Journal of Prevention, Diagnosis and Management of Human Diseases
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1202, DOI: 10.55529/jpdmhd.62.73.86

Research Paper

Explainable machine learning versus logistic regression for diabetes risk prediction in Pima Indian women: a calibration- and decision-focused internal validation study

Mainazarova Elmira[1*], Arzykulova Ainagul[2], Mamashov Nurbek[3]

[1*]Department of Public Health and Family Medicine, Jalal-Abad International University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.

[2]Department of Clinical Disciplines, Jalal-Abad International University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.

[3]Department of Clinical Disciplines, Jalal-Abad International University, Osh State University, Osh International Medical University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.

Article Info

Article History:
Received: 08 May 2026
Revised: 17 July 2026
Accepted: 24 July 2026
Published: 09 September 2026

Keywords:
Diabetes Mellitus
Clinical Prediction Model
Machine Learning
Explainable Artificial Intelligence
Decision Curve Analysis

ABSTRACT

Background: Machine learning (ML) models for diabetes prediction frequently emphasise accuracy while overlooking calibration and clinical utility. This study compared logistic regression (LR) with ML approaches and laboratory-based with non-laboratory predictor sets.

Methods: The Pima Indians Diabetes Database comprising 768 women aged ≥21 years, including 268 (34.9%) with diabetes, was analysed. Five models were evaluated: glucose-only LR, non-laboratory LR, full LR, random forest and gradient boosting. Missing physiologically implausible values were imputed within training folds. Model performance was assessed using repeated 5-fold cross-validation (10 repetitions), discrimination (AUROC), calibration slope, calibration-in-the-large, Brier score and decision curve analysis. SHapley Additive exPlanations (SHAP) assessed predictor contributions. Reporting followed TRIPOD+AI guidelines.

Results: Full LR achieved an AUROC of 0.837 (95% CI 0.809–0.864), comparable to random forest (0.838; 95% CI 0.809–0.865) and gradient boosting (0.831; 95% CI 0.800–0.859). Glucose-only and non-laboratory LR achieved AUROCs of 0.790 and 0.749, respectively. LR and random forest demonstrated acceptable calibration (slopes 0.93 and 1.07), whereas gradient boosting showed overextreme predictions (0.69). All models provided greater net benefit than testing all women across thresholds of 0.10–0.60. Plasma glucose and body mass index were the strongest predictors.

Conclusions: ML approaches provided no meaningful discrimination advantage over LR. Gradient boosting required recalibration. Non-laboratory prediction may support screening decisions where laboratory access is limited. External validation in diverse populations is essential before clinical implementation.

Journal of Prevention, Diagnosis and Management of Human Diseases (JPDMHD)

Corresponding Author:
Mainazarova Elmira
Department of Public Health and Family Medicine, Jalal-Abad International University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.
Email: elmira.sidikovna@gmail.com

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Diabetes mellitus is one of the fastest-growing chronic diseases worldwide. The International Diabetes Federation estimated that 537 million adults aged 20–79 years were living with diabetes in 2021, a figure projected to reach 783 million by 2045 [1]. The Global Burden of Disease Study 2021 estimated 529 million people with diabetes in 2021 and projected 1.31 billion by 2050 [2]. In India, the national Indian Council of Medical Research–India Diabetes (ICMR-INDIAB) study reported a diabetes prevalence of 11.4% and a prediabetes prevalence of 15.3% [3]. Because lifestyle intervention and metformin substantially reduce progression to type 2 diabetes in people at high risk [4], identifying such people early is a public-health priority.

Diagnosis relies on laboratory measures such as the 2-hour plasma glucose value of an oral glucose tolerance test (OGTT), fasting plasma glucose or glycated haemoglobin [5]. Where laboratory testing is scarce or costly, questionnaire-based scores such as the Finnish Diabetes Risk Score (FINDRISC) [6] and the Indian Diabetes Risk Score [7] are used to decide who should be tested. ML methods are increasingly proposed as more accurate alternatives. However, a systematic review of clinical prediction studies found no performance benefit of ML over logistic regression (LR) when comparisons were made at low risk of bias [8], and poor calibration – predicted risks that are systematically too high or too low – can make an otherwise discriminating model harmful in practice [9]. There is also growing concern that opaque 'black-box' models are used in high-stakes decisions when interpretable models would perform equally well [10].

The Pima Indians Diabetes Database is one of the most widely analysed clinical datasets in ML research, including studies comparing LR with ML classifiers [11]. Many analyses of this dataset report classification accuracy at a single threshold, treat physiologically impossible zero values as real measurements, or tune and evaluate models on the same data. These practices can overstate performance and give no information on calibration or clinical usefulness. The TRIPOD+AI statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence) now sets out reporting standards for prediction models developed with regression or ML methods [12].

This study therefore had three objectives: (i) to compare the discrimination, calibration and clinical net benefit of LR with random forest and gradient boosting under identical, pre-specified validation conditions; (ii) to quantify how much predictive information is lost when laboratory measurements are unavailable, by comparing a non-laboratory model with glucose-only and full models; and (iii) to examine whether explanations from an ML model agree with the associations estimated by LR.

2. METHODOLOGY

2.1 Data Source and Participants

We performed a secondary analysis of the Pima Indians Diabetes Database, originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases and analysed by [13]. The Pima Indian community living near Phoenix, Arizona, USA, has one of the highest recorded incidences of diabetes in the

world [14]. According to the dataset documentation, several constraints were applied when the records were selected from a larger database; all participants are women at least 21 years old of Pima Indian heritage. The dataset contains 768 records, with no identifying information. The copy analysed was obtained from a public mirror (see Data availability); its class distribution (500 without and 268 with diabetes) matches the documentation exactly.

2.2 Outcome

The outcome was the binary diabetes status recorded in the database. As documented with the dataset, diabetes was defined by World Health Organization criteria: a 2-hour post-load plasma glucose of at least 200 mg/dL at any survey examination, or diabetes found during routine medical care [13]. Smith et al. framed the task as forecasting the onset of diabetes [13], but the exact interval between predictor measurement and outcome ascertainment is not documented in the public release; we consider the consequences of this in the Discussion.

2.3 Candidate Predictors

All eight recorded variables were considered: number of pregnancies; plasma glucose concentration 2 hours after an oral glucose load (mg/dL); diastolic blood pressure (mm Hg); triceps skinfold thickness (mm); 2-hour serum insulin (μU/mL); body mass index (BMI, kg/m^2); the diabetes pedigree function, a score summarising diabetes history in relatives [13]; and age (years). No variable selection was performed, so that each model's predictor set was fixed before any analysis.

2.4 Handling of Implausible Values and Missing Data

Values of zero for plasma glucose, diastolic blood pressure, skinfold thickness, serum insulin and BMI are physiologically impossible in living adults and were recoded as missing. This affected glucose in 5 women, diastolic blood pressure in 35, BMI in 11, skinfold thickness in 227 (29.6%) and insulin in 374 (48.7%). Only 392 women (130 with diabetes) had complete data, so complete-case analysis would have discarded almost half of the cohort. All 768 women were therefore retained (Figure 1). For the logistic and random-forest models, each missing value was replaced by a single imputed value using iterative (chained-equations) imputation [15] with Bayesian ridge regression, as implemented in scikit-learn. The imputation model used only the predictors of the model concerned, not the outcome, so that it could be applied in the same way to new patients; for the glucose-only model it is therefore equivalent to mean imputation. The imputation model was fitted on the training folds only and then applied to the corresponding test fold, so that no information from test data entered model development. The gradient-boosting model used its built-in handling of missing values.

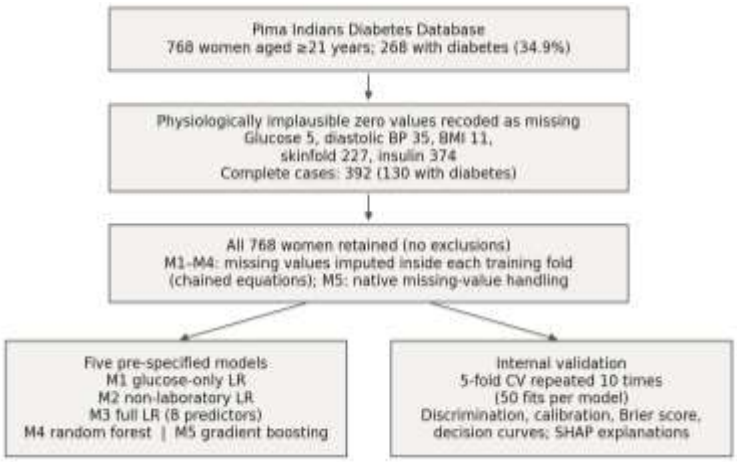


Figure 1. Study Flow and Analysis Design. BP, Blood Pressure; BMI, Body Mass Index; CV, Cross-Validation; LR, Logistic Regression; SHAP, SHapley Additive ExPlanations

2.5 Model Development

Five models were specified before analysis Table 1. M1 (glucose-only LR) represents a single laboratory test. M2 (non-laboratory LR) uses only variables obtainable by history-taking and simple examination. M3 (full LR) uses all eight predictors with linear terms. M4 is a random forest [16] and M5 a gradient-boosting machine [17], both using all eight predictors. Logistic regression models were fitted after standardising predictors with a negligible penalty (inverse regularisation strength C = 10^6), which is equivalent in practice to unpenalised maximum likelihood. To avoid optimistic estimates from tuning on the evaluation data, ML hyperparameters were fixed in advance at conventional values and were not tuned.

Table 1. Pre-Specified Models

Model	Algorithm	Predictors	Missing Data	Key Settings
M1	Logistic regression	2-h plasma glucose	Fold-wise imputation	Effectively unpenalised (C = 10^6)
M2	Logistic regression	Pregnancies, diastolic BP, BMI, pedigree function, age	Fold-wise imputation	Effectively unpenalised (C = 10^6)
M3	Logistic regression	All eight	Fold-wise imputation	Effectively unpenalised (C = 10^6)
M4	Random forest	All eight	Fold-wise imputation	500 trees; minimum 5 samples per leaf
M5	Gradient boosting (histogram-based)	All eight	Native handling	200 iterations; learning rate 0.05; maximum depth 3; minimum 20 samples per leaf

BP, blood pressure; BMI, body mass index. Single imputation by chained equations with Bayesian ridge regression, fitted on training folds only.

2.6 Internal Validation and Performance Measures

Performance was estimated by stratified 5-fold cross-validation repeated 10 times with different random partitions (50 model fits per model; 250 in total), so that every woman received an out-of-sample prediction in each repeat [18]. Calibration measures and the Brier score were computed within each repeat and averaged, and the mean and standard deviation of the AUROC across repeats are also reported. AUROC point estimates and 95% confidence intervals (CIs), paired differences in AUROC, threshold-based measures and decision curves were computed from each woman's prediction averaged across the 10 repeats. For CIs, 2,000 bootstrap resamples were drawn, and the same resamples were used for all models so that paired differences relative to M3 could be estimated. Because averaging predictions from repeated fits acts as a small ensemble, these pooled estimates can be slightly more favourable than those of a single fitted model, particularly for gradient boosting; both AUROC estimates are shown in Table 3.

Discrimination was measured by the area under the receiver operating characteristic curve (AUROC). Calibration was assessed with the calibration slope (ideal value 1; values below 1 indicate predictions that are too extreme) and calibration-in-the-large (ideal value 0), and displayed with smoothed calibration curves [9]. Overall accuracy was summarised by the Brier score (lower is better). Sensitivity, specificity and predictive values were reported at two illustrative thresholds: 0.20, representing a screening setting in which missed cases are costly, and 0.50. Clinical usefulness was evaluated by decision curve analysis, which weighs true positives against false positives across a range of threshold probabilities and compares each model with testing all women or none [19], [20].

2.7 Explainability and Associations

To explain the gradient-boosting model, SHapley Additive exPlanations (SHAP) values were computed with the path-dependent TreeSHAP algorithm on a model fitted to the full dataset [21], [22]. SHAP values express each predictor's contribution to an individual's predicted log-odds. To compare these explanations with conventional associations, a logistic model with all eight standardised predictors was fitted to 20 multiply imputed datasets (chained equations with posterior sampling), and odds ratios per standard deviation were pooled using Rubin's rules [15].

2.8 Sample Size

The dataset includes 268 outcome events for eight candidate predictors, about 34 events per predictor. We applied the three criteria of Riley et al. for binary outcomes [23], using an anticipated Cox–Snell R^2 of 0.31 (the apparent value of an eight-predictor logistic model fitted to the data with median imputation) and the observed outcome proportion of 0.349. The required sample sizes were 192 to keep the expected shrinkage factor at 0.9, 180 to limit optimism in model fit to 0.05, and 349 to estimate the overall outcome risk within ±0.05. The largest of these (349) is well below the 768 women available.

2.9 Software and Reproducibility

Analyses were performed in Python 3.11 with scikit-learn 1.8.0 [24], statsmodels 0.15.0 [25] and shap 0.51.0 [22]. A fixed random seed (2026) was used throughout. The complete analysis script, which regenerates every table and figure from the raw data file, is provided (see Code availability).

3. RESULTS AND DISCUSSION

3.1 Participant Characteristics

Of the 768 women, 268 (34.9%) had diabetes. Women with diabetes had higher median 2-hour plasma glucose (140 versus 107 mg/dL), BMI (34.3 versus 30.1 kg/m^2) and age (36 versus 27 years), and more pregnancies, than those without diabetes (Table 2). All eight characteristics differed between groups ($P < 0.001$). Missingness was substantial for serum insulin (48.7%) and skinfold thickness (29.6%) but low for the remaining variables.

Table 2. Characteristics of Participants by Diabetes Status

Characteristic	All (n=768)	No Diabetes (N=500)	Diabetes (N=268)	P Value	Missing, N (%)
Number of pregnancies	3 (1–6)	2 (1–5)	4 (2–8)	<0.001	0 (0.0)
2-h plasma glucose (mg/dL)	117.0 (99.0–141.0)	107.0 (93.0–125.0)	140.0 (119.0–167.0)	<0.001	5 (0.7)
Diastolic blood pressure (mm Hg)	72.0 (64.0–80.0)	70.0 (62.0–78.0)	74.5 (68.0–84.0)	<0.001	35 (4.6)
Triceps skinfold thickness (mm)	29.0 (22.0–36.0)	27.0 (19.0–33.0)	32.0 (27.0–39.0)	<0.001	227 (29.6)
2-h serum insulin (μU/mL)	125.0 (76.2–190.0)	102.5 (66.0–161.2)	169.5 (127.5–239.2)	<0.001	374 (48.7)
Body mass index (kg/m^2)	32.3 (27.5–36.6)	30.1 (25.6–35.3)	34.3 (30.9–38.9)	<0.001	11 (1.4)
Diabetes pedigree function	0.37 (0.24–0.63)	0.34 (0.23–0.56)	0.45 (0.26–0.73)	<0.001	0 (0.0)
Age (years)	29 (24–41)	27 (23–37)	36 (28–44)	<0.001	0 (0.0)

Values are median (interquartile range) of observed values. P values from the Mann–Whitney U test. Missing values are zeros recoded as missing.

3.2 Discrimination

Table 3. Discrimination, Overall Accuracy and Calibration (5-Fold Cross-Validation Repeated 10 Times)

Model	AUROC, Mean ± SD Across Repeats	AUROC (95% CI)[a]	ΔAUROC vs M3 (95% CI)[a]	Brier Score	Calibration Slope	Calibration-In-The-Large
M1 Glucose-only LR	0.790 ± 0.001	0.790 (0.757–0.822)	−0.047 (−0.068 to −0.026)	0.172	0.98	0.00
M2 Non-laboratory LR	0.748 ± 0.004	0.749 (0.714–0.781)	−0.088 (−0.115 to −0.062)	0.191	0.94	0.00
M3 Full LR	0.835 ± 0.002	0.837 (0.809–0.864)	Reference	0.157	0.93	0.00
M4 Random forest	0.835 ± 0.004	0.838 (0.809–0.865)	+0.001 (−0.013 to 0.015)	0.156	1.07	+0.01
M5 Gradient boosting	0.822 ± 0.006	0.831 (0.800–0.859)	−0.006 (−0.023 to 0.010)	0.164	0.69	+0.04

[a]Based on predictions averaged across repeats, with 2,000 paired bootstrap resamples. Brier score, calibration slope and calibration-in-the-large are means across the 10 repeats. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; LR, logistic regression; SD, standard deviation.

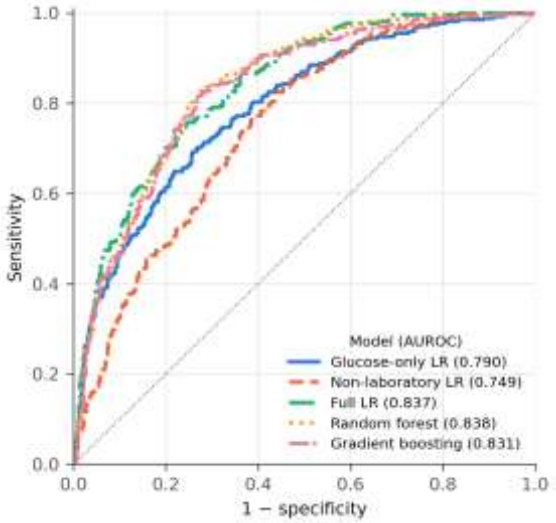


Figure 2. Receiver Operating Characteristic Curves of the Five Models (Cross-Validated Predictions Averaged over 10 Repeats). The Diagonal Indicates no Discrimination. AUROC, Area under the Receiver Operating Characteristic Curve; LR, Logistic Regression

3.3 Calibration and Overall Accuracy

The logistic models were well calibrated overall Figure 3: calibration slopes were 0.98 (M1), 0.94 (M2) and 0.93 (M3), with calibration-in-the-large close to zero, although the non-laboratory model

overestimated risk at predicted probabilities above about 0.5. The random forest had a slope of 1.07, indicating predictions slightly too moderate. Gradient boosting had a slope of 0.69, meaning its predicted risks were too extreme – too high for women at high risk and too low for women at low risk. Brier scores (means across the 10 repeats) were lowest for the random forest (0.156) and full LR (0.157), followed by gradient boosting (0.164), the glucose-only model (0.172) and the non-laboratory model (0.191). Using averaged predictions, the Brier score of the full LR model was 0.156 (95% CI 0.142–0.171), overlapping with that of the random forest (0.155; 0.143–0.168).

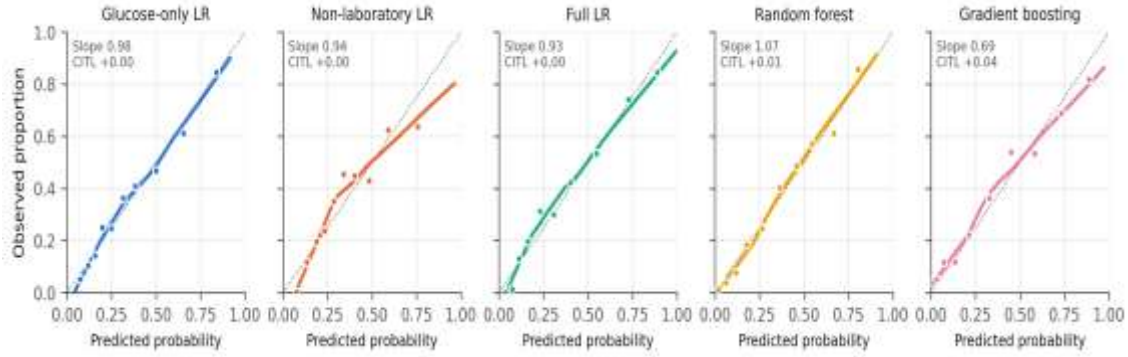


Figure 3. Calibration of the Five Models. Lines are Locally Weighted (LOWESS) Smooths of Observed Outcome against Predicted Probability; Points are Observed Proportions within Deciles of Predicted Risk; the Dotted Line Indicates Perfect Calibration. Slope and Calibration-in-the-Large (CITL) are Means across the 10 Cross-Validation Repeats

3.4 Classification at Illustrative Thresholds

At a threshold of 0.20, the full LR model identified 88.8% of women with diabetes with a specificity of 56.8% and a negative predictive value of 90.4% Table 4. The non-laboratory model achieved a similar sensitivity (93.3%) but at the cost of a much lower specificity (38.8%), meaning that more women without diabetes would be referred for confirmatory testing. At a threshold of 0.50, all models had specificities above 83% but sensitivities of only 40.7%–59.0%.

Table 4. Classification Performance at Two Illustrative Risk Thresholds

Model	Sens. (0.20)	Spec. (0.20)	PPV (0.20)	NPV (0.20)	Sens. (0.50)	Spec. (0.50)	PPV (0.50)	NPV (0.50)
M1 Glucose-only LR	88.4%	48.6%	48.0%	88.7%	49.3%	87.8%	68.4%	76.3%
M2 Non-laboratory LR	93.3%	38.8%	45.0%	91.5%	40.7%	86.2%	61.2%	73.1%
M3 Full LR	88.8%	56.8%	52.4%	90.4%	56.0%	88.0%	71.4%	78.9%
M4 Random forest	92.2%	54.4%	52.0%	92.8%	56.3%	86.2%	68.6%	78.6%
M5 Gradient boosting	88.8%	61.2%	55.1%	91.1%	59.0%	83.6%	65.8%	79.2%

Computed from cross-validated predictions averaged across repeats. NPV, negative predictive value; PPV, positive predictive value; Sens., sensitivity; Spec., specificity.

3.5 Clinical Usefulness

Decision curves are shown in Figure 4. For all threshold probabilities from 0.10 to 0.60, every model had a higher net benefit than testing all women or testing none. At a threshold of 0.20, the net benefit of the full LR model was 0.240, compared with 0.186 for testing all women; this is a net gain equivalent to

5.3 true-positive cases per 100 women. At most thresholds between about 0.18 and 0.43, the random forest and gradient boosting had modestly higher net benefit than full LR (for example, at 0.30: 0.206 and 0.210 versus 0.189). The non-laboratory model had lower net benefit than the full and ML models at every threshold, was similar to the glucose-only model up to a threshold of about 0.25, and fell below it at higher thresholds.

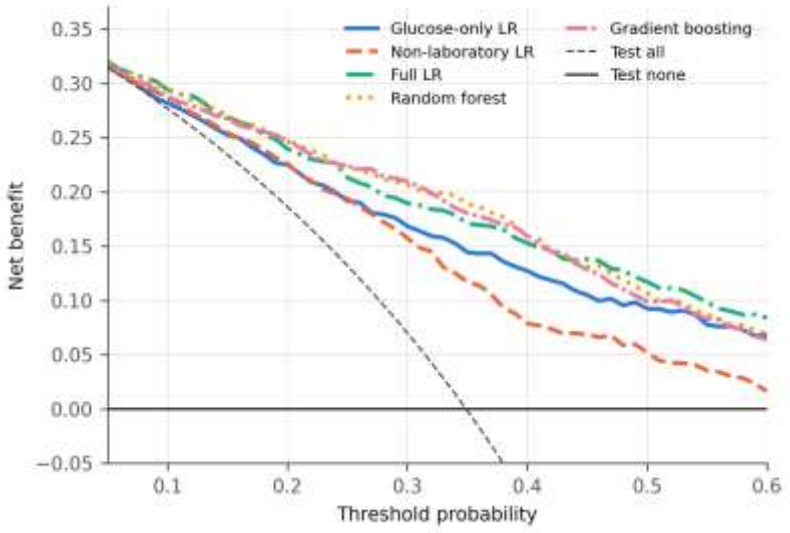


Figure 4. Decision Curve Analysis. Net Benefit is plotted against the Threshold Probability at which a Woman would be referred for Diagnostic Testing. 'Test All' Assumes Every Woman is tested; 'Test None' assumes no One is. LR, Logistic Regression

3.6 Associations and Model Explanations

In the multiply imputed logistic model Table 5, higher 2-hour plasma glucose (odds ratio per standard deviation 3.04 (95% CI 2.37–3.91)), BMI (1.86 (95% CI 1.42–2.43)), number of pregnancies (1.52 (95% CI 1.23–1.88)) and diabetes pedigree function (1.33 (95% CI 1.10–1.61)) were independently associated with diabetes. Age, diastolic blood pressure, skinfold thickness and serum insulin were not, after adjustment for the other predictors.

Table 5. Multivariable Logistic Regression: Odds Ratios per Standard Deviation (20 Imputations, Rubin's Rules)

Predictor	1 SD Equals	Odds Ratio (95% CI)	P Value
Number of pregnancies	3.4	1.52 (1.23–1.88)	<0.001
2-h plasma glucose (mg/dL)	30.5	3.04 (2.37–3.91)	<0.001
Diastolic blood pressure (mm Hg)	12.4	0.91 (0.74–1.12)	0.374
Triceps skinfold thickness (mm)	10.5	1.03 (0.79–1.35)	0.814
2-h serum insulin (μU/mL)	112.0	0.99 (0.77–1.28)	0.962
Body mass index (kg/m^2)	6.9	1.86 (1.42–2.43)	<0.001
Diabetes pedigree function	0.33	1.33 (1.10–1.61)	0.004
Age (years)	11.8	1.16 (0.93–1.45)	0.192

Odds ratios express the change in odds of diabetes for a one-standard-deviation increase in each predictor, adjusted for all others. SD, standard deviation (mean over imputed datasets); CI, confidence interval.

The SHAP analysis of the gradient-boosting model ranked 2-hour plasma glucose (mean absolute SHAP value 0.99 log-odds) and BMI (0.76) as the most influential predictors, followed by age (0.42) and the diabetes pedigree function (0.32); diastolic blood pressure had the smallest influence (0.11) Figure 5. The dependence plots Figure 6 show that the contribution of glucose rose steadily across its range, whereas the contribution of BMI rose steeply up to about 30 kg/m^2 and changed little between about 30 and 40 kg/m^2. The ordering of the two dominant predictors therefore agreed with the logistic model. Age ranked higher in the SHAP analysis than its adjusted odds ratio suggested, which may reflect non-linear effects and interactions captured by the tree model, as well as the correlation between age and number of pregnancies, which the two methods apportion differently. Beyond the two dominant predictors, the rankings from the two approaches therefore differed.

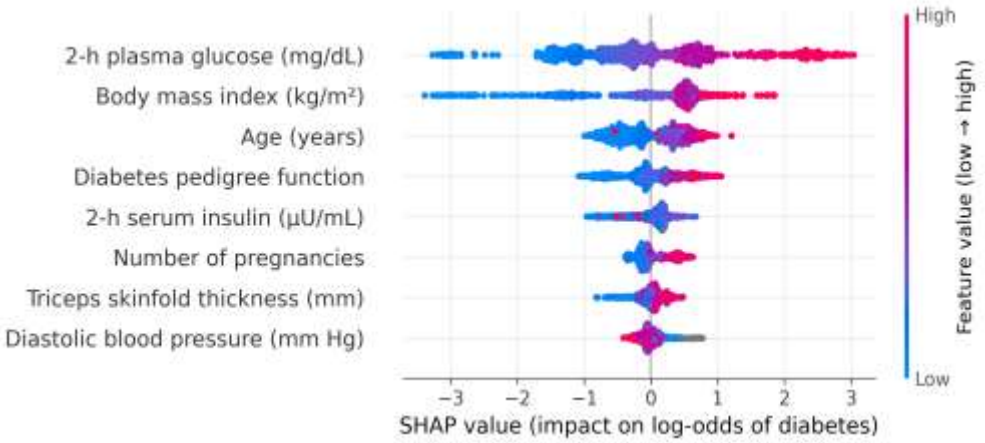


Figure 5. SHAP Summary for the Gradient-Boosting Model. Each Point is One Woman; Its Horizontal Position Shows How Much That Predictor Raised or Lowered Her Predicted Log-Odds of Diabetes, and Its Colour Shows the Predictor Value (Grey Indicates a Missing Value). Predictors Are Ordered by Mean Absolute SHAP Value

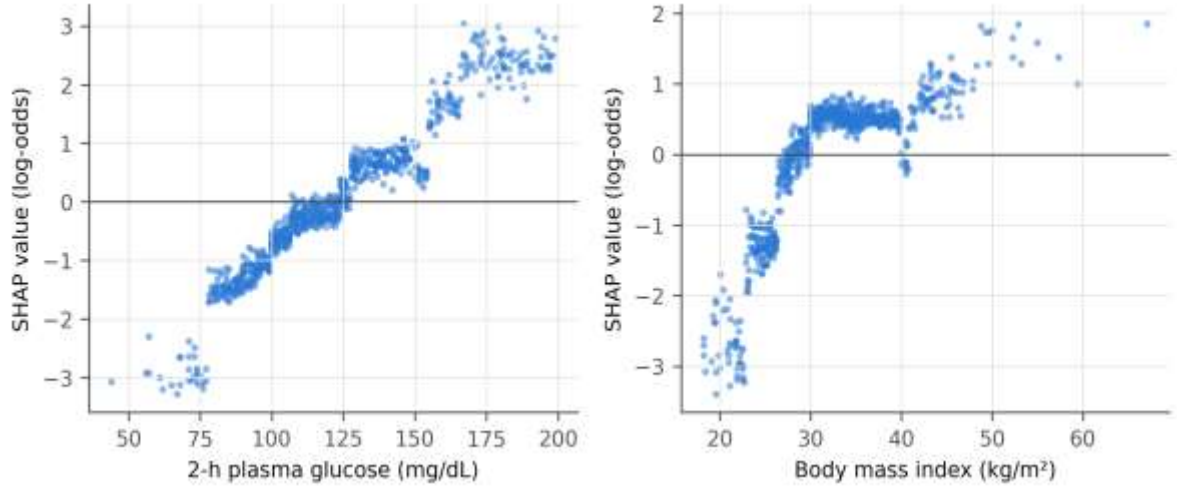


Figure 6. SHAP Dependence Plots for 2-Hour Plasma Glucose and Body Mass Index in the Gradient-Boosting Model. Values above Zero Increase, and Values below Zero Decrease, the Predicted Log-Odds of Diabetes. The Step-Like Pattern Reflects the Tree-Based Structure of the Model

3.7 Discussion

3.7.1 Principal Findings

In this internal validation study, a transparent logistic regression model using eight routinely recorded variables discriminated as well as random forest and gradient boosting (AUROC 0.837 versus 0.838 and 0.831 using averaged predictions; 0.835 versus 0.835 and 0.822 for single repeats), and it was well calibrated. Gradient boosting was noticeably miscalibrated, so its predicted probabilities could not be used directly for individual risk communication without recalibration. Removing laboratory measurements reduced discrimination from 0.837 to 0.749, but the non-laboratory model remained reasonably calibrated and gave greater net benefit than testing all women. SHAP explanations of the ML model agreed with the logistic model on the two dominant predictors, 2-hour plasma glucose and BMI.

3.7.2 Comparison with Previous Work

Our results agree with the systematic review by Christodoulou et al., which found no performance benefit of ML over LR for clinical prediction when studies were at low risk of bias [8]. In the original analysis of this dataset, Smith et al. reported a sensitivity and specificity of 76% on a single held-out set of 192 women [13]. Our estimates are not directly comparable because we used repeated cross-validation over all women, treated impossible zero values as missing and reported calibration and net benefit, which single-threshold accuracy does not capture. The markedly different calibration of gradient boosting, despite similar discrimination, illustrates the point made by Van Calster et al. that discrimination and calibration are distinct properties and that a model can rank patients well while misestimating their absolute risk [9].

3.7.3 Implications for Practice

For settings in which OGTT or other laboratory tests are limited, our findings support a two-step approach: a simple non-laboratory model or an established score such as FINDRISC or the Indian Diabetes Risk Score [6], [7] to decide who should be tested, followed by laboratory confirmation. The decision curves suggest that such pre-selection yields more benefit than universal testing across a wide range of thresholds, although the appropriate threshold should be chosen with clinicians and local health-system constraints in mind. Where an ML model is preferred, our results show that calibration must be checked and, if necessary, corrected before predicted probabilities are shown to patients or clinicians [9], [18]. Finally, the agreement between SHAP explanations and regression coefficients on the dominant predictors supports the view that, when an interpretable model performs as well as a complex one, the interpretable model should be preferred for high-stakes decisions [10].

3.7.4 Strengths and Limitations

The strengths of this study include pre-specification of all models, fold-wise imputation that prevents information leakage, repeated cross-validation, and the joint assessment of discrimination, calibration, net benefit and explainability in line with TRIPOD+AI [12]. The analysis is fully reproducible from a public dataset and a single script, and a completed TRIPOD+AI checklist is provided as Supplementary File S2.

The study also has important limitations. First, validation was internal only; performance in other populations is unknown, and the cohort – women of one Indigenous community, with a very high diabetes incidence [14] – limits transportability to men, to other ethnic groups and to South Asian populations. Second, the data were collected before 1988 [13], and selection constraints applied to the records are not fully documented. Third, the outcome definition includes a 2-hour plasma glucose threshold of 200 mg/dL, while 2-hour plasma glucose is also a predictor, and the timing between measurement and outcome is not documented; no woman in the dataset has a recorded 2-hour plasma glucose of 200 mg/dL or more (maximum 199 mg/dL). The glucose-based models may therefore partly reflect current rather than future diabetes, and their performance should not be interpreted as purely prognostic. Fourth, insulin and skinfold values were missing for 48.7% and 29.6% of women; single imputation assumes that missingness depends only on observed data and does not propagate imputation uncertainty into the cross-validated estimates. Fifth, ML hyperparameters were not tuned; tuning within nested cross-validation might improve

ML performance slightly, although the systematic evidence suggests that large gains are unlikely [8]. Sixth, the bootstrap confidence intervals, based on averaged cross-validated predictions, do not capture all sources of variability, and the decision curves are presented without confidence intervals. Finally, SHAP values describe how the model uses the data and should not be interpreted causally.

4. CONCLUSION

In a cohort of 768 Pima Indian women, a transparent logistic regression model predicted diabetes as accurately as random forest and gradient boosting, with better or similar calibration; the tree-based models offered a modestly higher net benefit only at intermediate thresholds. Gradient boosting required recalibration, and both explainability and regression analyses identified 2-hour plasma glucose and BMI as the main drivers of predicted risk. A non-laboratory model offers a less accurate but potentially useful first step where laboratory testing is limited. Prediction studies in diabetes should report calibration and clinical usefulness alongside discrimination, and models should be externally validated in contemporary, diverse populations before clinical use.

Acknowledgments

The author thanks the National Institute of Diabetes and Digestive and Kidney Diseases, the original owner of the database, the Pima Indian community whose participation made the data possible, and the researchers whose works have been cited in this study.

Funding Information

This research has received no external funding.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mainazarova Elmira	✓	✓	✓		✓				✓		✓			
Arzykulova Ainagul		✓		✓		✓		✓		✓		✓		
Mamashov Nurbek			✓				✓				✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The author is affiliated with HM Publishers, the publisher of this journal. The manuscript was handled by an independent editor and peer reviewers, and the author had no role in its editorial assessment. No other conflict of interest is declared.

Informed Consent

Not Applicable. The study used only publicly available, anonymised data.

Ethical Approval

Not Applicable. The study is a secondary analysis of an anonymised, publicly available dataset that contains no personal identifiers; no new data were collected and there was no contact with participants.

Data Availability

The Pima Indians Diabetes Database is publicly available; the copy analysed was obtained from a public repository (MD5 checksum 56a8d8ae619fcc223941e54f361b8406) together with its

documentation file. No new datasets were created. The complete analysis script is available from the author on reasonable request.

Use of Artificial Intelligence

An AI model (Claude, Anthropic) assisted in writing the analysis code, producing the figures and drafting the manuscript. The authors reviewed and re-ran all code, verified every result and reference, and take full responsibility for the content.

REFERENCES

[1] H. Sun et al., "IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," Diabetes Res Clin Pract, vol. 183, Art. no. 109119, 2022. doi.org/10.1016/j.diabres.2021.109119

[2] GBD 2021 Diabetes Collaborators, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021," Lancet, vol. 402, no. 10397, pp. 203–234, 2023. doi.org/10.1016/S0140-6736(23)01301-6

[3] R. M. Anjana et al., "Metabolic non-communicable disease health report of India: the ICMR-INDIAB national cross-sectional study (ICMR-INDIAB-17)," Lancet Diabetes Endocrinol, vol. 11, no. 7, pp. 474–489, 2023. doi.org/10.1016/S2213-8587(23)00119-5

[4] W. C. Knowler et al., "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," N Engl J Med, vol. 346, no. 6, pp. 393–403, 2002. doi.org/10.1056/NEJMoa012512

[5] American Diabetes Association Professional Practice Committee, "2. Diagnosis and classification of diabetes: Standards of Care in Diabetes—2024," Diabetes Care, vol. 47, no. Suppl 1, pp. S20–S42, 2024. doi.org/10.2337/dc24-S002

[6] J. Lindström and J. Tuomilehto, "The diabetes risk score: a practical tool to predict type 2 diabetes risk," Diabetes Care, vol. 26, no. 3, pp. 725–731, 2003. doi.org/10.2337/diacare.26.3.725

[7] V. Mohan, R. Deepa, M. Deepa, S. Somannavar, and M. Datta, "A simplified Indian Diabetes Risk Score for screening for undiagnosed diabetic subjects," J Assoc Physicians India, vol. 53, pp. 759–763, 2005.

[8] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," J Clin Epidemiol, vol. 110, pp. 12–22, 2019. doi.org/10.1016/j.jclinepi.2019.02.004

[9] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, "Calibration: the Achilles heel of predictive analytics," BMC Med, vol. 17, Art. no. 230, 2019. doi.org/10.1186/s12916-019-1466-7

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat Mach Intell, vol. 1, no. 5, pp. 206–215, 2019. doi.org/10.1038/s42256-019-0048-x

[11] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," Int J Environ Res Public Health, vol. 18, no. 14, Art. no. 7346, 2021. doi.org/10.3390/ijerph18147346

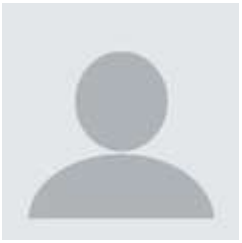
[12] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024. doi.org/10.1136/bmj-2023-078378

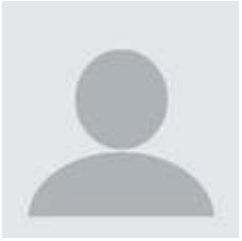
[13] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in Proceedings of the Annual Symposium on Computer Application in Medical Care. Washington, DC: IEEE Computer Society Press;, pp. 261–265, 1988.

[14] W. C. Knowler, P. H. Bennett, R. F. Hamman, and M. Miller, "Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota," Am J Epidemiol, vol. 108, no. 6, pp. 497–505, 1978. doi.org/10.1093/oxfordjournals.aje.a112648
[15] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," J Stat Softw, vol. 45, no. 3, pp. 1–67, 2011. doi.org/10.18637/jss.v045.i03
[16] L. Breiman, "Random forests," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001. doi.org/10.1023/A:1010933404324
[17] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Ann Stat, vol. 29, no. 5, pp. 1189–1232, 2001. doi.org/10.1214/aos/1013203451
[18] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: seven steps for development and an ABCD for validation," Eur Heart J, vol. 35, no. 29, pp. 1925–1931, 2014. doi.org/10.1093/eurheartj/ehu207
[19] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," Med Decis Making, vol. 26, no. 6, pp. 565–574, 2006. doi.org/10.1177/0272989X06295361
[20] B. Van Calster et al., "Reporting and interpreting decision curve analysis: a guide for investigators," Eur Urol, vol. 74, no. 6, pp. 796–804, 2018. doi.org/10.1016/j.eururo.2018.08.038
[21] S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, in Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, CA, 2017. arXiv:1705.07874. arxiv.org/abs/1705.07874
[22] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat Mach Intell, vol. 2, no. 1, pp. 56–67, 2020. doi.org/10.1038/s42256-019-0138-9
[23] R. D. Riley et al., "Calculating the sample size required for developing a clinical prediction model," BMJ, vol. 368, Art. no. m441, 2020. doi.org/10.1136/bmj.m441
[24] F. Pedregosa et al., "Scikit-learn: machine learning in Python," J Mach Learn Res, vol. 12, pp. 2825–2830, 2011.
[25] S. Seabold and J. Perktold, Statsmodels: econometric and statistical modeling with Python, in Proceedings of the 9th Python in Science Conference. Austin, TX;, pp. 92–96, 2010. doi.org/10.25080/Majora-92bf1922-011

How to Cite: Mainazarova Elmira, Arzykulova Ainagul, Mamashov Nurbek. (2026). Explainable machine learning versus logistic regression for diabetes risk prediction in Pima Indian women: a calibration- and decision-focused internal validation study. Journal of Prevention, Diagnosis and Management of Human Diseases (JPDMHD), 6(2), 73-86. https://doi.org/10.55529/jpdmhd.62.73.86

BIOGRAPHIES OF AUTHORS

	**Mainazarova Elmira**, is an associate professor in the Department of Public Health and Family Medicine at Jalal-Abad International University, Manas, Kyrgyzstan. Their work concerns population health and primary care, including preventive medicine and the early detection of disease in the community. They are interested in how screening and risk-assessment tools behave in routine practice and in what their results mean for the patients to whom they are applied. Email: elmira.sidikovna@gmail.com
	**Arzykulova Ainagul**, is a lecturer in the Department of Clinical Disciplines at Jalal-Abad International University, Manas, Kyrgyzstan. Their interests lie in clinical practice and clinical teaching, with attention to diagnostic reasoning and patient management. They are concerned with how new diagnostic methods are validated before they reach the ward and with what their reported performance means for everyday care.



Mamashov Nurbek, is an associate professor in the Department of Clinical Disciplines at Jalal-Abad International University, Manas, Kyrgyzstan, and holds an appointment at Osh State University. Their research addresses the clinical care of patients, from presentation and investigation through to management and follow-up. They take a particular interest in the interpretation of diagnostic results under uncertainty and in the training of junior clinicians.