FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 23, December, 2026, pp 432 – 439
DOI: https://doi.org/10.33003/fjs-2026-1023-5807

Machine Learning for Diabetes Prediction: A Systematic Review and Implications for Nigeria

[*1]Alhaji Abdullahi Gwani, [2]Siok Kun Sek, [1]Usman Waziri

[1]Department of Mathematical Sciences Saadu Zungur University Gadau Bauchi State Nigeria
[2]School of Mathematical Sciences, Universiti Sains Malaysiam Penang Malaysia

*Corresponding authors' email: aagwani1@gmail.com

Received: 25 Sept. 2026
Accepted: 17 Oct. 2026
Published: 20 Nov. 2026

ABSTRACT

Machine learning (ML) is widely proposed for early diabetes prediction, but the reliability of reported performance is uncertain. This systematic review describes the data sources, algorithms, validation designs and reporting practices of recent ML diabetes prediction studies and draws implications for Nigeria and other low-resource settings. PubMed was searched in October 2026 for studies published between 2019 and 2026 yielding 570 records; nine additional records were identified from other sources including google scholar. All records were screened, and 17 studies met the inclusion criteria. Data on data source, sample size, algorithms, validation, performance, a study reported more than one algorithm. seven reporting items were extracted and synthesised. Gradient boosting (12 studies), random forests (11), regression models (9) were the most frequent algorithms, and four studies used the Pima Indians Diabetes Database. Reported area under the curve ranged from 0.73 to 0.996 (median 0.85), with the highest values in large public datasets where glycaemic variables dominated predictions. Explainability was reported in 10 studies, yet only one evaluated model in an independent population, one reported calibration and one assessed fairness. One study used African data (Rwanda) and none used Nigerian data. Reported accuracy therefore, probably overstates real-world performance. Priorities include representative Nigerian datasets, leakage-free preprocessing, external validation, calibration and fairness assessment, and models that avoid diagnostic variables when early detection is the aim. Findings are limited by single-database searching and partial screening.

Keywords: Diabetes Mellitus; Machine Learning; Risk Prediction; Systematic Review; Model Validation; Nigeria.

INTRODUCTION

Diabetes mellitus is a chronic metabolic disease whose complications impose a heavy burden on patients and health systems. Detecting people at high risk before diagnosis allows lifestyle and clinical interventions to begin early, which is why risk prediction is an active research field. In Nigeria the burden appears to be growing. A national survey in 1992 reported a prevalence of 2.2%, and a later systematic review and meta-analysis pooled a prevalence of 5.77%, with the south-south zone the most affected (Uloko et al., 2018). A more recent meta-analysis estimated the pooled prevalence of type 2 diabetes at 7.0%, with 11.35% in the south-south zone and 2.03% in the north-central zone, and observed that most contributing studies came from southern zones (Olamoyegun et al., 2024).

Machine learning (ML) has received much interest as it can learn risk patterns from frequently acquired data. Initial evaluations of the field showed that supervised methods were dominant in diabetes research (Kavakiotis et al., 2017). More recent surveys have charted three decades of machine learning and artificial intelligence research on type 2 diabetes prediction (Kiran et al., 2025) and explored the datasets, algorithms and metrics applied, pointing to the need of ethical considerations and interdisciplinary collaboration (Khokhar et al., 2024). Focused reviews consistently report promising discrimination but little evidence of real-world application. A meta-analysis of models to predict progression from prediabetes to type 2 diabetes revealed pooled C-indices of 0.76 in training sets and 0.84 in validation sets, but concluded that external validation was missing (Wang et al., 2026). A scoping analysis of ML models for prediction of type 2 diabetes risk in primary care identified only four qualifying papers and insufficient information regarding external validation, calibration or prospective application (Ravi & Olickal, 2026). A systematic assessment of artificial intelligence models for diabetic retinopathy risk showed AUC values from 0.68 to 0.97 in eight research, however external validation was predominantly absent (Teruna et al., 2026).

These findings are important since high reported accuracy does not necessarily mean clinical usefulness. A systematic review of clinical prediction models did not find any performance benefit of ML over logistic regression (Christodoulou et al., 2019) and methodological standards such as TRIPOD, PROBAST and sample-size guidance have been published to improve development and reporting (Collins et al., 2015; Moons et al., 2019; Riley et al., 2020). Calibration, a measure of whether anticipated risks agree with observed hazards, is frequently overlooked (Van Calster et al., 2019). Information leakage between training and test data can inflate performance (Kaufman et al., 2012).

This systematic review aimed to evaluate the data sources, algorithms, validation and reporting practices of recent ML-based diabetes prediction studies and to examine their implications for Nigeria. It has been observed that there have been few assessments of the extent to which contemporary diabetes prediction studies meet these conditions, how much their stated effectiveness rely on glycaemic predictors that are inherently diagnostic, or what the evidence means for sub-Saharan Africa. The systematic review therefore answers four questions: (RQ1) which data sources and sample sizes are utilised; (RQ2) which algorithms are employed; (RQ3) what

performance is stated and what drives it; and (RQ4) how often validation, calibration, explainability and fairness are documented. We next explore consequences for Nigeria.

MATERIALS AND METHODS

Design and reporting

This systematic review was conducted in accordance with the PRISMA 2020 guidelines (Page et al., 2021). Formal ethical approval was not required as this study utilized secondary, publicly available literature and did not involve human or animal participants.

Eligibility Criteria

Eligibility criteria are summarised in Table 1. The review targeted studies that develop or evaluate ML models to predict or detect diabetes in general or screening populations. Studies of gestational diabetes and studies predicting outcomes or complications among people who already have diabetes were excluded, because their predictors and clinical use differ.

Table 1: Eligibility Criteria.

Element	Criterion
Population	Adults or mixed-age people without known diabetes at baseline, or general datasets labelled for diabetes status
Index models	Machine learning or deep learning classifiers (for example random forest, gradient boosting, support vector machine, neural network)
Comparators	Any, including logistic regression or other ML models
Outcome	Diagnosis of existing (undiagnosed) diabetes, or prediction of incident type 1, type 2 or unspecified diabetes
Study design	Primary model development or validation studies reporting at least one performance metric
Limits	Journal articles with English language, published 2019–2026
Exclusions	Gestational diabetes; complications or outcomes among people with diabetes; commentaries, letters, editorials, errata, retractions and reviews

Information Sources and Search

PubMed was searched in October 2026 for peer-reviewed journal articles with record containing the keywords diabetes and machine learning in the title, alongside title terms for predict, detect, diagnoses, or classify, while excluding reviews, comments, letters, editorials, published errata, and studies with retinopathy or complications in the title, restricted to 2019 to 2026. It returned 570 records. Two broader preliminary searches were used to refine the string; eligible studies encountered during those searches (n = 9) were retained and reported as records from other sources. The 9 records were found in reference lists and other scholarly sites (including SPRINGER *scientific reports* and others) in addition to PubMed. These records were evaluated alongside PubMed records for relevancy. First, duplicate entries were deleted, then title and abstract screening, then full-text review of suitable research. To reduce duplication, improve evidence identification, and assure systematic and reproducible research selection, this technique was used.

Selection Process

Titles and abstracts and full text of all 570 PubMed records, together with the nine records from other sources, were screened against the criteria in Table 1. Screening was limited to the 24 database records for relevance by PubMed, plus the nine records from other sources including SPRINGER *scientific reports* and others Full texts were retrieved, so all judgements rest on double reviewers, titles, abstracts and indexing data and finally the full text. This restriction is a major limitation and is discussed in Section 3.9.

Data Extraction and Reporting Appraisal

For each included study the following were extracted: data source and sample size, algorithms, validation design, headline performance and the data source category. Seven reporting items adapted from Transparent Reporting of a multivariable prediction model for Individual Prognosis (Collins et al., 2015; Moons et al., 2019) were coded as reported in the abstract and later full text of the extracted record, validation in an independent population, calibration, explainability, explicit control of data leakage, handling of class imbalance, quantification of uncertainty or statistical testing, and resampling-based validation (cross-validation or bootstrap).

Synthesis

Studies were synthesised narratively and with descriptive counts. Meta-analysis was not attempted because data sources, outcomes, validation designs and metrics differed too much to be pooled. Figures were produced with Python (matplotlib) from the coded extraction table.

RESULTS AND DISCUSSION

Study Selection

The database search identified 570 records and an additional nine acceptable records were identified by preliminary searches (Figure 1). Sixteen of 24 assessed database records were excluded: four were related to gestational diabetes solely, 11 predicted outcomes or complications in patients with pre-existing diabetes (eg frailty, neuropathy, kidney disease or mortality) and one record could not be recovered. The database contained eight records. Included were 17 studies and nine records from other sources including SPRINGER *scientific reports* and others.

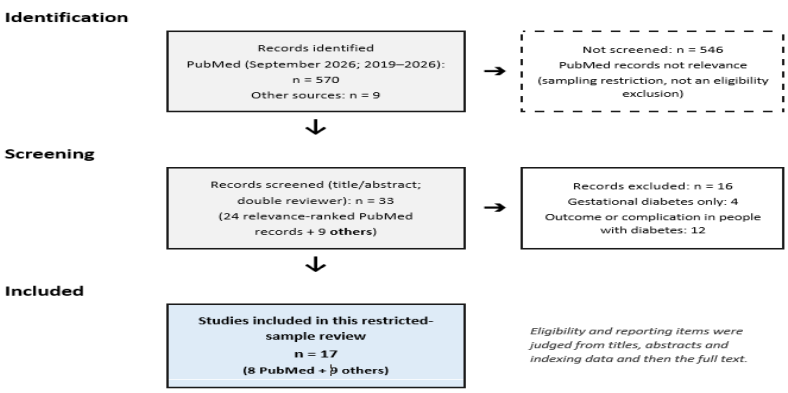

Figure 1: PRISMA 2020-Style Flow Diagram of Study Selection

Characteristics of Included Studies

Characteristics of the 17 included studies are summarised in Table 2. Thirteen were published in 2026, one in 2025, two in 2022 and one in 2020 (Figure 2a). This skew reflects the relevance and the recent increase of the literature and not a thorough representation of prior years. Public benchmark or online datasets were the primary data source for seven research, five studies used hospital or electronic health record data, two studies used survey or cohort data, two studies used big multi-source data, and one study could not specify the source (Figure 2b).

Table 2: Features of The 17 Included Articles

Study	Data source	Algorithms	Validation	Headline result
Bi et al. (2026)	NHANES (n = 15,528) and Chinese health-examination database (n = 4,952)	LR, SVM, RF, KNN, CART, GBDT, LightGBM, XGBoost	Two independent cohorts	XGBoost AUC 0.814 (0.803 without liver enzymes)
Dadashpour et al. (2026)	Kaggle dataset (n = 99,982)	Bayesian ridge, LR, XGBoost, ANN, RF	10-fold stratified cross-validation	XGBoost accuracy 96%, AUC 0.98
Kamkar et al. (2026)	MASHAD cohort, Iran (n = 6,093)	LR, RF, XGBoost, LightGBM	Grid search with CV; independent test set	RF AUC 0.73; sensitivity 0.618, specificity 0.707
Yata et al. (2026)	Pima database (n = 768)	Adversarial neural network vs LR	Stratified split; five random seeds	ROC-AUC 0.785 to 0.790 with debiasing
Salas-Ramos et al. (2026)	Primary care clinic, Mexico (n = 1,903)	SVM, LR, RF, MLP, KNN, ensembles, XGBoost; SMOTE	Not stated	XGBoost with SMOTE: sensitivity 79%, specificity 51.9%
Majyambere et al. (2026)	Public dataset (n = 520); Rwanda (n = 270,943)	MLP, SVM, KNN, XGBoost, NB	K-3, K-fold cross-validation	Accuracy 95.19% (MLP); 97.14% (XGBoost)
Lim et al. (2026)	Home-health nursing records, Korea (7,896 records)	Five ML models incl. RF; NLP	10-fold CV	RF AUC 0.985
Aich et al. (2026)	CDC survey; Pima database	Copula-based feature selection; four classifiers	DeLong tests; permutation importance	No significant difference from strong baselines on Pima
Nasrin et al. (2026)	Kaggle dataset (about 100,000)	LR, RF, MPA-optimised XGBoost	Fold-wise CV; Wilcoxon, McNemar, DeLong	Accuracy 96.72%, AUC 0.9956, Brier 0.0262
Nanda Kishore & Navaneethan (2026)	Multi-source (353,680 records)	Stacking of RF, XGBoost, LightGBM	Leakage-free; held-out test (n = 70,736); CV	AUC 0.8481; 0.9781 for HbA1c/glucose sub-model
Johnson et al. (2025)	Sylhet, Obesity, NIDDK-DF; Pima, LMCH	ML, ensemble and DL with generative synthetic data	Not stated	Accuracy 95.19% (Sylhet); 98.37% (Pima)
Xiao et al. (2026)	12 hospitals, China (training n = 11,382; test n = 1,026)	XGBoost	5-fold CV; held-out test	AUC 0.772 (95% CI 0.703–0.841)
Khan et al. (2026)	Pima database (n = 768)	LR, NB, AdaBoost, XGBoost; SMOTE	Stratified 10-fold CV; held-out test	Accuracy about 94%
Mohsen et al. (2022)	Practice Fusion EHR dataset	RF, SVC, AdaBoost, ensemble, AutoML	5-fold CV	AutoML best across measures
Shin et al. (2022)	Tertiary hospital, Korea (62 and 27 variables)	Gradient boosting, RF	Stratified 10-fold CV	AUC 0.928 (62 variables); 0.842–0.880 (27)
Kopitar et al. (2020)	6 different Sample datasets from 6 categories (Total n = 363)	Glmnet, RF, XGBoost, LightGBM vs regression	100 bootstrap iterations	No clinically relevant gain over simple regression
Shobha et al. (2026)	Public datasets (EHRs data size of 10000)	KNN, LR, NB, RF, XGBoost; multiclass subtypes	hold-out validation strategy	XGBoost accuracy 0.97

AUC, area under the receiver operating characteristic curve; CV, cross-validation; DL, deep learning; LR, logistic regression; MLP, multilayer perceptron; NB, naive Bayes; RF, random forest; SMOTE, synthetic minority oversampling technique; SVM, support vector machine. Results are as reported in the study.

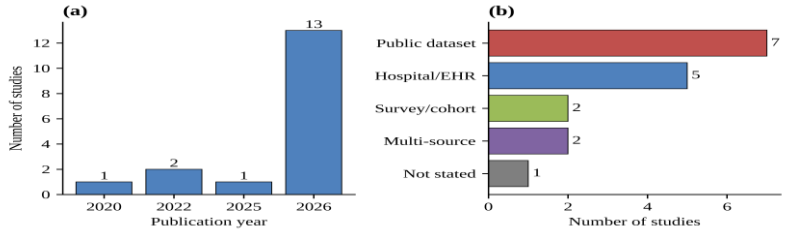

Figure 2: Included Studies by (a) Publication Year and (b) Main Data Source Category

Data Sources and Sample Sizes

The Pima Indians Diabetes Database was used in four of 17 research (24%) either alone or in combination with other sources (Aich et al., 2026; Johnson et al., 2025; Khan et al., 2026; Yata et al., 2026). The Smith et al. (1988) data set includes only 768 women from a single population. Estimates of performance from this data set are therefore uncertain and not generalisable. Two studies analysed public Kaggle data of around 100,000 records (Dadashpour et al., 2026; Nasrin et al., 2026), while two studies employed very big multi-source or program data of 270,943 and 353,680 records (Majyambere et al., 2026; Nanda Kishore & Navaneethan, 2026). Of the 11 studies that did report sample numbers, they ranged from 768 to 353,680 (Figure 3). Six research did not provide sample sizes. Large size is no guarantee of representativeness: a large public online dataset may tell us little about the population in which a model will be deployed. Sample size calculation for prediction models is seldom stated, however guidance is available (Riley et al., 2020).

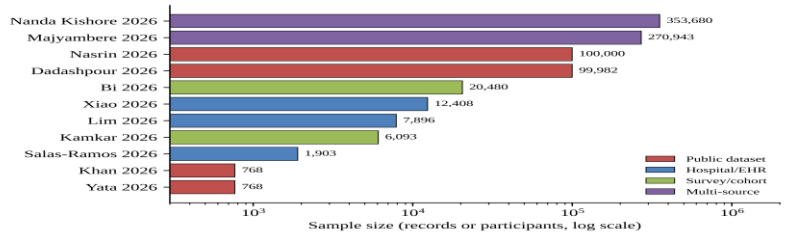

Figure 3: Reported Sample Size of Included Studies (Log Scale). Bars Are Coloured by Data Source Category

Algorithms

In 12 research gradient boosting approaches, and in particular XGBoost, were reported, in 11 random forests and in nine regression models (Figure 4). The neural networks were mentioned five times, support vector machines and k-nearest neighbours four times each, and stacking or ensemble techniques four times. The algorithms are named after these personalities. Gradient boosting was the best classifier in five investigations (Bi et al., 2026; Dadashpour et al., 2026; Majyambere et al., 2026; Nasrin et al., 2026; Shobha et al., 2026) and random forest in two studies (Kamkar et al., 2026; Lim et al., 2026). This is in line with the prevailing belief that tree ensembles work well on tabular clinical data (Breiman, 2001; Chen & Guestrin, 2016). Yet the evidence for a systematic benefit over simpler models is not strong. In agreement with the review of Christodoulou et al. (2019), Kopitar et al. (2020) reported that the replacement of a simple regression model by more advanced models did not result in any clinically relevant improvement in the detection of undiagnosed type 2 diabetes. Similarly, Nanda Kishore and Navaneethan (2026) found that a basic average baseline performed somewhat better than a stacking ensemble on discrimination. Such results advocate for an always included regularised logistic regression baseline.

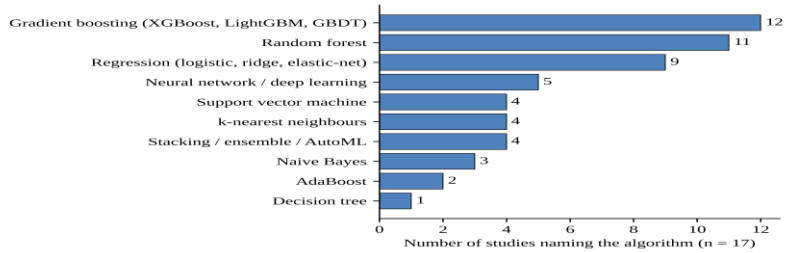

Figure 4: Number of Included Studies Naming Each Algorithm

Reported Performance and What Drives It

An AUC for the headline model was reported in nine studies. Values varied from 0.73 to 0.996, median 0.85 (Figure 5). The greatest values were obtained from big public datasets (0.98 in Dadashpour et al., 2026 and 0.9956 in Nasrin et al., 2026) and from retrospective hospital data (0.985 in Lim et al., 2026 and 0.928 in Shin et al., 2022). The lowest was from a prospective cohort predicting incident diabetes (Kamkar et al., 2026, 0.73) and multi-centre health-checkup data to detect undiagnosed diabetes (Xiao et al., 2026, 0.772).

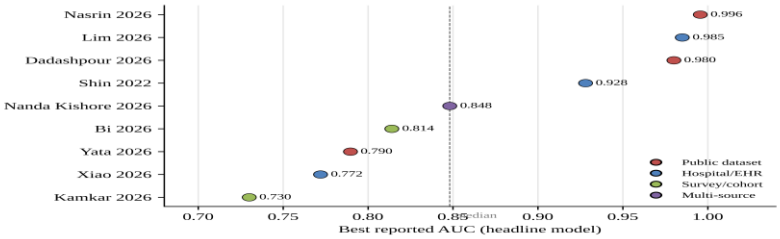

Figure 5: Best Reported AUC of The Headline Model in The Nine Studies Reporting It. The Dashed Line Marks the Median (0.848)

Several studies point to glycaemic predictors as a main driver of this spread. In the two Kaggle-based studies, SHAP analysis ranked HbA1c and blood glucose among the most influential predictors, together with age (Dadashpour et al., 2026; Nasrin et al., 2026). Nanda Kishore and Navaneethan (2026) quantified the effect directly: a sub-model trained on complete HbA1c and glucose data reached an AUC of 0.9781, whereas the fused multi-source ensemble reached 0.8481, a gap of 0.130. In the multi-centre Chinese study, fasting blood glucose alone accounted for 50.6% of the model's feature importance (Xiao et al., 2026), and Shin et al. (2022) found that accuracy rose when fasting glucose was included. Because HbA1c and glucose are themselves diagnostic criteria, models that rely on them are closer to confirming diabetes than to predicting it. Their high AUC values are real but may be of little use for early risk identification, and they should not be compared directly with models that predict incident disease from non-glycaemic data. We therefore recommend that authors state clearly whether their aim is detection of undiagnosed diabetes or prediction of future diabetes and justify including glycaemic variables accordingly.

Two other features of the evidence deserve comment. First, performance was often reported on the same dataset used for tuning or on a single split, which favours optimistic estimates. Second, class imbalance strongly affects the usefulness of results: in a Mexican primary care sample, base models detected people at risk with sensitivities of only 25% to 40%, and oversampling with XGBoost raised sensitivity to 79% at the cost of specificity of 51.9% (Salas-Ramos et al., 2026). Accuracy alone would have hidden this trade-off.

Validation and Reporting Quality

Reporting of methodological safeguards was uneven. Resampling-based validation such as cross-validation or bootstrapping was reported in nine of 17 records and explainability in ten. In contrast, explicit control of data leakage was stated in three studies (Khan et al., 2026; Nanda Kishore & Navaneethan, 2026; Yata et al., 2026), handling of class imbalance in five, and uncertainty or statistical testing in six. Only one study evaluated models in two independent cohorts from different countries (Bi et al., 2026), only one reported calibration through the Brier score (Nasrin et al., 2026), and only one assessed fairness (Yata et al., 2026). Decision-curve analysis, which estimates clinical net benefit (Vickers & Elkin, 2006), was mentioned by Nanda Kishore and Navaneethan (2026) but was absent from most others.

These results mirror the gaps identified in other reviews. External validation was lacking in models of progression from prediabetes to diabetes (Wang et al., 2026), in primary care risk models (Ravi & Olickal, 2026) and in retinopathy risk models (Teruna et al., 2026) (Table 3). Weak reporting of leakage control is especially worrying because common preprocessing steps, including imputation, scaling, feature selection and oversampling with SMOTE (Chawla et al., 2002), can leak information if applied before data are split (Kaufman et al., 2012). The few studies that did report such control show that it is feasible: Khan et al. (2026) applied SMOTE strictly within cross-validation folds, Yata et al. (2026) fitted the scaler on training data only, and Nanda Kishore and Navaneethan (2026) fitted normalisation, imputation and SMOTE exclusively on training data. Without these safeguards, the very high accuracy figures of 95% to 99% seen in several studies cannot be accepted at face value.

Table 3: Recent Secondary Evidence On ML-Based Diabetes Prediction Used for Context

Review	Focus	Evidence base	Key message
Khokhar et al. (2024)	Diabetes prediction	Systematic literature review of datasets, algorithms, training and metrics	Datasets include NHANES and Pima; CNN, SVM, LR, XGBoost compared; ethics emphasised
Kiran et al. (2025)	Type 2 diabetes prediction	Bibliometric and literature analysis, 1991–2024	Documents the growth of ML and AI research in this field
Teruna et al. (2026)	Retinopathy risk in type 2 diabetes	1,040 records; eight studies	AUC 0.68–0.97; external validation largely absent
Wang et al. (2026)	Prediabetes to diabetes	16 studies; 1,368,130 people	Pooled C-index 0.76 (training), 0.84 (validation); external validation lacking
Ravi & Olickal (2026)	Primary care risk models	186 records; four studies	Moderate-to-good performance; limited external validation, calibration, implementation

Interpretability, Fairness and Class Balance

Explainability was the best-reported item, mostly through SHAP (Lundberg & Lee, 2017) and feature importance. This is encouraging, but explanations describe model behaviour and do not establish causation, and some authors argue that inherently interpretable models should be preferred for high-stakes decisions (Rudin, 2019). Nanda Kishore and Navaneethan (2026) found weak and non-significant

agreement between SHAP, permutation importance and LIME rankings, which cautions against treating any single explanation method as definitive. Fairness was addressed in one study. Yata et al. (2026) used adversarial debiasing on the Pima data and found that recall in the over-50 age group rose from 0.556 to 0.778, but the recall parity gap on the main test partition widened from 0.0996 to 0.2153, and across five random seeds the mean gap fell only slightly from 0.328 to 0.303 with high variability. Their conclusion that single train-test splits are insufficient for fairness assessment applies more widely. The importance of such checks was shown by Obermeyer et al. (2019), who found that a widely used health-care algorithm under-served Black patients because of a biased proxy outcome.

Implications for Nigeria and Other Low-Resource Settings

Only one of the studies in the review used data from sub-Saharan Africa. Majyambere et al. (2026) proposed a two-stage explainable model in Rwanda where the first stage used symptoms using a public dataset of 520 records, while the second stage included demographic, anthropometric and vital-sign data from 270,943 Rwandan records, with XGBoost achieving 97.14% accuracy. The authors hope to integrate the models into the mUzima mobile application so that community health professionals may detect and refer high-risk individuals, which is the kind of deployment pathway needed in Nigeria. The records indicate no external validation and so great accuracy on a single program dataset needs to be verified in independent data before use. None of the 17-research used Nigerian data.

This is the Nigerian epidemiological evidence. This is why local data matters. Pooled prevalence ranges from 2.03% in the north-central zone to 11.35% in the south-south zone (Olamoyegun et al., 2024), so the predictive value of any model will vary by location, and a model trained in one zone or country may be miscalibrated in another. At the primary care level, readily available documented risk factors such as family history, urban living, physical inactivity, obesity and unhealthy diet (Olamoyegun et al., 2024; Uloko et al., 2018) are available without laboratory tests, which supports building parsimonious models from age, family history, body mass index, waist circumference and blood pressure. In cases where a point-of-care glucose reading is utilised, its role should be disclosed transparently as described in Section 3.5.

Recommended Workflow, Strengths and Limitations

Drawing together the evidence, Figure 7 outlines a development and evaluation workflow suited to Nigerian settings. Its central points are to define the clinical question before choosing predictors, to build representative multi-centre data, to prevent leakage by fitting all preprocessing within training folds, to compare complex models with a regularised logistic regression baseline, to report calibration and uncertainty and validate externally and temporally, to assess fairness across sex, age and region, and to pilot prospectively in primary care.

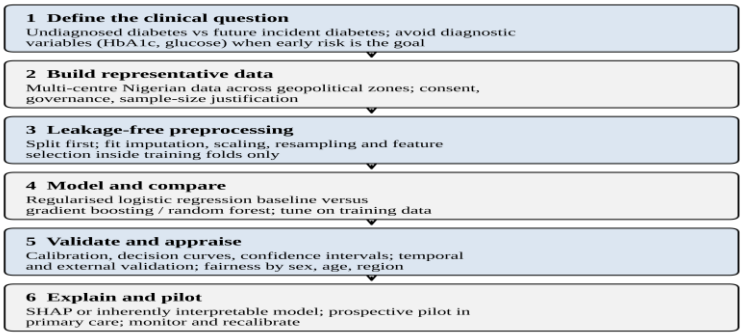

Figure 7: Recommended Workflow for Developing and Evaluating Diabetes Prediction Models for Nigerian Primary Care

Strengths of this evaluation are a visible and reproducible search string, unambiguous eligibility criteria, systematic extraction and an emphasis on practical reporting elements. Its restrictions are very great and must be borne in mind. One database was searched. We only reviewed 24 of the 570 database entries by relevance, therefore the sample is not exhaustive and favours recent publications, as 13 of the 17 included research were published in 2026. Screening and Extraction Some items may be labelled as not reported in the full texts. Heterogeneity hindered meta-analysis and presented measures are not directly comparable across trials. The results should thus be considered as a structured description of a recent sampling of the literature rather than a complete census. Strengthening of the review would involve searching in new databases (Scopus, Web of Science, IEEE Xplore, African Journals Online), double screening all records and grading full texts.

CONCLUSION

Recent studies report excellent ML performance for diabetes prediction, but the evidence base is weaker than the headline figures suggest. In this sample of 17 studies, the highest accuracies came from large public datasets and from models dominated by glycaemic variables that are themselves diagnostic, while explicit leakage control, calibration, independent validation and fairness assessment were rarely reported. Only one study used African data and none used Nigerian data. For Nigeria, progress is more likely to come from representative multi-regional datasets, leakage-free and transparent methods, models built on low-cost predictors, and prospective primary care pilots than from further comparisons of classifiers on benchmark data. Researchers and journals can help by requiring Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) style reporting, external validation and a clear statement of whether a model detects existing diabetes or predicts it.

ACKNOWLEDGEMENT

The authors are appreciative to the Tertiary Education Trust Fund (TETFund) for its generous financial support toward the effective completion of this research under the Institutional

Based Research (IBR) Grant (Reference No.:TETF/DR&D/CE/UNI/GADAU/IBR/2015/VOL.I). We also want to acknowledge their profound gratitude to the Center for Eccelence Research and Innovation (CERI) and the management of Sa'adu Zungur University (SAZU) for their immense administrative support, institutional backing and conducive research atmosphere granted during the tenure of this project.

REFERENCES

Aich, A., Murshed, M. M., Hewage, S., & Mayeaux, A. (2026). A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-41874-9

Bi, Y., Yuan, X., Wang, Y., Sun, C., & Tian, S. (2026). Machine-learning-based prediction model of type 2 diabetes using liver enzymes: A cross-sectional study. *Frontiers in Endocrinology*, *17*, Article 1905144. https://doi.org/10.3389/fendo.2026.1905144

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, *350*, Article g7594. https://doi.org/10.1136/bmj.g7594

Dadashpour, M., Yousefi, M., Effati, S., & Hafezi, S. G. (2026). Evaluation and comparison of machine learning methods for type 2 diabetes classification and associated factors. *Healthcare Informatics Research*, *32*(3), 214–223. https://doi.org/10.4258/hir.2026.32.3.214

Johnson, E. S., Gandhi, U., Reddy, U. S., & Mangalanathan, U. (2025). Generative AI models for type 2 diabetes mellitus risk prediction. *Health Systems*, *15*(1), 66–89. https://doi.org/10.1080/20476965.2025.2546904

Kamkar, N., Behzadi, S., Mahdavizadeh, V., Maali Tafti, M., Mansoori, A., Poudineh, M., Esmaily, H., & Ghayour-Mobarhan, M. (2026). A machine learning-based framework for predicting type 2 diabetes mellitus using hematological indices. *American Heart Journal Plus: Cardiology Research and Practice*, *69*, Article 100854. https://doi.org/10.1016/j.ahjo.2026.100854

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), Article 15. https://doi.org/10.1145/2382577.2382579

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Khan, T. F., Saeed, M., Hussain, M., Khan, L., Khan, M. Z., Nadeem, A., & Alghamdi, T. (2026). A comparative and interpretable machine learning framework for reliable diabetes risk prediction. *Scientific Reports*, *16*. https://doi.org/10.1038/s41598-026-61038-z

Khokhar, P. B., Gravino, C., & Palomba, F. (2024). *Advances in artificial intelligence for diabetes prediction: Insights from a systematic literature review* [Preprint]. https://doi.org/arXiv

Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscionek, B., & Russell, D. (2025). Machine learning and artificial intelligence in type 2 diabetes prediction: A comprehensive 33-year bibliometric and literature analysis. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1557467

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1), Article 11981. https://doi.org/10.1038/s41598-020-68771-z

Lim, D., Kim, A., Lee, H., & Woo, K. (2026). Enhancing community-based nursing decision support: Machine learning models for diabetes risk prediction using home health nursing notes. *Public Health Nursing*, *43*(4), 1043–1055. https://doi.org/10.1111/phn.70122

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *https://doi.org/30*,4765–4774

Majyambere, S., Lindgren, T., Twizere, C., & Ntakirutimana, I. (2026). Early type 2 diabetes risk prediction using explainable machine learning in a two-stage approach. *Frontiers in Digital Health*, *8*, Article 1743619. https://doi.org/10.3389/fdgth.2026.1743619

Mohsen, F., Biswas, M. R., Ali, H., Alam, T., Househ, M., & Shah, Z. (2022). Customized and automated machine learning-based models for diabetes type 2 classification. *Studies in Health Technology and Informatics*, *295*, 517–520. https://doi.org/10.3233/SHTI220779

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, *170*(1), W1–W33. https://doi.org/10.7326/M18-1377

Nanda Kishore, M., & Navaneethan, C. (2026). Quantifying the imputation paradox and XAI inconsistency in multi-

source diabetes prediction: A 353,680-record leakage-free stacking ensemble with dynamic routing architecture. *Frontiers in Artificial Intelligence*, *9*, Article 1898153. https://doi.org/10.3389/frai.2026.1898153

Nasrin, A., Asif, M. B., Naidu, R., Asif, A. H., Chohan, M. S., Khan, G. A., Arifuzzaman, M., Azad, A. K. M., & Martuza, M. A. (2026). An optimized and explainable machine learning framework for diabetes prediction using Marine Predators Algorithm and SHAP. *Journal of Personalized Medicine*, *16*(8), Article 423. https://doi.org/10.3390/jpm16080423

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Olamoyegun, M. A., Alare, K., Afolabi, S. A., Aderinto, N., & Adeyemi, T. (2024). A systematic review and meta-analysis of the prevalence and risk factors of type 2 diabetes mellitus in Nigeria. *Clinical Diabetes and Endocrinology*, *10*(1), 1–13. https://doi.org/10.1186/s40842-024-00209-1

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Ravi, R., & Olickal, J. J. (2026). Machine learning-based risk prediction models for type 2 diabetes in primary care: A scoping review. *Frontiers in Public Health*, *14*, Article 1946070. https://doi.org/10.3389/fpubh.2026.1946070

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, Article m441. https://doi.org/10.1136/bmj.m441

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salas-Ramos, B. M., Garcia-Calvillo, I., Navarro-Acosta, J. A., & Mendoza-Frías, G. I. (2026). Predicting individuals at risk of diabetes using machine learning. *Cureus*, *18*(5), Article e109761. https://doi.org/10.7759/cureus.109761

Shin, J., Kim, J., Lee, C., Yoon, J. Y., Kim, S., Song, S., & Kim, H.-S. (2022). Development of various diabetes prediction models using machine learning techniques. *Diabetes & Metabolism Journal*, *46*(4), 650–657. https://doi.org/10.4093/dmj.2021.0115

Shobha, T., Pradeep, S., Patil, S., Subramanya, J., & Mitra, G. B. S. (2026). A data-driven machine learning model for effective diabetes diagnosis. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-63505-z

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* https://doi.org/pp. 261–265

Teruna, M. M. S., Razak, T. R., Yasin, S. M., Ali, A. A. R., Salleh, M. M. M., Yusoff, M. Z. M., Khebir, M. H. A., Rameli, N. A. C., Sallehhudin, M. I. M., Asraff, M. S. M., & Isa, M. R. (2026). Artificial intelligence driven prognostic models for predicting diabetic retinopathy risk in type 2 diabetes mellitus: A systematic review. *The Medical Journal of Malaysia*, *81*(5), https://doi.org/759–768

Uloko, A. E., Musa, B. M., Ramalan, M. A., Gezawa, I. D., Puepet, F. H., Uloko, A. T., Borodo, M. M., & Sada, K. B. (2018). Prevalence and risk factors for diabetes mellitus in Nigeria: A systematic review and meta-analysis. *Diabetes Therapy*, *9*(3), 1307–1316. https://doi.org/10.1007/s13300-018-0441-1

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*, Article 230. https://doi.org/10.1186/s12916-019-1466-7

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

Wang, Y., Wang, C., & Wang, J. (2026). Prediction models for progression from prediabetes to diabetes: A systematic review and meta-analysis. *Frontiers in Endocrinology*, *17*, Article 1888466. https://doi.org/10.3389/fendo.2026.1888466

Xiao, H., Xi, Q., Zeng, P., Hao, J., He, Q., Wang, X., & Zhang, C. (2026). An XGBoost-based model for detecting undiagnosed type 2 diabetes using routine physical and lifestyle data from a multi-center Chinese population. *Frontiers in Medicine*, *13*, Article 1798509. https://doi.org/10.3389/fmed.2026.1798509

Yata, V. K., Jena, S., Indracanti, M., Chitta, S., & Kolliputi, N. (2026). Adversarial debiasing for age-equitable diabetes prediction: Performance-fairness trade-offs and partition dependency in machine learning. *Frontiers in Digital Health*, *8*, Article 1816806. https://doi.org/10.3389/fdgth.2026.1816806

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.