G. Soldatos, et al., External validation and updating of a prediction model for the diagnosis of gestational diabetes mellitus, medRxiv. (2021).

[26] S. Thangaratinam, S.D. Cooray, N. Sukumar, M.S.B. Huda, R. Devlieger, K. Benhalima, et al., ENDOCRINOLOGY IN THE TIME OF COVID-19: diagnosis and management of gestational diabetes mellitus, Eur. J. Endocrinol. 183 (2) (2020) G49–G56, https://doi.org/10.1530/EJE-20-0401, PMID 32454456.

[27] M. Awad, R. Machine Learning Khanna, Efficient learning machines: theories, concepts, and applications for engineers and system designers [internet]. In: Berkeley, CA: Apress. p. 1-18; 2015 Awad M, Khanna R, editors [cited Feb 23 2023]. 10.1007/978-1-4302-5990-9 1.

[28] TM. Machine learning , MacGraw-Hill , New York, 1997.

[29] J.A. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Method. 19 (1) (2019) 1.

[30] Q. Bi, K.E. Goodman, J. Kaminsky, J. Lessler, What is machine learning? A primer for the epidemiologist, Am. J. Epidemiol. 188 (12) (2019) 2222–2239, https://doi.org/10.1093/aje/kwz189, PMID 31509183.

[31] H. Sufriyana, A. Husnayain, Y.L. Chen, C.Y. Kuo, O. Singh, T.Y. Yeh, et al., Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: systematic review and meta-analysis, JMIR Med. Inform. 8 (11) (2020 Nov 17) e16503.
[32] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B. Van Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clin. Epidemiol. 110 (2019 Jun) 12–22, https://doi.org/10.1016/j.jclinepi.2019.02.004, PMID 30763612.
[33] Y. Ye, Y. Xiong, Q. Zhou, J. Wu, X. Li, X. Xiao, Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: A retrospective cohort study, J. Diabetes Res. 12 (2020) (2020 Jun) 4168340, https://doi.org/10.1155/2020/4168340, PMID 32626780.
[34] ADIPS GDM Guidelines V18.11.2014_000.pdf [Internet] [cited Feb 23 2023]. Available from: https://www.adips.org/downloads/2014ADIPSGDMGuidelinesV18.11.2014_000.pdf.
[35] International Association of Diabetes and Pregnancy Study Groups Consensus Panel, Metzger BE, Gabbe SG, Persson B, Buchanan TA, Catalano PA et al. International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. Diabetes Care. 2010;33(3):676-82. 10.2337/dc09-1848, PMID 20190296.
[36] World Health Organization. Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy [internet]. World Health Organization; 2013 [cited Feb 23 2023]. Report No. : WHO/NMH/MND/13.2. Available from: https://apps.who.int/iris/handle/10665/85975.
[37] Australian standard classification of cultural and ethnic groups (ASCCEG). Australian Bureau of Statistics [internet]; 2019 [cited Feb 23 2023]. Available from: https://www.abs.gov.au/statistics/classifications/australian-standard-classification-cultural-and-ethnic-groups-ascceg/latest-release.
[38] H. Kaur, H.S. Pannu, A.K. Malhi, A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions, ACM Comput. Surv. 52 (4) (2020) 1–36, https://doi.org/10.1145/3343440.
[39] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16: 321-57.
[40] R.D. Riley, J. Ensor, K.I.E. Snell, F.E. Harrell, G.P. Martin, J.B. Reitsma, et al., Calculating the sample size required for developing a clinical prediction model, BMJ 18 (368) (2020 Mar), m441, https://doi.org/10.1136/bmj.m441, PMID 32188600.
[41] Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, Roobol MJ, Steyerberg EW. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol. 2018 Dec;74(6):796-804. 10.1016/j.eururo.2018.08.038. Epub 2018 Sep 19. PMID: 30241973; PMCID: PMC6261531.
[42] E.W. Steyerberg, Clinical prediction models: A practical approach to development, validation, and updating, 2nd ed 2019 ed., Springer, Cham, Switzerland, 2019, p. 591 p..
[43] Z. Zhang, L. Yang, W. Han, Y. Wu, L. Zhang, C. Gao, et al., Machine learning prediction models for gestational diabetes mellitus: meta-analysis, J. Med. Internet Res. 24 (3) (2022 Mar 16) e26634.
[44] D.D. Miller, E.W. Brown, Artificial intelligence in medical practice: the question to the answer? Am. J. Med. 131 (2) (2018) 129–133, https://doi.org/10.1016/j.amjmed.2017.10.035, PMID 29126825.
[45] I.H. Sarker, Machine learning: algorithms, real-world applications and research directions, SN Comput Sci. 2 (3) (2021 May) 160, https://doi.org/10.1007/s42979-021-00592-x, PMID 33778771.
[46] P. Royston, D.G. Altman, W. Sauerbrei, Dichotomizing continuous predictors in multiple regression: a bad idea, Stat. Med. 25 (1) (2006 Jan 15) 127–141, https://doi.org/10.1002/sim.2331, PMID 16217841.
[47] Tomiyama AJ, Hunger JM, Nguyen-Cuu J, Wells C. Misclassification of cardiometabolic health when using body mass index categories in NHANES 2005-2012. Int J Obes (Lond) May 1;40(5):883–6. 2016;40(5):883-6. 10.1038/ijo.2016.17, PMID 26841729.
[48] S.B. Heymsfield, D. Gallagher, L. Mayer, J. Beetsch, A. Pietrobelli, Scaling of human body composition to stature: new insights into body mass index, Am. J. Clin. Nutr. 86 (1) (2007 Jul) 82–91, https://doi.org/10.1093/ajcn/86.1.82, PMID 17616766.
[49] I. Tsakiridis, S. Giouleka, A. Mamopoulos, A. Kourtis, A. Athanasiadis, D. Filopoulou, et al., Diagnosis and management of gestational diabetes mellitus: an overview of national and international guidelines, Obstet. Gynecol. Surv. 76 (6) (2021 Jun) 367–381, https://doi.org/10.1097/OGX.0000000000000899, PMID 34192341.
[50] C.E. Powe, Early pregnancy biochemical predictors of gestational diabetes mellitus, Curr. Diab. Rep. 17 (2) (2017 Feb 22) 12, https://doi.org/10.1007/s11892-017-0834-y, PMID 28229385.
[51] K. Benhalima, P. Van Crombrugge, C. Moyson, J. Verhaeghe, S. Vandeginste, H. Verlaenen, et al., Estimating the risk of gestational diabetes mellitus based on the 2013 WHO criteria: a prediction model based on clinical and biochemical variables in early pregnancy, Acta Diabetol. 57 (6) (2020 Jun 1) 661–671, https://doi.org/10.1007/s00592-019-01469-5, PMID 31915927.
[52] A. Syngelaki, R. Kotecha, A. Pastides, A. Wright, K.H. Nicolaides, First-trimester biochemical markers of placentation in screening for gestational diabetes mellitus, Metabolism 64 (11) (2015 Nov 1) 1485–1489, https://doi.org/10.1016/j.metabol.2015.07.015, PMID 26362726.
[53] Cultural diversity of Australia. Australian Bureau of Statistics [internet]; 2022 [cited Mar 30 2023]. Available from: https://www.abs.gov.au/articles/cultural-diversity-australia.