International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 10 | Oct 2026 www.irjet.net p-ISSN: 2395-0072

Evaluating Classical Machine Learning Models for Three-Class Diabetes Status Prediction and External Validation using Primary Data from Mumbai

Dr. Supriya Kamale[1], Aarya Sanglikar[2]

[1]*Assistant Professor, D.Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*

[2]*Student, Department of MCA, D.Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*

Abstract - *India has one of the highest numbers of people living with diabetes globally. Prediabetes is a critical phase where prevention is most effective, but detection is often overlooked. This study assesses five classical machine learning models - logistic regression, random forest, extra trees, gradient boosting and histogram-based gradient boosting - for predicting three diabetes statuses: No Diabetes, Prediabetes and Diabetes. The models are trained using data from US Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) which includes 229,781 participants. The random forest model, which demonstrated the strongest performance, was evaluated using primary data gathered from 93 participants with complete cases in Mumbai. The model showed an accuracy of 0.762 on the CDC dataset, balanced accuracy of 0.489 and macro F1 score of 0.448 (recall: 0.800 No Diabetes, 0.656 Diabetes, 0.011 Prediabetes). On the Mumbai data, the accuracy was 0.677, balanced accuracy was 0.432 and macro F1 was 0.404. None of the 18 prediabetic participants in Mumbai were identified by the model. The findings reveal that accuracy can be misleading due to class imbalance in the data. The study shows that models trained on US survey data may not work well in South Asian urban populations, emphasizing the need for locally relevant, laboratory-confirmed clinical indicators.*

Key Words: Diabetes, Prediabetes, Machine Learning, External Validation, Class Imbalance, India, BRFSS

1. INTRODUCTION

Diabetes mellitus has evolved into one of the most pressing global public health challenges of the twenty-first century, with South Asia experiencing a particularly dramatic surge in disease incidence. According to the Indian Council of Medical Research - India Diabetes (ICMR - INDIAB) national cross-sectional study of 113,043 adults, India currently harbors an estimated 101 million individuals living with diabetes - representing a national prevalence of 11.4% - alongside an alarming 136 million individuals in the prediabetic phase [1]. Prediabetes, characterized by blood glucose or glycated hemoglobin levels that are elevated above normal baseline values but remain below the diagnostic threshold for overt diabetes, represents the critical physiological window where preventative lifestyle modifications and clinical interventions demonstrate maximum efficacy [2]. Despite a larger window opportunity for early intervention, prediabetes remains largely undiagnosed. In the United States National Health and Nutrition Examination Survey it was found that only approximately 11% of individuals with prediabetes had ever been diagnosed with prediabetes by a healthcare provider [2].

To address the bottleneck of costly routine laboratory screening across large cohorts, population health surveillance frameworks such as the US Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) - collect extensive self-reported health, socioeconomic, and lifestyle metrics [3], [4]. Prior computational research has increasingly leveraged these non-invasive, large-scale survey indicator sets to construct predictive risk stratification classifiers using classical machine learning algorithms [5]. For instance, Xie et al. trained six standard classification models on 138,146 respondents from the 2014 BRFSS dataset, demonstrating areas under the receiver operating characteristic curve between 0.718 and 0.795, while nothing that the sensitivity of survey-based models still needed improvement [5].

However, there are two significant methodological flaws in the majority of the work assessing machine learning risk models which are:

1. **Binary Framing and Class Imbalance Distortion:** Many survey-based studies, including that of Xie et al., simplify diabetes prediction into a binary task and do not predict prediabetes as a separate target class [5]. In instances where prediabetes is included within imbalanced datasets, algorithms optimized strictly for global accuracy can achieve high headline scores simply by predicting the dominant majority class while failing entirely on minority diagnostic categories [6].
2. **Lack of External Validation and Demographic Transportability:** Machine learning models are typically trained using Western population surveys, such as the 2015 BRFSS dataset, and evaluated using internal validation methods [3], [5], [7]. While

internal cross-validation helps measure how consistently a model performs, it does not check whether the model can be applied successfully in different populations with varying social, demographic and biological traits [7].

A significant concern arises when models derived from United States survey data, such as the 2015 BRFSS, are applied to populations in South Asia [3]. The World Health Organization notes that Asian populations exhibit distinct body composition characteristics relative to Western groups, implying that they confront elevated risks of metabolic disorders and type 2 diabetes at lower Body Mass Index values (with Asian public health action thresholds established at 23.0 and 27.5kg/m^2) [8]. Moreover, regional screening tools, like the Indian Diabetes Risk Score, place strong emphasis on waist circumference and family history of diabetes [9], which are not included in standardized Western surveys like the BRFSS [3], [4].

2. PROBLEM STATEMENT

Prediabetes is widespread but often goes unnoticed. In India alone, it is estimated that 136 million people have prediabetes, and 101 million have diabetes [1]. While population heavy surveys can help with low-cost risk screening, there are three main methodological issues that limit the current research [5].

First, many machine learning studies treat diabetes prediction as a simple yes - or - no task [5] or depend on clinical lab results [10], [11]. Using non-invasive survey data to classify three categories - no diabetes, prediabetes and diabetes - is not well - researched [12]. Second, in situations where the classes are highly imbalanced, overall accuracy may prove deceptive and might fail to capture how the model performs on the less frequent prediabetic cases [6], [13]. Third, models developed using US survey data, like the 2015 BRFSS, are rarely tested on international groups [7], [14], even though there are big differences in BMI thresholds, risk factors, and body composition in South Asian populations [8], [9].

The core problem addressed in this study is evaluating how effectively classical machine learning models trained on US survey data classify three-class diabetes status, how accurately they detect prediabetes under class imbalance, and whether the selected model holds up when externally validated on an independent primary sample from Mumbai [7], [3].

3. LITERATURE REVIEW

To establish the academic background for this work, prior research in classical machine learning for diabetes prediction, imbalanced data handling, and external validation is reviewed below.

3.1 Classical Machine Learning Models for Diabetes Prediction

Xie et al. (2019) [5] evaluated six machine learning models including support vector machines, decision trees, logistic regression, random forests, neural networks, and Gaussian naive Bayes using 138,146 respondents from the 2014 BRFSS dataset. They obtained areas under the curve ranging between 0.718 and 0.795, observing that neural networks yielded the best AUC while decision trees achieved the highest sensitivity, though survey-based models still required sensitivity improvements.

Kavakiotis et al. (2017) [10] conducted a systematic review of machine learning and data mining methods in diabetes research, identifying prediction and diagnosis as the most popular categories where 85% of utilized algorithms were supervised and predominantly applied to clinical datasets.

Zou et al. (2018) [15] utilized hospital physical-examination data from Luzhou, China, containing 14 attributes to train decision trees, random forests, and neural networks using 5-fold cross-validation and repeated balanced sampling, successfully comparing methods and ranking features for binary diabetes prediction.

Lai et al. (2019) [11] analyzed 13,309 Canadian patients aged 18 to 90 with laboratory values using logistic regression, gradient boosting, decision trees, and random forests, demonstrating that gradient boosting and logistic regression outperformed other models when adjusting thresholds and class weights.

Deberneh and Kim (2021) [12] examined Korean population data using logistic regression, random forests, support vector machines, XGBoost, and ensembles, with features selected via ANOVA, chi-squared, and recursive elimination to forecast normal, prediabetes, or diabetes status the following year, establishing a three-class precedent using clinical variables.

3.2 Imbalanced Data Handling and Prediabetes Detection

He and Garcia (2009) [6] reviewed learning from imbalanced data, establishing that underrepresented classes and severe class skew require dedicated principles, algorithms, and specialized evaluation tools.

Ramezankhani et al. (2016) [16] applied probabilistic neural networks, naive Bayes, and decision trees with and without SMOTE on 6,647 non-diabetic adults from the Tehran Lipid and Glucose Study over 10 years, finding that naive Bayes achieved the best Youden's index and that oversampling effects heavily depend on the chosen classifier.

Choi et al. (2014) [17] used artificial neural networks and support vector machines to screen for prediabetes defined by

fasting plasma glucose in a Korean population, showing that both models performed better than traditional logistic-regression screening scores.

Sokolova and Lapalme (2009) [13] systematically analyzed 24 performance measures across binary, multi-class, and hierarchical tasks, concluding that metrics differ based on confusion-matrix changes and that evaluation measures must closely match task objectives.

Saito and Rehmsmeier (2015) [18] compared ROC and precision-recall plots for binary classifiers on imbalanced data, demonstrating that ROC plots can be deceptive under class skew and that precision-recall plots provide much more informative evaluations for rare classes.

3.3 Model Transportability and External Validation

Steyerberg and Harrell (2016) [7] provided a methodological commentary on prediction models, strongly recommending internal bootstrap validation, internal-external validation, and fully independent external validation to ensure real-world reliability.

Debray et al. (2015) [19] established a framework for interpreting external validation studies using deep vein thrombosis models, noting that case-mix comparisons determine whether a sample tests model reproducibility or true demographic transportability.

Collins et al. (2024) [14] reviewed how clinical prediction models should be evaluated, from development to external validation, warning that validated models can still experience substantial drops in discrimination or calibration when deployed in new data settings.

Van Calster et al. (2019) [20] analyzed the calibration of risk prediction models as the "Achilles heel" of predictive analytics, noting that models developed in high-incidence settings can severely overestimate risk in lower-incidence settings and emphasizing the need for external calibration curves.

Mohan et al. (2005) [9] developed the Indian Diabetes Risk Score (IDRS) from the CURES study in Chennai using age, abdominal obesity, family history, and physical activity, achieving, at a cut-off score of 60, 72.5% sensitivity and 60.1% specificity for undiagnosed diabetes while highlighting India-specific predictors missing in standard Western surveys.

4. RESEARCH GAP

None of the fifteen studies reviewed combines all of the following methodological elements:

Three-Class Survey-Based Outcome: While prior research has explored machine learning for diabetes prediction, most models treat the target as a binary outcome or rely heavily on invasive clinical and laboratory measurements. The single study incorporating prediabetes as an independent class relies on clinical longitudinal data from a Korean cohort [12]. Consequently, predicting a three-class status (No Diabetes, Prediabetes, and Diabetes) using non-invasive survey indicators alone remains unaddressed.

Balanced Minority-Class Evaluation: Under severe class skew, standard headline accuracy metrics can mask prediction failures on minority classes like prediabetes. Few of the studies reviewed evaluate multi-class models using balanced metrics, per-class recall, and explicit precision-recall comparisons against a majority-class baseline.

Independent External Validation in South Asia: While internal cross-validation ensures reproducibility, prediction models are rarely tested via fully independent external validation [7] in new geographic or demographic populations. Furthermore, none of the studies reviewed externally validated a model trained on a Western health survey in a South Asian cohort, where distinct body composition profiles and lower BMI risk thresholds apply [8].

Transparent Pre-Specified Reporting: Reporting guidance for prediction model studies calls for transparent reporting of how models were developed and evaluated [21]. This study explicitly separates the pre-specified model selection rule from post-hoc adjustments made after the external results were seen.

5. RESEARCH QUESTIONS

RQ1. How accurately can classical machine learning models (logistic regression, random forest, extra trees, gradient boosting and histogram-based gradient boosting) classify individuals as No Diabetes, Prediabetes or Diabetes from structured health, lifestyle, demographic and socioeconomic indicators?

RQ2. How does model performance differ across the three diabetes-status categories and how well is Prediabetes detected compared with No Diabetes and Diabetes when classes are heavily imbalanced?

RQ3. Does the model selected on the CDC BRFSS data retain its predictive performance when applied, without retraining, to an independent primary sample of participants from Mumbai?

RQ4. How do the Mumbai and CDC samples differ in their demographic, lifestyle and health characteristics, and could these differences explain any change in model performance?

6. RESEARCH OBJECTIVES

1. To develop and evaluate classical machine learning models for classifying individuals into three diabetes-status categories (No Diabetes, Prediabetes and Diabetes) using structured health, lifestyle, demographic and socioeconomic indicators.
2. To compare the performance of the developed machine learning models across the three diabetes-status categories using appropriate evaluation metrics, with particular emphasis on Prediabetes detection and minority-class performance.
3. To externally validate the selected machine learning model using independently collected primary data from approximately 100 participants in Mumbai, by comparing the model-predicted diabetes status with the participants' reported diabetes status.

7. HYPOTHESES

HYPOTHESIS 1:

H_0: Classical machine learning models trained on the BRFSS indicators do not classify three-class diabetes status better than chance (balanced accuracy of 0.333 or lower on the CDC holdout set).

H_1: Classical machine learning models trained on the BRFSS indicators classify three-class diabetes status better than chance.

HYPOTHESIS 2:

H_0: For at least one model, Prediabetes F1-score is not lower than Diabetes F1-score.

H_1: For every model, Prediabetes F1-score is lower than Diabetes F1-score.

HYPOTHESIS 3:

H_0: The balanced accuracy of the selected model on the Mumbai sample is equal to or higher than its balanced accuracy on the CDC holdout set.

H_1: The balanced accuracy of the selected model on the Mumbai sample is lower than its balanced accuracy on the CDC holdout set.

HYPOTHESIS 4:

H_0: Overall accuracy of the selected model on the Mumbai sample does not exceed the majority-class baseline.

H_1: Overall accuracy of the selected model on the Mumbai sample exceeds the majority-class baseline.

8. RESEARCH METHODOLOGY

8.1 Study Design and Reporting

This is a model development and external validation study, reported following the TRIPOD+AI guidance for prediction model studies [21]. All analyses were run as one sequential Python 3.13.16 notebook using scikit-learn 1.6.1 [22], pandas 2.2.3, NumPy 2.1.3, Matplotlib 3.10.0, seaborn 0.13.2, joblib 1.6.0 and openpyxl 3.1.5, with a fixed random seed. The Mumbai data were not used for training, hyperparameter tuning or model selection in the primary analysis.

8.2 Development Data

The development data were the publicly available Diabetes Health Indicators dataset, a cleaned file derived by its curator from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey of the US Centers for Disease Control and Prevention (CDC) [3], [4], [23]. The raw file contained 253,680 records. The outcome variable has three categories: 0 = no diabetes, 1 = prediabetes and 2 = diabetes. The 21 predictors were HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age (13 categories), Education (6 categories) and Income (8 categories). The outcome and all predictors are self-reported, and there were no missing values.

Exact duplicate records (n = 23,899; 9.4%) were removed, leaving 229,781 records. A further 1,834 feature combinations appeared with conflicting outcome labels. These are not exact duplicates, so they were retained, and they indicate some label noise. After cleaning, 82.71% of records were No Diabetes, 2.01% Prediabetes and 15.27% Diabetes (Table-1).

8.3 External Validation Data (Mumbai)

Primary data were collected from 111 residents of Mumbai using a structured questionnaire that captured the same 21 predictors as the development data, together with each participant's reported diabetes status (No Diabetes, Prediabetes or Diabetes). Responses were recoded into the BRFSS category codes before analysis, so that the model pipeline could be applied unchanged. The coding of all 21 predictors, with the corresponding Mumbai questionnaire response, is given in Appendix A (Table A-1).
Of the 111 participants, 109 reported a diabetes status. Data checks found no values outside the BRFSS ranges and no non-binary values in binary fields. Among the labeled participants, missing predictor values were HighChol (11), DiffWalk (6), PhysActivity (3), and one each for HighBP, CholCheck, Stroke and HeartDiseaseorAttack. The 93 participants with complete data (59 No Diabetes, 18

Prediabetes and 16 Diabetes) formed the primary analysis set, and all 109 labeled participants were used in the sensitivity analysis (Table-1). The Mumbai data were never used for training, model selection or hyperparameter tuning in the primary analysis.

Table-1: Class distribution of the datasets

Class distribution of the datasets				
Dataset	**n**	**No Diabetes**	**Prediabetes**	**Diabetes**
CDC cleaned	229,781	190,055	4,629	35,097
CDC training	183,824	152,043	3,703	28,078
CDC holdout	45,957	38,012	926	7,019
Mumbai, all labeled	109	72	19	18
Mumbai, complete cases	93	59	18	16

8.4 Preprocessing and Data Partition

All preprocessing was placed inside a single scikit-learn pipeline so that it was fitted on training data only. Seven continuous or ordinal predictors (BMI, MentHlth, PhysHlth, GenHlth, Age, Education and Income) were imputed with the median and standardized to zero mean and unit variance. The 14 binary predictors were imputed with the most frequent value and were not scaled. The pipeline was fitted on the CDC training set and then applied unchanged to the holdout and Mumbai data. Because the CDC data contained no missing values, imputation affected only the Mumbai sensitivity analysis. The cleaned CDC data were split by stratified random sampling on the three-class outcome into a training set (80%, n = 183,824) and a holdout test set (20%, n = 45,957), using a fixed random seed of 42.

8.5 Candidate Models and Model Selection

Five classical models were compared: logistic regression, random forest [24], extra trees [25], gradient boosting [26] and histogram-based gradient boosting. Logistic regression used balanced class weights and a maximum of 2,000 iterations. Random forest and extra trees each used 300 trees, a minimum leaf size of 5 and balanced class weights. Histogram-based gradient boosting used 200 iterations, a learning rate of 0.1, at most 31 leaf nodes, L2 regularization of 1.0 and balanced class weights. Gradient boosting used 100 trees, a learning rate of 0.1 and a maximum depth of 3. It does not accept a class-weight argument in scikit-learn and was therefore trained without class weighting. All other settings were the scikit-learn defaults, and no hyperparameter tuning was performed for the five candidate models.

The models were compared by three-fold stratified cross-validation on the training set (shuffled, random seed 42). The selection rule, fixed before the holdout or Mumbai data were evaluated, was the highest mean cross-validated macro F1-score, which gives equal weight to the three classes. Accuracy, balanced accuracy, macro precision and macro recall were also recorded. Per-class precision, recall and F1-score of every model were computed from out-of-fold predictions on the training set and, for reporting only, on the holdout set.

8.6 Evaluation Metrics

Because 82.7% of the development data belong to one class, overall accuracy is reported only alongside a majority-class baseline, which is the accuracy obtained by always predicting No Diabetes. The main metrics were balanced accuracy, defined as the mean of the three class recalls [27], macro F1-score, and per-class precision, recall and F1-score, supported by confusion matrices. For the large CDC holdout set, point estimates are reported. For the Mumbai sample, 95% confidence intervals for the overall metrics were obtained by non-parametric bootstrap with 2,000 resamples [28], and intervals for per-class recall were obtained by the Wilson method [29], because the bootstrap interval collapses to zero when no case of a class is detected.

8.7 Sensitivity, Secondary and Exploratory Analyses

A sensitivity analysis applied the frozen model to all 109 labeled Mumbai participants. Missing predictor values were imputed by the model pipeline using the median or most frequent value of the CDC training data, and not from the Mumbai data.

A secondary Prediabetes-focused analysis compared four random forest strategies (300 trees, minimum leaf size 5) using out-of-fold predictions on the CDC training set only. The strategies were balanced class weights versus custom weights (No Diabetes 1, Prediabetes 8, Diabetes 2.5), each with and without a Prediabetes probability threshold. Under a threshold, a participant was classified as Prediabetes when the predicted probability of Prediabetes reached the threshold, and otherwise as No Diabetes or Diabetes according to the larger probability. Thresholds from 0.05 to 0.50 in steps of 0.05 were compared and the best was chosen by out-of-fold macro F1-score. The strategy with the highest out-of-fold macro F1-score was then evaluated once on the CDC holdout and once on the Mumbai sample. Because it was applied after the primary Mumbai result had been seen, it is reported as secondary.

Three further analyses were run on the Mumbai sample after the primary result was known: a random forest with custom class weights, a hand-set Prediabetes probability threshold

of 0.15, and an extra trees model found by randomized search on the CDC training data. The search drew six random combinations from 100 or 150 trees, maximum depth of 4, 6 or 8, and minimum leaf size of 2 or 5, with balanced class weights and three-fold stratified cross-validation on macro F1-score. These three analyses are reported as exploratory and post-hoc and were not used for model selection.

9. DATA ANALYSIS: TOOLS AND TECHNOLOGIES

All data handling, modeling and evaluation were done in Python 3.13.16 in a single sequential notebook. The CDC data (CSV) and the Mumbai responses (Excel) were loaded, cleaned and checked using pandas 2.2.3, NumPy 2.1.3 and openpyxl 3.1.5. Preprocessing pipelines, the five classifiers, stratified data splitting, cross-validation and the performance metrics were implemented with scikit-learn 1.6.1 [22]. Bootstrap confidence intervals and Wilson intervals were computed with NumPy [28], [29]. Confusion-matrix figures were produced with Matplotlib 3.10.0 and seaborn 0.13.2, and the final fitted pipeline, including the preprocessing, was saved and reloaded with joblib 1.6.0 so that exactly the same model was applied to the Mumbai data. The main analytical techniques were a stratified 80/20 holdout split, three-fold stratified cross-validation with out-of-fold predictions, class weighting, probability thresholding for Prediabetes, majority-class baselines, and complete-case and imputation-based analyses. Results were exported as CSV tables and PNG figures.

10. MAJOR FINDINGS

The main findings of the study are as follows.

1. **Selected model.** Random forest had the highest cross-validated macro F1-score (0.441) and was selected. On the CDC holdout set it reached an accuracy of 0.762, a balanced accuracy of 0.489 and a macro F1-score of 0.448, which is above the chance level of 0.333 for balanced accuracy but modest.
2. **Prediabetes was almost undetectable.** The selected model identified only 1.1% of Prediabetes cases in the CDC holdout set. Logistic regression and histogram-based gradient boosting identified about 30% of cases, but with a precision of only about 0.03, roughly 1.6 times the 2.0% base rate.
3. **Accuracy was misleading.** Always predicting No Diabetes gave an accuracy of 0.827 on the CDC data. Gradient boosting reached 0.838 on the holdout set while predicting no Prediabetes cases and achieving a balanced accuracy of only 0.392.
4. **External validation was weak.** In the Mumbai sample (n = 93), the selected model achieved an accuracy of 0.677 (95% CI 0.581 to 0.774) against a majority-class baseline of 0.634, a balanced accuracy of 0.432 (0.354 to 0.516) and a macro F1-score of 0.404 (0.304 to 0.489). None of the 18 participants with Prediabetes was identified, and 5 of the 16 with Diabetes were identified.
5. **Other analyses did not change the conclusion.** Using all 109 labeled participants gave an accuracy of 0.697 (baseline 0.661) and a macro F1-score of 0.399. The Prediabetes-focused strategy chosen on the CDC data performed worse externally (macro F1-score 0.356).
6. **The two samples differed strongly.** Mumbai participants were younger (mean age category 3.8 versus 8.1), reported less high blood pressure (20% versus 45%) and included a much larger share of Prediabetes (19.4% versus 2.0%).
7. **An exploratory model scored higher but is not evidence of better performance.** A post-hoc extra trees model reached a macro F1-score of 0.559 in Mumbai. It was not the pre-selected model, had a lower cross-validated score, and its confidence interval overlaps that of the primary result.

The outcome of each hypothesis is summarized in Table-2.

Table-2: Outcome of the research hypotheses

Outcome of the research hypotheses		
Hypothesis	**Result**	**Evidence**
Hypothesis 1: Better than chance	H_0 rejected (modest effect)	Balanced accuracy 0.489 on the CDC holdout against 0.333 for chance
Hypothesis 2: Prediabetes F1 lower than Diabetes F1	H_0 rejected	Prediabetes F1 0.000 to 0.059 against Diabetes F1 0.297 to 0.472 for every model on the holdout set
Hypothesis 3: Lower balanced accuracy in Mumbai	Directionally supported, not conclusive	0.432 (95% CI 0.354 to 0.516) against 0.489; the interval includes 0.489. Diabetes recall (0.31; 0.14 to 0.56) was clearly below the holdout value of 0.656
Hypothesis 4: Accuracy above the baseline	H_0 not rejected	0.677 (0.581 to 0.774) against 0.634; the interval includes the baseline

11. RESULTS AND ANALYSIS

11.1 Participants and Class Distribution

After the removal of duplicates, the CDC development data contained 229,781 respondents. In Mumbai, 111 participants were recruited, 109 reported a diabetes status and 93 had complete data. Prediabetes made up 2.0% of the development data but 19.4% of the Mumbai primary set (Table-1).

11.2 Model Comparison on CDC Data

In three-fold cross-validation on the training set, random forest had the highest macro F1-score (0.441) and was selected, ahead of extra trees (0.432), logistic regression (0.421), histogram-based gradient boosting (0.416) and gradient boosting (0.396) (Table-3). Gradient boosting had the highest accuracy (0.835), which is no better than the 0.827 obtained by always predicting No Diabetes, and the lowest balanced accuracy (0.387), because it identified no Prediabetes cases. Logistic regression and histogram-based gradient boosting had the highest balanced accuracy (0.509 and 0.507) and the highest Prediabetes recall (0.303 and 0.282), but their Prediabetes precision was only about 0.03. No Diabetes was classified well by every model (F1-score 0.746 to 0.908), largely because it makes up 82.7% of the data. Full per-class cross-validation results are given in Appendix B (Table B-1).

Table-3: Overall performance of the models (CV = cross-validation on the training set; Test = CDC holdout)

Overall performance of the models						
Model	**CV Acc.**	**CV Bal.**	**CV F1**	**Test Acc.**	**Test Bal.**	**Test F1**
LR	0.624	0.509	0.421	0.629	0.514	0.426
RF	0.758	0.483	0.441	0.762	0.489	0.448
ET	0.703	0.489	0.432	0.708	0.494	0.436
HGB	0.610	0.507	0.416	0.618	0.516	0.422
GB	0.835	0.387	0.396	0.838	0.392	0.402
Baseline	0.827	0.333	0.302	0.827	0.333	0.302

LR = logistic regression, RF = random forest (selected), ET = extra trees, HGB = histogram gradient boosting, GB = gradient boosting (no class weights), Baseline = always No Diabetes. Bal. = balanced accuracy, F1 = macro F1-score.

On the CDC holdout set, the selected random forest achieved an accuracy of 0.762, a balanced accuracy of 0.489 and a macro F1-score of 0.448. Recall was 0.800 for No Diabetes, 0.656 for Diabetes and only 0.011 for Prediabetes (10 of 926 cases identified; precision 0.043). The ranking of the models on the holdout set matched the cross-validation ranking. These holdout results were computed after model selection, for reporting only. The per-class holdout results most relevant to Prediabetes are in Table-4, and the complete results are in Appendix B (Table B-2).

Table-4: Per-class results on the CDC holdout set (focused on Prediabetes)

Per-class results on the CDC holdout set (focused on Prediabetes)					
Model	**No Diab. F1**	**Prediab. P**	**Prediab. R**	**Prediab. F1**	**Diab. F1**
LR	0.765	0.033	0.304	0.059	0.453
RF	0.853	0.043	0.011	0.017	0.472
ET	0.813	0.032	0.053	0.040	0.455
HGB	0.753	0.032	0.301	0.058	0.455
GB	0.910	0.000	0.000	0.000	0.297
Baseline	0.905	0.000	0.000	0.000	0.000

P = precision, R = recall.

In the Prediabetes-focused analysis on the CDC training data, adding a tuned probability threshold to a balanced random forest raised the out-of-fold Prediabetes recall from 0.006 to 0.066. The best strategy by macro F1-score was a random forest with custom class weights and a Prediabetes threshold of 0.20 (macro F1-score 0.449; Prediabetes recall 0.026). On the holdout set this strategy gave an accuracy of 0.811, a macro F1-score of 0.453, and Prediabetes precision, recall and F1-score of 0.031, 0.024 and 0.027. The gains in Prediabetes detection were small in every strategy (Table-5).

Table-5: Prediabetes-focused random forest strategies (out-of-fold, CDC training set)

Prediabetes-focused random forest strategies (out-of-fold, CDC training set)					
Strategy	**Acc.**	**Bal. acc.**	**Macro F1**	**Prediab. Recall**	**Prediab. F1**
Balanced weights, argmax	0.758	0.483	0.441	0.006	0.009
Balanced weights,Threshold 0.25	0.737	0.484	0.447	0.066	0.043
Custom weights, argmax	0.816	0.443	0.442	0.000	0.000
Custom weights, threshold	0.807	0.445	0.449	0.026	0.030

0.20 (selected)					

11.3 External Validation in Mumbai

The frozen random forest was applied to the 93 complete cases. The accuracy was 0.677 (95% CI 0.581 to 0.774), only four percentage points above the majority-class baseline of 0.634. The balanced accuracy was 0.432 (0.354 to 0.516) and the macro F1-score was 0.404 (0.304 to 0.489). Of the 59 participants without diabetes, 58 were classified correctly (recall 0.983). None of the 18 participants with Prediabetes was identified (recall 0; Wilson 95% CI 0 to 0.18): 15 were classified as No Diabetes and 3 as Diabetes. Of the 16 participants with Diabetes, 5 were identified (0.31; 0.14 to 0.56) and 11 were classified as No Diabetes (Fig -1). The sensitivity analysis on all 109 labeled participants gave similar results (accuracy 0.697 against a baseline of 0.661; macro F1-score 0.399). The Prediabetes-focused strategy chosen on the CDC data did not improve external performance (macro F1-score 0.356; 3 of 16 Diabetes cases identified) (Table-6).

Table-6: External validation results on the Mumbai sample

External validation results on the Mumbai sample						
Analysis	**n**	**Acc. (95% CI)**	**Bal. acc. (95% CI)**	**Macro F1 (95% CI)**	**Predi. Recall (Wilson 95% CI)**	**Diab. Recall (Wilson 95% CI)**
Primary: random forest, complete cases	93	0.677 (0.581 to 0.774)	0.432 (0.354 to 0.516)	0.404 (0.304 to 0.489)	0/18 = 0.00 (0.00 to 0.18)	5/16 = 0.31 (0.14 to 0.56)
Sensitivity: all labeled	109	0.697 (0.615 to 0.780)	0.421 (0.356 to 0.500)	0.399 (0.310 to 0.480)	0/19 = 0.00 (0.00 to 0.17)	5/18 = 0.28 (0.12 to 0.51)
Secondary: custom weights, threshold 0.20	93	0.656 (0.548 to 0.753)	0.390 (0.328 to 0.465)	0.356 (0.263 to 0.447)	0/18 = 0.00 (0.00 to 0.18)	3/16 = 0.19 (0.07 to 0.43)
Majority-class baseline (primary set)	93	0.634	0.333	N/A	0	0

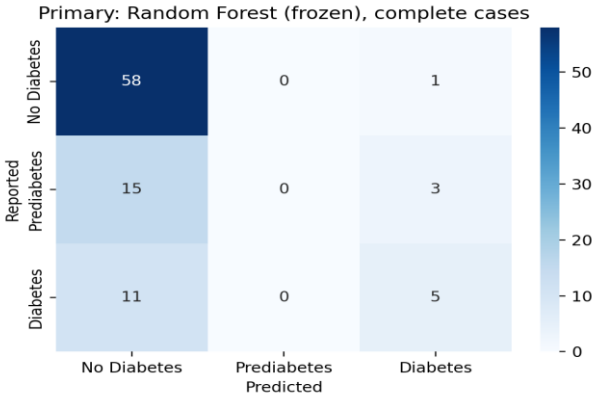


Fig -1: Confusion matrix of the primary Mumbai external validation

11.4 Difference Between the CDC and Mumbai Samples

The Mumbai sample was much younger (mean age category 3.8, about 30 to 39 years, against 8.1, about 55 to 59 years) and included more men (63.4% against 43.9%). Participants reported much less high blood pressure (20.4% against 45.4%) and high cholesterol (19.4% against 44.2%), and fewer had had a cholesterol check in the past five years (66.7% against 96.0%), eaten fruit daily (26.9% against 61.3%) or had health coverage (71.0% against 94.6%). The mean BMI was lower (26.6 against 28.7 kg/m^2) and the mean number of poor mental health days was higher (7.8 against 3.5) (Table-7).

Table-7: Predictor means in the CDC training set and the Mumbai complete cases

Predictor means in the CDC training set and the Mumbai complete cases		
Predictor	**CDC training**	**Mumbai**
Age category (1 to 13)	8.09	3.81
Sex	0.439	0.634
BMI (kg/m2)	28.69	26.56
HighBP	0.455	0.204
HighChol	0.442	0.194
CholCheck	0.960	0.667
Fruits	0.613	0.269
Veggies	0.795	0.538
AnyHealthcare	0.946	0.710
MentHlth (Days)	3.50	7.79
Income category (1 to 8)	5.89	5.07
Education category (1 to 6)	4.98	5.52

11.5 Exploratory Post-hoc Analyses

These analyses were run after the primary Mumbai result was known and are reported for transparency only. Re-weighting the classes of the random forest gave an accuracy of 0.667 and a macro F1-score of 0.363. A hand-set Prediabetes threshold of 0.15 gave an accuracy of 0.667 and

a macro F1-score of 0.371, and identified 1 of the 18 participants with Prediabetes. An extra trees model found by randomized search on the CDC training data (100 trees, minimum leaf size 5, maximum depth 8) reached an accuracy of 0.731 (95% CI 0.634 to 0.817), a balanced accuracy of 0.544 (0.439 to 0.654) and a macro F1-score of 0.559 (0.425 to 0.675), identifying 3 of 18 Prediabetes cases and 8 of 16 Diabetes cases. This configuration had a cross-validated macro F1-score of 0.421 on the CDC data, below the 0.441 of the random forest selected in advance, so the pre-specified rule would not have chosen it. Its confidence interval overlaps that of the primary result, and the five-point difference in accuracy corresponds to about five participants. These results should therefore not be interpreted as evidence of better external performance.

11.6 Why Prediabetes Is Hard to Detect

Several factors plausibly explain the poor Prediabetes detection. First, the predictors change in an ordered way across the three classes, with Prediabetes lying between the other two. High blood pressure was reported by 39.5%, 62.9% and 75.2% of respondents with No Diabetes, Prediabetes and Diabetes, and the mean BMI was 28.0, 30.7 and 32.0 kg/m^2. A class that lies between two others on most features has little region of its own for a classifier to find. Second, label quality matters. The BRFSS outcome records what respondents report having been told, and most people with prediabetes have never been told: only about 11% of US adults with prediabetes were aware of it in national survey data [2]. Many undiagnosed cases are therefore labeled No Diabetes, which adds noise to the Prediabetes class, and the 1,834 feature combinations with conflicting labels in the cleaned data are consistent with this. Third, because Prediabetes is only 2.0% of the data, even the models with the highest recall (about 30%) had a precision of about 0.03, roughly 1.6 times the base rate, so the extra recall came with about 30 false alarms for each true case. Class weighting and probability thresholds move a model along this recall and precision trade-off, as is expected for imbalanced learning [6], [11], but cannot add information that the predictors do not contain (Table-5).

The model-selection rule also affected the outcome. Selecting by cross-validated macro F1-score chose the random forest, which almost never predicted Prediabetes, whereas a rule that prioritized Prediabetes recall would have chosen logistic regression. Such a rule is a legitimate design choice for future work, but it was not tested on the Mumbai data, to avoid selecting a model after seeing the external result.

11.7 Why External Performance Was Poor

The fall in performance in Mumbai is consistent with population shift as well as chance. Comparing the case mix of the development and validation samples is the first step in interpreting external validation [19]. The Mumbai sample was far younger, reported much less high blood pressure and cholesterol, and had fewer recent cholesterol checks (Table-7). Age and blood pressure are among the strongest predictors in survey-based models, so a model that learned these relationships in an older US population is expected to under-predict disease in a young sample. The share of Prediabetes also differed (19.4% against 2.0%), and a model trained with a 2% prior rarely predicts the class; poor calibration of this kind is a known weakness of prediction models when they are moved to a setting with a different disease frequency [20]. A substantial fall in discrimination or calibration in new data is common even for validated models [14]. The Mumbai class proportions belong to this convenience sample and should not be read as prevalence estimates for the city; population estimates come from national surveys [1].

Body composition is a further source of shift. A WHO expert consultation found that the risk of type 2 diabetes in Asian populations rises at a lower BMI than in European populations and proposed action points of 23.0 and 27.5 kg/m^2 [8]. The mean BMI of the Mumbai sample (26.6) was lower than the CDC mean (28.7) but already above 23.0, so a model that learned a US BMI and risk relationship would understate risk for these participants. The Indian Diabetes Risk Score uses waist circumference and family history alongside age and physical activity [9], and neither of those two is among the 21 BRFSS indicators used here, which limits what any BRFSS-trained model can do in this setting.

11.8 Interpreting Accuracy

A classifier that always predicts No Diabetes would score 0.827 on the CDC data and 0.634 on the Mumbai complete cases. On the CDC data, gradient boosting reached 0.835 (0.838 on the holdout set) with a balanced accuracy of 0.387 to 0.392 and no Prediabetes predictions. Accuracy close to the baseline therefore says little about whether a model is useful for screening. Balanced accuracy, macro-averaged and per-class measures, reported with an explicit baseline, are more informative for a three-class problem [13], and precision should be reported together with recall for a rare class [18]. For data with this degree of imbalance, a fixed overall-accuracy target is not a meaningful success criterion, because the model with the highest accuracy here was also the one that ignored Prediabetes completely.

12. CONCLUSION

Five classical machine learning models were developed and compared to predict diabetes status among three classes (No Diabetes, Prediabetes and Diabetes) using 21 self-reported indicators from the CDC BRFSS data and tested on an independent sample of participants from Mumbai.

For the first objective, the models separated people with and without diabetes only moderately well. The selected random forest reached a balanced accuracy of 0.489 and a macro F1-

score of 0.448 on the CDC holdout set, above chance but modest. For the second objective, no model detected Prediabetes usefully: the random forest identified 1.1% of cases, and the models with the highest recall (about 30%) had a precision of only about 0.03. Overall accuracy was a misleading measure because the model with the highest accuracy (gradient boosting, 0.838) predicted no Prediabetes cases and scored no better than always predicting No Diabetes (0.827).

For the third objective, the frozen random forest did not transport to the Mumbai sample. Its accuracy (0.677) was only four percentage points above the majority-class baseline (0.634), its balanced accuracy was 0.432, none of the 18 participants with Prediabetes was identified, and 5 of the 16 with Diabetes were identified. Sensitivity and secondary analyses did not change this conclusion. The Mumbai sample differed strongly from the training data in age, blood pressure, cholesterol screening and the share of Prediabetes, which is consistent with population shift.

Overall, the findings indicate that a model that was designed to detect Prediabetes using data from surveys in the USA should not be applied to an Indian population and that these self-reported measures alone are not sufficient for the reliable detection of Prediabetes. Research of this type should include balanced and per class measures and provide baseline and confidence interval values and should not be based on accuracy.

13. LIMITATIONS

The findings should be read with the following limitations.

1. **Small external sample.** The Mumbai primary set had 93 participants, with 16 to 18 in each minority class, so every external estimate is imprecise (for example, Diabetes recall 0.31, 95% CI 0.14 to 0.56).
2. **Self-reported labels.** Diabetes status was self-reported in both datasets, so undiagnosed diabetes and prediabetes are misclassified as No Diabetes [2].
3. **Convenience sampling.** The Mumbai participants were young and may not represent the city or India, and their class proportions should not be read as prevalence estimates.
4. **Recoding and measurement differences.** Mumbai age, education and income responses were recoded to US-based BRFSS categories, which may introduce measurement differences between the datasets.
5. **Complete-case primary analysis.** Sixteen labeled participants were excluded from the primary analysis because of missing predictors; the sensitivity analysis only partly addresses this.
6. **Post-hoc analyses.** The secondary and exploratory analyses used the Mumbai data after the primary result was known and are open to selection bias.
7. **Model scope and comparison.** Only classical models with limited hyperparameter search were studied. Gradient boosting was trained without class weights, unlike the other four models, so its comparison is not like for like.
8. **Validation design.** One random 80/20 split of a single survey year was used, with three-fold cross-validation, and the calibration of the predicted probabilities was not assessed [20].

14. FUTURE SCOPE OF RESEARCH

These findings suggest a number of directions for future research.

1. **Laboratory-confirmed outcomes.** If the same analysis were to be repeated using the labels HbA1c or fasting glucose, then the issue of undiagnosed Prediabetes would not be present in the self-reported data [2].
2. **More multi-site Indian samples** of adequate size to provide stable estimates within each minority class.
3. **India-relevant predictors.** The inclusion of waist circumference and family history, both integral to the Indian Diabetes Risk Score [9] combined with Asian BMI action points [8] may enhance the predictions made in this context.
4. **Model updating and recalibration** for the Indian context, where calibration curves are published after obtaining
5. **A pre-specified, Prediabetes-weighted selection rule,** for example choosing the model by Prediabetes recall at a minimum precision, with cost-sensitive evaluation [11] and the rule fixed before any external data are seen.
6. **Replication of exploratory findings,** such as the extra trees result, in a new sample with a pre-registered analysis.
7. **Comparison with established screening tools,** such as the Indian Diabetes Risk Score [9], and follow-up studies of who progresses from Prediabetes to Diabetes.

ACKNOWLEDGEMENT

I am thankful to Prof. Supriya Kamale, Prof. Pranjali Patil and Prof. Prachi Surve for their mentorship and support during the course of this work. I thank all the participants from Mumbai who took the time to share their responses for this study, and D.Y. Patil Deemed to be University, Belapur, for providing the environment to carry out this work. Finally, I am deeply grateful to my family for their constant support and patience, and I warmly welcome the newest member of our family, Samar.

REFERENCES

1) 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.

2) Centers for Disease Control and Prevention, "Awareness of prediabetes - United States, 2005-2010," MMWR Morb. Mortal. Wkly. Rep., vol. 62, no. 11, pp. 209-212, Mar. 2013.

3) A. Teboul, "Diabetes Health Indicators Dataset," Kaggle (derived from the CDC BRFSS 2015 survey). [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset Accessed: Oct. 25, 2026.

4) Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System (BRFSS)." [Online]. Available: https://www.cdc.gov/brfss/ Accessed: Oct. 25, 2026.

5) Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," Prev. Chronic Dis., vol. 16, art. 190109, 2019.

6) H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263-1284, Sep. 2009.

7) E. W. Steyerberg and F. E. Harrell Jr., "Prediction models need appropriate internal, internal-external, and external validation," J. Clin. Epidemiol., vol. 69, pp. 245-247, 2016, doi:10.1016/j.jclinepi.2015.04.005.

8) WHO Expert Consultation, "Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies," Lancet, vol. 363, no. 9403, pp. 157-163, 2004.

9) 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.

10) Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104-116, 2017.

11) H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," BMC Endocr. Disord., vol. 19, art. 101, 2019, doi:10.1186/s12902-019-0436-6.

12) H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," Int. J. Environ. Res. Public Health, vol. 18, no. 6, art. 3317, 2021, doi:10.3390/ijerph18063317.

13) M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427-437, 2009, doi:10.1016/j.ipm.2009.03.002.

14) G. S. Collins et al., "Evaluation of clinical prediction models (part 1): from development to external validation," BMJ, vol. 384, art. e074819, 2024, doi:10.1136/bmj-2023-074819.

15) Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Front. Genet., vol. 9, art. 515, 2018, doi:10.3389/fgene.2018.00515.

16) A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, and D. Khalili, "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes," Med. Decis. Making, vol. 36, no. 1, pp. 137-144, 2016, doi:10.1177/0272989X14560647.

17) S. B. Choi et al., "Screening for prediabetes using machine learning models," Comput. Math. Methods Med., vol. 2014, art. 618976, 2014, doi:10.1155/2014/618976.

18) T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, art. e0118432, 2015, doi:10.1371/journal.pone.0118432.

19) T. P. A. Debray et al., "A new framework to enhance the interpretation of external validation studies of clinical prediction models," J. Clin. Epidemiol., vol. 68, no. 3, pp. 279-289, 2015, doi:10.1016/j.jclinepi.2014.06.018.

20) B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," BMC Med., vol. 17, art. 230, 2019, doi:10.1186/s12916-019-1466-7.

21) G. S. Collins, K. G. M. Moons, P. Dhiman et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, art. e078378, 2024, doi:10.1136/bmj-2023-078378.

22) F. Pedregosa et al., "Scikit-learn: machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

23) "CDC Diabetes Health Indicators," UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/dataset/891 Accessed: Oct. 25, 2026.

24) L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

25) P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn., vol. 63, no. 1, pp. 3-42, 2006.

26) H. Friedman, "Greedy function approximation: a gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189-1232, 2001.

27) H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in Proc. 20th Int. Conf. Pattern Recognit. (ICPR), 2010, pp. 3121-3124.

28) B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.

29) E. B. Wilson, "Probable inference, the law of succession, and statistical inference," J. Am. Stat. Assoc., vol. 22, no. 158, pp. 209-212, 1927.

APPENDIX A

Table A-1: BRFSS coding of the 21 predictors and the corresponding Mumbai recoding

Sr. No.	Predictor	BRFSS coding	Mumbai recoding
1.	HighBP	0 = no high blood pressure; 1 = high blood pressure	0 = no high blood pressure, high blood pressure only during pregnancy, borderline/prehypertensive; 1 = yes
2.	HighChol	0 = no high cholesterol; 1 = high cholesterol	0 = no high cholesterol; 1 = high cholesterol
3.	CholCheck	0 = no cholesterol check in past 5 years; 1= check in past 5 years	0 = no cholesterol check in past 5 years; 1= check in past 5 years
4.	BMI	Body mass index (continuous)	Derived mathematically from participant's reported height and weight
5.	Smoker	0 = fewer than 100 cigarettes in lifetime; 1 = at least 100	0 = not at all; 1 = Has smoked 100 cigarettes, smokes every day, smokes some days
6.	Stroke	0 = never told had a stroke; 1 = told had a stroke	0 = never told had a stroke; 1 = told had a stroke
7.	HeartDiseaseorAttack	0 = no; 1 = coronary heart disease or myocardial infarction	0 = no; 1 = coronary heart disease or myocardial infarction
8.	PhysActivity	0 = no physical activity in past 30 days (outside work); 1 = yes	0 = no physical activity in past 30 days (outside work); 1 = yes
9.	Fruits	0 = fruits less than once a day; 1 = once a day or more	0 = never, less than once a week, 1-3 times per week, 4-6 times per week; 1 = once a day, more than once a day
10.	Veggies	0 = vegetables less than once a day; 1 = once a day or more	0 = never, less than once a week, 1-3 times per week, 4-6 times per week; 1 = once a day, more than once a day
11.	HvyAlcoholConsump	0 = no; 1 = heavy drinker (men more than 14 drinks/week, women more than 7)	0 = does not consume alcohol, 1-7 drinks/week, 8-14 drinks/week(only men); 1 = 8-14 drinks/week(only women), 15-21 drinks/week, 22-28 drinks/week, more than 28 drinks per week
12.	AnyHealthcare	0 = no health coverage; 1 = any health coverage	0 = no health coverage; 1 = any health coverage
13.	NoDocbcCost	0 = no; 1 = could not see a doctor in past 12 months because of cost	0 = no; 1 = could not see a doctor in past 12 months because of cost
14.	GenHlth	1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor	1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor
15.	MentHlth	Days of poor mental health in past 30 days (0 to 30)	Days of poor mental health in past 30 days (0 to 30)
16.	PhysHlth	Days of poor physical health in past 30 days (0 to 30)	Days of poor physical health in past 30 days (0 to 30)

17.	DiffWalk	0 = no; 1 = serious difficulty walking or climbing stairs	0 = no; 1 = serious difficulty walking or climbing stairs
18.	Sex	0 = female; 1 = male	0 = female; 1 = male
19.	Age	1 = 18 to 24; 2 = 25 to 29; 3 = 30 to 34; 4 = 35 to 39; 5 = 40 to 44; 6 = 45 to 49; 7 = 50 to 54; 8 = 55 to 59; 9 = 60 to 64; 10 = 65 to 69; 11 = 70 to 74; 12 = 75 to 79; 13 = 80 or older (years)	1 = 18 to 24; 2 = 25 to 29; 3 = 30 to 34; 4 = 35 to 39; 5 = 40 to 44; 6 = 45 to 49; 7 = 50 to 54; 8 = 55 to 59; 9 = 60 to 64; 10 = 65 to 69; 11 = 70 to 74; 12 = 75 to 79; 13 = 80 or older (years)
20.	Education	1 = never attended or kindergarten only; 2 = grades 1 to 8; 3 = grades 9 to 11; 4 = grade 12 or equivalent; 5 = some college or technical school; 6 = college graduate	1 = no formal education; 2 = primary school(up to class 5), middle school(classes 6-8); 3 = secondary school (classes 9-10); 4 = higher secondary / junior college(11-12); 5 = diploma/ITI/polytechnic; 6 = bachelors degree, postgraduate or higher
21.	Income	1 = less than $10,000; 2 = $10,000 to under $15,000; 3 = $15,000 to under $20,000; 4 = $20,000 to under $25,000; 5 = $25,000 to under $35,000; 6 = $35,000 to under $50,000; 7 = $50,000 to under $75,000; 8 = $75,000 or more (US dollars, annual)	1 = less than 1 lakh per year; 2 = 1 lakh to 2.5 lakh; 3 = 2.5 lakh to 5 lakh; 4 = 5 lakh to 7.5 lakh; 5 = 7.5 lakh to 10 lakh; 6 = 10 lakh to 15 lakh; 7 = 15 lakh to 25 lakh; 8 = More than 25 lakh

APPENDIX B

Table B-1: Out-of-fold performance on the CDC training set (three-fold cross-validation)

Metric	LR	RF	ET	HGB	GB	Base
Accuracy	0.624	0.758	0.703	0.610	0.835	0.827
Balanced Accuracy	0.509	0.483	0.489	0.507	0.387	0.333
Macro F1	0.421	0.441	0.432	0.416	0.396	0.302
No Diabetes Precision	0.941	0.912	0.924	0.945	0.850	0.827
No Diabetes Recall	0.639	0.796	0.721	0.616	0.974	1.000
No Diabetes F1	0.761	0.850	0.810	0.746	0.908	0.905
Prediabetes Precision	0.032	0.026	0.032	0.030	0.000	0.000
Prediabetes Recall	0.303	0.006	0.053	0.282	0.000	0.000

Prediabetes F1	0.059	0.009	0.040	0.054	0.000	0.000
Diabetes Precision	0.358	0.362	0.329	0.350	0.553	0.000
Diabetes Recall	0.584	0.648	0.692	0.624	0.187	0.000
Diabetes F1	0.444	0.464	0.446	0.448	0.279	0.000

LR = logistic regression, RF = random forest (selected), ET = extra trees, HGB = histogram-based gradient boosting, GB = gradient boosting (trained without class weights), Base = always predicting No Diabetes., F1 = F1-score.

Table B-2: Performance of all models on the CDC holdout set

Metric	**LR**	**RF**	**ET**	**HGB**	**GB**	**Base**
Accuracy	0.629	0.762	0.708	0.618	0.838	0.827
Balanced Accuracy	0.514	0.489	0.494	0.516	0.392	0.333
Macro F1	0.426	0.448	0.436	0.422	0.402	0.302
No Diabetes Precision	0.943	0.915	0.927	0.949	0.852	0.827
No Diabetes Recall	0.644	0.800	0.725	0.624	0.977	1.000
No Diabetes F1	0.765	0.853	0.813	0.753	0.910	0.905
Prediabetes Precision	0.033	0.043	0.032	0.032	0.000	0.000
Prediabetes Recall	0.304	0.011	0.053	0.301	0.000	0.000
Prediabetes F1	0.059	0.017	0.040	0.058	0.000	0.000
Diabetes Precision	0.367	0.368	0.337	0.358	0.590	0.000
Diabetes Recall	0.593	0.656	0.704	0.623	0.198	0.000

Diabetes F1	0.453	0.472	0.455	0.455	0.297	0.000

LR = logistic regression, RF = random forest (selected), ET = extra trees, HGB = histogram-based gradient boosting, GB = gradient boosting (trained without class weights), Base = always predicting No Diabetes., F1 = F1-score.