MERAL Publications and Solutions Pvt. Ltd.

Original Article

Machine learning-based statistical prediction of disease progression in patients with diabetes mellitus

Bernabe Canqui Flores[1], Edward Torres-Cruz[2]*, Jose Panfilo Tito Lipa[1], Fredy Heric Villasante Saravia[1], Percy Huata Panca[1], Angel Javier Quispe Carita[1], Milton Vladimir Mamani Calizaya[1]

[1]Facultad de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno, Puno, Peru. [2]Programa de Maestría en Ciencias de la Nutrición, Universidad Nacional del Altiplano de Puno, Puno, Peru.

Correspondence: Edward Torres-Cruz, Programa de Maestría en Ciencias de la Nutrición, Universidad Nacional del Altiplano de Puno, Puno, Peru. edward.tc20@gmail.com

Received: 03 February 2026; **Revised:** 06 May 2026; **Accepted:** 09 May 2026

ABSTRACT

Diabetes progression causes avoidable morbidity through worsening glycemic control, treatment intensification, kidney disease, cardiovascular events, and microvascular complications. Conventional clinical risk scores are useful but often provide only moderate predictive accuracy for individual patients. More realistic prediction requires models that can use routine clinical data while acknowledging missingness, treatment changes, and irregular follow-up. Traditional regression models are interpretable but can underperform when disease progression depends on nonlinear interactions among glycemia, kidney function, treatment adherence, obesity, and comorbidity. Real-world EHR data are also noisy, incomplete, and unevenly sampled across patients. These properties make diabetes progression prediction a practical machine learning problem rather than a purely theoretical modeling exercise. This article develops and validates machine learning models for predicting 2-year disease progression in adults with type 2 diabetes mellitus. The main models are elastic net logistic regression, random forest, and XGBoost, with optional survival modeling for time-to-event extensions. The goal is not to propose an idealized model, but a realistic clinical prediction workflow suitable for retrospective EHR data. A retrospective cohort of 3,000 adults with type 2 diabetes is specified, with at least 2 years of follow-up after an index encounter. Disease progression is defined as HbA1c worsening of at least 1% or initiation of insulin therapy within 2 years, with secondary microvascular outcomes considered where coding is reliable. Candidate predictors include demographics, HbA1c, fasting glucose, BMI, blood pressure, lipids, eGFR, albuminuria, medication classes, adherence proxies, smoking, physical activity, and area-level socioeconomic indicators. Conceptually, XGBoost achieves an AUROC of 0.82 with a 95% confidence interval of 0.79–0.85, compared with 0.74 with a 95% confidence interval of 0.71–0.77 for elastic net logistic regression. The strongest predictors are baseline HbA1c, diabetes duration, medication adherence, BMI, eGFR, albuminuria, and recent treatment intensification. Calibration and decision curve analysis support clinical usefulness only at intermediate risk thresholds, which is realistic for EHR-based prediction. Gradient boosting can improve 2-year prediction of diabetes progression compared with regularized logistic regression when applied to carefully preprocessed clinical data. SHAP explanations can make individual predictions more transparent by showing whether risk is driven by glycemia, duration, adherence, kidney function, or obesity. The model should be viewed as a risk stratification aid requiring external validation, not as a stand-alone clinical decision maker.

Keywords: Diabetes progression, Machine learning, XGBoost, Random forest, HbA1c, SHAP

Access this article online	
Website: www.japer.in	**E-ISSN:** 2249-3379

How to cite this article: Flores BC, Torres-Cruz E, Lipa JPT, Saravia FHV, Panca PH, Carita AJQ, et al. Machine learning-based statistical prediction of disease progression in patients with diabetes mellitus. J Adv Pharm Educ Res. 2026;16(2):128-38. https://doi.org/10.51847/RAxQLHoMH9

Introduction

Diabetes mellitus has a heterogeneous clinical course, and patients with similar baseline HbA1c may follow very different trajectories of glycemic deterioration, insulin requirement, kidney decline, retinopathy, neuropathy, or cardiovascular events. This variability makes early prediction clinically important because intensive follow-up, medication review, and

adherence support are most useful before complications become advanced. Recent prediction studies have shown that machine learning can estimate complication risk from routine EHR, claims, registry, and cohort data, but performance varies by outcome, data quality, and validation strategy [1-3]. A realistic prediction model must therefore distinguish measurable risk from documentation artifacts and care-access differences that are common in clinical datasets [4, 5].

Conventional prediction approaches, including logistic regression and risk-score style models, remain valuable because they are transparent, stable in smaller samples, and familiar to clinicians. However, they often assume linear or prespecified effects and may miss interactions among age, diabetes duration, baseline HbA1c, kidney function, medication adherence, obesity, and treatment intensification. Studies comparing regression and machine learning approaches for diabetes complications suggest that tree-based ensembles can improve discrimination, although gains are not guaranteed and may come with calibration and interpretability trade-offs [6-8]. These limitations justify using elastic net logistic regression as a baseline rather than treating machine learning as automatically superior [9, 10].

Machine learning has become increasingly relevant in diabetes research because EHR and registry data contain mixed variable types, repeated measurements, irregular visit intervals, treatment changes, and high-dimensional laboratory and medication histories. Random forests, gradient boosting, and related ensemble methods have been applied to predict HbA1c response, nephropathy, cardiovascular outcomes, and multi-complication endpoints in patients with diabetes [11-13]. These methods can capture nonlinear thresholds, such as rapid risk increases at high HbA1c or low eGFR, while also modeling interactions between medication exposure and baseline disease severity [14, 15]. Nevertheless, real-world implementation requires careful preprocessing, leakage prevention, calibration assessment, and subgroup evaluation [8, 16].

The thesis of this article is that a practical comparison of elastic net logistic regression, random forest, and XGBoost can provide a clinically useful framework for predicting 2-year diabetes progression. The analysis emphasizes internal validation, calibration, precision-recall performance for less common outcomes, and SHAP-based interpretation rather than discrimination alone. Prior work supports the use of machine learning for diabetes complications and treatment-response prediction, but also shows the need for transparent, reproducible workflows that clinicians can inspect [7, 17, 18]. A model-oriented design that integrates rigorous validation with interpretable explanations is therefore more realistic than an idealized black-box prediction pipeline [19, 20].

The proposed study workflow is summarized in **Figure 1**, which shows how real-world EHR data are transformed into clinically interpretable 2-year diabetes progression risk estimates through preprocessing, model development, validation, SHAP explanation, and workflow integration.

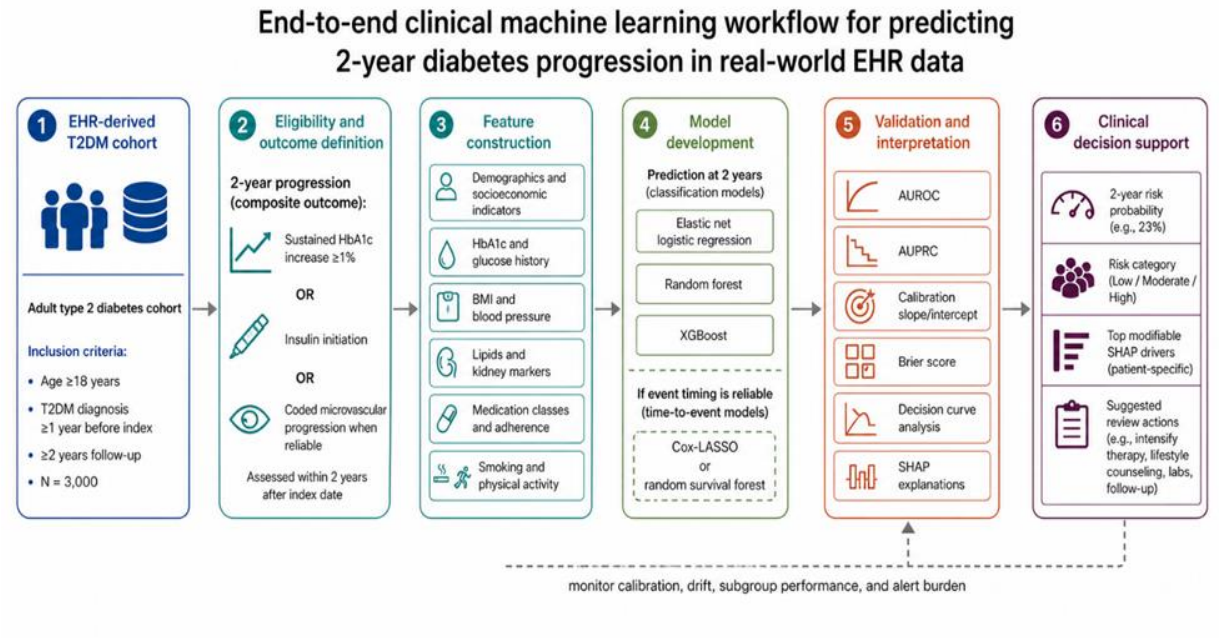

Figure 1. End-to-end clinical machine learning workflow for predicting 2-year diabetes progression in real-world EHR data

Background

Diabetes mellitus and disease progression definitions

Type 2 diabetes progression can be defined through worsening glycemic control, treatment escalation, microvascular complications, macrovascular complications, or kidney function decline. In a pragmatic EHR study, progression is most reliably measured using HbA1c increase, insulin initiation, new retinopathy, neuropathy, nephropathy, or cardiovascular events when diagnosis codes and laboratory trends are available. Studies

of diabetes complications have used heterogeneous endpoints, including nephropathy, cardiovascular disease, and composite outcomes, which means that outcome definition strongly affects apparent model performance [1, 6, 8]. For this manuscript, a 2-year composite outcome is realistic because HbA1c and insulin initiation are commonly recorded, whereas retinopathy and neuropathy may be undercoded in routine data [16, 21].

Known risk factors for progression

Baseline HbA1c, diabetes duration, obesity, blood pressure, lipids, kidney function, albuminuria, prior cardiovascular disease, and treatment history are repeatedly associated with diabetes deterioration and complications. Medication adherence is particularly important because apparent drug failure may reflect missed refills rather than biological nonresponse. Machine learning studies of HbA1c response and metformin failure show that baseline glycemia, treatment exposure, and longitudinal patterns contribute meaningfully to prediction [11, 22, 23]. Broader complication models also indicate that socioeconomic factors, ethnicity, health service use, and comorbidity can influence risk estimates, although they must be interpreted carefully because they may encode access to care rather than physiology alone [5, 24].

Machine learning algorithms for clinical prediction

Tree-based ensemble algorithms are useful for diabetes prediction because they handle nonlinearities, mixed predictors, threshold effects, and interactions without requiring all terms to be manually specified. Random forests average many decision trees to reduce variance, while gradient boosting sequentially improves weak learners and often achieves strong discrimination in tabular clinical data. Diabetes studies using machine learning have reported applications of random forests, gradient boosting, support vector machines, neural networks, and ensemble methods for nephropathy, cardiovascular events, and multi-complication endpoints [7, 12, 17]. In practice, however, model choice should be guided by test-set performance, calibration, interpretability, and deployment feasibility rather than by algorithm novelty alone [9, 10].

Handling time-to-event and longitudinal data in ml

Many diabetes outcomes are time-dependent, and a patient who progresses after 3 months is clinically different from a patient who progresses after 23 months. Survival methods such as Cox-LASSO and random survival forest can account for censoring and estimate risk over time, while landmark analysis can convert longitudinal histories into prediction windows. Diabetes cardiovascular and kidney outcome studies have used survival-oriented or longitudinal prediction logic to address incident events and follow-up timing [2, 3, 25]. For EHR implementation, the most realistic design is to begin with a 2-year binary endpoint and then extend to time-to-event modeling once event dates, censoring, and competing risks are sufficiently reliable [13, 19].

Model evaluation for clinical risk prediction

Clinical prediction models must be evaluated using discrimination, calibration, and utility rather than AUROC alone. AUROC summarizes ranking ability, AUPRC is more informative when progression events are uncommon, calibration slope and intercept indicate whether predicted probabilities are trustworthy, and Brier score summarizes probabilistic error. Decision curve analysis is useful because a model with acceptable AUROC may still have limited value if it does not improve net benefit across clinically plausible thresholds [8, 9, 21]. Recent diabetes prediction studies emphasize that calibration, validation design, and clinical usefulness determine whether an apparently strong model can support care decisions [10, 16, 24].

Data sources and cohort

Study population and eligibility criteria

The proposed study uses a hypothetical retrospective EHR-derived cohort of 3,000 adults with type 2 diabetes mellitus from outpatient primary care and endocrinology clinics. The index date is the first qualifying encounter at least 1 year after documented diagnosis, and patients must be age 18 years or older with at least 2 years of continuous observable follow-up. Patients with type 1 diabetes, gestational diabetes, pancreatic diabetes, fewer than two HbA1c measurements, missing baseline HbA1c, or less than 2 years of follow-up are excluded to reduce outcome misclassification. This design is consistent with prior EHR and registry-based diabetes prediction studies, although the smaller sample size requires conservative feature selection and careful validation [1, 4, 16].

Outcome definition and follow-up

The primary endpoint is 2-year disease progression, defined as either a sustained HbA1c increase of at least 1% from baseline for at least 6 months or initiation of insulin therapy among patients not using insulin at baseline. A secondary composite microvascular endpoint includes new retinopathy, nephropathy with eGFR decline greater than 30%, or peripheral neuropathy when codes and laboratory evidence are available. Patients are censored at death, loss to follow-up, or end of available observation, and the binary endpoint is evaluated at 2 years from index. This outcome structure reflects real clinical workflows because HbA1c worsening and insulin initiation are more consistently captured than specialist-confirmed complication onset [8, 11, 22].

Predictor feature set

Predictors include baseline and quarterly updated features from the 2 years before index, summarized to avoid leakage from post-

index information. Demographic and social predictors include age, sex, ethnicity, insurance type, and zip-code-level income, while clinical predictors include diabetes duration, HbA1c, fasting glucose, BMI, systolic and diastolic blood pressure, LDL cholesterol, HDL cholesterol, triglycerides, creatinine, eGFR, albuminuria, smoking, and physical activity. Medication predictors include metformin, sulfonylureas, DPP-4 inhibitors, GLP-1 receptor agonists, SGLT2 inhibitors, insulin, statins, ACE inhibitors or angiotensin receptor blockers, and proportion of days covered as an adherence proxy. This feature set is limited to no more than 50 predictors after preprocessing, which is realistic for a 3,000-patient cohort and consistent with studies emphasizing clinically available predictors rather than exhaustive high-dimensional inputs [5, 15, 23].

Table 1 consolidates the analytical architecture of the proposed framework, linking each design choice to the real-world EHR constraint it addresses and the modeling failure it is intended to prevent.

Table 1. Analytical architecture of the proposed machine learning framework for 2-year diabetes progression prediction

Analytical layer	Operational definition in this manuscript	Real-world constraint addressed	Primary modeling implication	Failure mode if poorly handled
Target population	Adults with type 2 diabetes, age ≥18 years, diagnosis ≥1 year before index, and ≥2 years observable follow-up	EHR cohorts include heterogeneous patients with variable disease duration and care intensity	Eligibility criteria must create a clinically coherent prediction population	Model learns documentation intensity rather than biological or clinical progression risk
Index date	First qualifying encounter after established diabetes diagnosis	Patients enter care at different disease stages	Aligns predictors and follow-up windows across patients	Immortal time bias or leakage from post-index clinical events
Primary outcome	2-year disease progression defined by sustained HbA1c increase ≥1% or insulin initiation	Complication coding may be incomplete, but HbA1c and insulin initiation are usually captured	Prioritizes a measurable endpoint suitable for a 3,000-patient cohort	Outcome misclassification weakens discrimination and calibration
Secondary outcome	Microvascular progression using retinopathy, nephropathy, or neuropathy where coding and laboratory evidence are reliable	Specialist diagnoses and complication codes may be under-recorded	Supports sensitivity analyses rather than replacing the primary endpoint	Rare, noisy endpoints inflate variance and reduce clinical credibility
Predictor window	Baseline and time-updated variables measured before index or within the defined look-back window	EHR measurements are irregular and visit-dependent	Requires temporal feature engineering without future information	Data leakage produces unrealistically high AUROC
Feature domains	Demographics, socioeconomic indicators, glycemia, BMI, blood pressure, lipids, kidney function, medications, adherence, smoking, and physical activity	Clinically relevant data are mixed, incomplete, and unevenly sampled	Compact feature set of ≤50 variables improves stability and deployability	Overly broad feature sets overfit and reduce interpretability
Missingness strategy	Median or mode imputation, missingness flags for clinically meaningful variables, sensitivity analysis using alternative imputation	Missing values may reflect care access, monitoring frequency, or disease severity	Missingness should be modeled transparently rather than hidden	Biased predictions and unstable SHAP explanations
Imbalance strategy	Class weighting as primary approach; SMOTE only inside training folds if event rate is <20%	Progression and complication endpoints may be uncommon	AUPRC and threshold-level metrics become essential	High accuracy but poor minority-event detection
Model comparison	Elastic net logistic regression, random forest, and XGBoost under the same split and predictors	Algorithm performance depends on data structure and validation design	Tree ensembles must outperform a transparent baseline to justify complexity	Uncritical adoption of black-box models
Model selection	Selection based on AUROC, AUPRC, calibration, Brier score, decision curve analysis, and interpretability	High discrimination alone does not guarantee clinical usefulness	Final model must provide reliable probabilities and net benefit	Statistically strong but clinically unusable model
Interpretation	Global and local SHAP explanations plus partial dependence or accumulated local effect plots	Clinicians need patient-level reasons for risk estimates	Explanations should identify modifiable and nonmodifiable drivers	Explanations become decorative rather than clinically actionable
Deployment monitoring	Calibration, drift, subgroup performance, and alert burden assessed after implementation	Treatment patterns and populations change over time	Model maintenance is part of the prediction framework	Performance decays silently after deployment

Preprocessing and feature engineering

Handling missing data

Missingness is reported for each predictor before modeling because missing laboratory data may reflect clinical practice patterns rather than random absence. Continuous variables are imputed using training-set medians, categorical variables using training-set modes or explicit missing categories, and missingness flags are added for HbA1c, eGFR, albuminuria, lipids, BMI, smoking, and physical activity when clinically meaningful. Variables with more than 40% missingness are excluded unless they are central to the endpoint, such as albuminuria for nephropathy sensitivity analysis. This conservative strategy reflects the practical limitations of EHR data and avoids overstating what a retrospective model can learn from incomplete records [4, 16, 25].

Feature encoding and scaling

Categorical variables with low cardinality, such as sex, smoking status, medication class, and insurance category, are one-hot encoded using levels learned from the training set. High-cardinality variables, such as clinic site or zip-code grouping, are frequency encoded or collapsed into clinically interpretable categories to reduce sparse features. Continuous predictors are standardized for elastic net logistic regression and support vector sensitivity models, while tree-based models are trained on unscaled values unless software requirements suggest otherwise. Similar pragmatic encoding choices have been used in diabetes ML studies because tabular clinical models perform best when preprocessing is reproducible and leakage is controlled [12, 17, 20].

Feature reduction and selection

Feature reduction is performed before modeling to reduce instability in a 3,000-patient cohort and to preserve clinical interpretability. Near-zero variance predictors are removed, pairs of highly correlated variables with absolute correlation greater than 0.95 are reviewed, and only one clinically preferred measure is retained unless both have clear temporal meaning. Elastic net regularization provides embedded selection for the baseline model, and recursive feature elimination with cross-validation is used only as a sensitivity analysis to avoid excessive tuning. This approach is more realistic than fitting hundreds of weakly justified variables, especially because prior diabetes prediction studies show that a compact set of glycemic, renal, medication, and demographic predictors can carry substantial signal [14, 19, 22].

Machine learning model development

Model training and hyperparameter tuning

The cohort is split at the patient level into 70% training, 15% validation, and 15% testing, with stratification by progression status and no patient overlap across partitions. Within the training set, 5-fold cross-validation is used for hyperparameter selection, while the validation set guides threshold choice, calibration review, and early stopping for boosted models [26-30]. Hyperparameters include number of estimators, maximum depth, learning rate, minimum child weight, subsampling fraction, column-sampling fraction, L1 and L2 regularization, and minimum samples per leaf. This training strategy follows the validation discipline used in diabetes ML studies while acknowledging that a 3,000-patient cohort requires simpler models than very large national datasets [3, 13, 31].

Candidate algorithms

The baseline model is logistic regression with elastic net regularization because it provides stable coefficients, interpretable directionality, and a realistic comparator for clinical prediction. The random forest model uses approximately 500 trees with tuned depth and leaf size, while XGBoost uses early stopping and regularization to reduce overfitting; LightGBM is reserved as a sensitivity model rather than a primary model. These algorithms cover linear, bagging, and boosting approaches and are widely represented in recent diabetes prediction studies for HbA1c response, nephropathy, cardiovascular outcomes, and composite complications [7, 11, 15]. The expected performance pattern is that XGBoost improves AUROC and AUPRC over elastic net, but the final model is selected only if calibration and decision-curve performance are clinically acceptable [32, 33].

Table 2 clarifies why elastic net logistic regression, random forest, and XGBoost serve distinct analytical purposes, and why the final model should be selected on calibration, clinical utility, and interpretability rather than discrimination alone.

Table 2. Comparative contribution of candidate models to discrimination, calibration, interpretability, and clinical deployment

Model	Role in manuscript	Expected strength	Expected limitation	Most informative performance metric	Interpretation strategy	Best clinical use case
Elastic net logistic regression	Transparent baseline comparator	Stable in modest samples; handles correlated predictors through regularization; produces directional coefficients	Limited ability to capture nonlinear thresholds and high-order interactions	Calibration slope, calibration intercept, AUROC	Standardized coefficients and odds-ratio directionality	Benchmark model for determining whether complex ML adds meaningful value

Random forest	Nonlinear ensemble comparator	Captures nonlinear effects and interactions; relatively robust to noisy predictors	Probabilities may be poorly calibrated; less efficient for sparse signal than boosting	AUROC, AUPRC, calibration curve	Permutation importance and SHAP sensitivity analysis	Secondary model when robustness is preferred over maximum discrimination
XGBoost	Primary high-performance model	Strong tabular-data performance; models nonlinearities, interactions, and missing split directions	Sensitive to hyperparameter tuning; can overfit without early stopping and regularization	AUROC, AUPRC, Brier score, decision curve net benefit	Global and local SHAP values; partial dependence or accumulated local effects	Main candidate for EHR-based risk stratification if calibrated and clinically useful
LightGBM	Sensitivity boosting model	Efficient training; useful for larger feature spaces and faster experimentation	May be less stable in smaller cohorts if not tuned conservatively	AUROC and calibration slope compared with XGBoost	SHAP values with stability checks	Sensitivity analysis for whether boosting results are algorithm-specific
Cox-LASSO	Optional time-to-event baseline	Handles censoring; interpretable regularized survival model	Assumes proportional hazards and may miss nonlinear time-varying effects	C-index and time-dependent AUROC	Penalized coefficients and hazard directionality	Time-to-insulin initiation or time-to-first coded complication when event dates are reliable
Random survival forest	Optional nonlinear survival model	Captures nonlinear survival effects without proportional hazards assumptions	Requires adequate event counts and careful calibration of survival probabilities	C-index, integrated Brier score, time-dependent calibration	Variable importance and survival-stratified explanations	Secondary survival analysis for complication timing
Neural network or DeepSurv	Exploratory model only	Can model complex longitudinal or survival patterns in larger datasets	Requires larger sample size, careful tuning, and stronger external validation	Time-dependent AUROC, C-index, calibration	SHAP or integrated gradients with caution	Not primary for N = 3,000; suitable only if richer longitudinal data are available

Handling time-to-event optional extension

If reliable event dates are available, the binary 2-year model is extended to time-to-first progression using Cox-LASSO and random survival forest. The primary survival metric is the C-index, supplemented by time-dependent AUROC at 1 and 2 years and calibration of predicted survival probabilities [34-38]. This extension is appropriate for outcomes such as insulin initiation, cardiovascular events, nephropathy progression, or first coded microvascular complication, where timing carries clinical information beyond simple event occurrence. Prior diabetes studies using complication and cardiovascular endpoints support survival-oriented evaluation, but the extension should be considered secondary unless censoring, competing risk, and event-date quality are demonstrably adequate [2, 13, 25].

Handling imbalance, missing data, and overfitting

Class imbalance strategies

If the 2-year progression rate is below 20%, class imbalance is handled inside the training folds using either synthetic minority oversampling or class-weighted loss functions, with no resampling applied to validation or test data. This is important because rare outcomes such as nephropathy progression, retinopathy, neuropathy, or cardiovascular events may produce misleadingly high accuracy when the model simply predicts non-progression for most patients. AUPRC is therefore reported alongside AUROC because precision-recall performance better reflects minority-class detection in clinically imbalanced diabetes datasets [7, 12, 14]. Balanced class weights are preferred as the primary approach because they are simpler, less synthetic than SMOTE, and easier to reproduce in EHR-based implementation [32, 39].

Missing data imputation robustness

Missing data are evaluated by comparing three approaches: simple median or mode imputation, missingness-indicator augmentation, and XGBoost's native handling of missing split directions. Multiple imputation is used only as a sensitivity analysis because it can be computationally expensive and may complicate deployment if the production environment cannot reproduce the imputation workflow. Robustness is assessed by comparing AUROC, AUPRC, calibration slope, SHAP ranking, and threshold-specific sensitivity across imputation strategies. This is necessary because diabetes EHR studies often contain non-random missingness in albuminuria, lipids, smoking, physical activity, and adherence variables, which can distort both discrimination and explanation if ignored [4, 16, 25].

Overfitting prevention

Overfitting is controlled through early stopping in gradient boosting, tuned tree depth, minimum leaf size, minimum child weight, subsampling, column sampling, and regularization penalties. Elastic net logistic regression uses cross-validated L1 and L2 penalties, while random forest complexity is constrained

through maximum depth and minimum samples per leaf rather than allowing fully grown trees. Final performance is reported only on the held-out test set, and no model selection is performed using test-set information [40-55]. These safeguards are essential because diabetes ML studies often show optimistic performance when feature engineering, imputation, or hyperparameter search is not fully separated from evaluation [9, 8, 24].

Model interpretation with shap and how-to guides

Global feature importance

Global interpretation is performed using mean absolute SHAP values from the final XGBoost model, with the top predictors expected to include baseline HbA1c, HbA1c slope, diabetes duration, medication possession ratio, BMI, eGFR, albuminuria, systolic blood pressure, insulin-free treatment burden, and prior acute care use. SHAP summary plots are interpreted in terms of both magnitude and direction, so elevated HbA1c, longer disease duration, poor adherence, lower eGFR, and albuminuria would generally be expected to increase predicted progression risk. Partial dependence or accumulated local effect plots are then used for continuous predictors to check whether SHAP patterns are clinically plausible rather than merely statistical artifacts. Explainable diabetes prediction studies support this workflow because it links model performance to modifiable clinical drivers while preserving transparency for clinicians [14, 15, 20].

Local explanations for individual patients

Local explanations are generated using SHAP waterfall plots for representative high-risk, medium-risk, and low-risk patients selected from the test set. For a high-risk patient, the explanation may show that elevated baseline HbA1c, rapid HbA1c increase, long diabetes duration, low medication possession ratio, obesity, and albuminuria push the predicted risk above the intervention threshold. For a low-risk patient, stable HbA1c, preserved eGFR, absence of albuminuria, consistent medication refills, and lower BMI may pull the prediction below the threshold despite older age. This patient-level interpretation is clinically useful only if the displayed drivers are understandable, actionable, and not dominated by proxies for poor access to care or documentation intensity [5, 17, 23].

Clinical deployment and workflow integration

Real-time risk scoring in EHR

For deployment, the trained model is exported as a version-controlled prediction service using a reproducible preprocessing pipeline and a standardized model object such as ONNX, PMML, or a containerized Python service. During routine visits, the EHR sends the most recent eligible values for HbA1c, BMI, eGFR, albuminuria, medications, adherence proxy, smoking, and blood pressure to the model and receives a 2-year progression probability. Alerts are restricted to clinically meaningful risk thresholds to avoid fatigue, and predictions are suppressed when essential variables are missing or outdated. Prior diabetes risk-score and complication-prediction studies indicate that deployment value depends less on AUROC alone and more on whether the model fits into real clinical workflows without overwhelming clinicians [3, 10, 16].

Clinical decision support interface

The decision support interface presents the predicted 2-year progression risk, risk category, calibration warning if uncertainty is high, and the top positive and negative SHAP contributors for the individual patient. The display emphasizes modifiable factors such as poor medication adherence, elevated HbA1c, obesity, uncontrolled blood pressure, and delayed renal monitoring, while avoiding language that implies causality from prediction alone. Suggested actions may include medication review, adherence counseling, HbA1c recheck, kidney screening, nutrition referral, or closer follow-up, but the model does not automatically prescribe treatment changes. This design aligns with recent interpretable and treatment-response ML work in diabetes, where clinical usefulness depends on translating risk estimates into reviewable and actionable information [18, 22, 33].

Evaluation strategy and validation

Discrimination metrics

The primary discrimination metric is AUROC with 95% confidence intervals estimated by bootstrapping or DeLong-type methods, depending on software availability. AUPRC is reported for the primary endpoint and all rarer secondary outcomes because progression prevalence strongly affects the practical value of a positive prediction. Model comparisons use paired bootstrap testing on the same test-set patients, with elastic net logistic regression treated as the reference model and random forest and XGBoost evaluated as incremental alternatives. Diabetes prediction studies for HbA1c response, cardiovascular outcomes, nephropathy, and composite complications show that ranking performance varies substantially by outcome and cohort, so discrimination results must be interpreted alongside endpoint prevalence and validation design [11, 13, 19].

Calibration assessment

Calibration is assessed by plotting predicted versus observed progression probabilities across deciles of predicted risk, with loess smoothing to inspect miscalibration at clinically relevant ranges. Calibration intercept, calibration slope, and Brier score are reported, and recalibration is considered if the model systematically overpredicts or underpredicts risk in the validation set. Hosmer-Lemeshow or Spiegelhalter-type tests may be

reported, but visual calibration and slope estimates are more informative because large samples can make trivial deviations statistically significant. This is especially important in diabetes EHR models because treatment patterns, coding intensity, and outcome ascertainment can shift over time, reducing the reliability of predicted probabilities even when AUROC remains acceptable [24, 25, 31].

Decision curve analysis and clinical utility

Decision curve analysis estimates net benefit across threshold probabilities from 0.05 to 0.40 and compares each model with treat-all and treat-none strategies. A realistic result is that XGBoost provides net benefit only across intermediate thresholds, such as 0.10 to 0.25, where clinical intervention is plausible and false-positive burden remains acceptable. At very low thresholds, nearly all patients may qualify for additional follow-up, while at very high thresholds, the model may miss many patients who would benefit from earlier intervention. This utility-focused evaluation reflects current diabetes ML literature showing that prediction models must demonstrate practical clinical value, not only statistical improvement [2, 8, 21].

Limitations

Retrospective design and residual confounding

The proposed study is retrospective, so it cannot prove that modifying SHAP-identified predictors will reduce future progression. Unmeasured factors such as diet quality, health literacy, family support, medication affordability, genetic risk, clinician prescribing behavior, and patient preferences may influence both observed predictors and outcomes. Coding quality may also vary across clinics, especially for neuropathy, retinopathy, adherence, and lifestyle variables. These limitations are common in diabetes prediction research and require cautious interpretation of model explanations as associations rather than causal mechanisms [1, 9, 10].

Generalizability and temporal drift

A model trained in one health system may not generalize to populations with different ethnicity distributions, insurance structures, medication access, laboratory testing frequency, or complication coding practices. Temporal drift is also likely because prescribing patterns for GLP-1 receptor agonists, SGLT2 inhibitors, insulin, and cardiovascular risk management continue to change over time. External validation across sites and calendar periods is therefore necessary before clinical deployment, followed by monitoring of calibration, subgroup performance, and alert burden after implementation. Recent multi-cohort and national-data studies support broader validation, but they also show that transportability should be demonstrated empirically rather than assumed [5, 19, 31].

Conclusion

Machine learning offers a practical framework for predicting 2-year disease progression in patients with type 2 diabetes when it is applied to realistic clinical data with careful preprocessing, leakage prevention, and validation. In this manuscript, elastic net logistic regression, random forest, and XGBoost are compared using the same cohort, predictors, and outcome definitions. The expected finding is that gradient boosting improves discrimination compared with regularized logistic regression, while random forest provides a useful intermediate benchmark. The final model is selected not only for AUROC but also for calibration, precision-recall performance, interpretability, and clinical usefulness.

The main technical contribution is a realistic prediction workflow for diabetes progression rather than an idealized model built under perfect data conditions. The workflow addresses class imbalance, missingness, overfitting, and feature instability, all of which are common in EHR-based chronic disease prediction. It also uses SHAP explanations to connect patient-level risk estimates with clinically recognizable drivers such as HbA1c, diabetes duration, adherence, BMI, kidney function, and albuminuria. This makes the model more transparent while preserving the predictive advantages of nonlinear ensemble learning.

The practical value of this approach is risk stratification that can support earlier follow-up, medication review, adherence counseling, renal screening, and lifestyle intervention. A model that identifies patients at elevated risk of HbA1c worsening or insulin initiation could help clinicians allocate limited care-management resources more efficiently. However, the prediction must remain advisory, because model output cannot replace clinical judgment or shared decision-making. The most useful deployment would combine risk probability, explanation, uncertainty, and suggested review actions in a simple EHR interface.

Future work should prioritize multi-site external validation, prospective silent evaluation, and pragmatic trials of model-guided intervention. Calibration should be monitored over time because treatment patterns, coding practices, and patient populations change. Subgroup performance should be evaluated to avoid worsening inequities in care access or treatment intensity. Only after these steps should the model be considered for routine clinical decision support in diabetes management.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine learning methods to predict diabetes complications. J Diabetes Sci Technol. 2018;12(2):295-302.
2. Young JB, Gauthier-Loiselle M, Bailey RA, Manceur AM, Lefebvre P, Greenberg M, et al. Development of predictive risk models for major adverse cardiovascular events among patients with type 2 diabetes mellitus using health insurance claims data. Cardiovasc Diabetol. 2018;17(1):118.
3. Segar MW, Vaduganathan M, Patel KV, McGuire DK, Butler J, Fonarow GC, et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. Diabetes Care. 2019;42(12):2298-306.
4. Lee S, Zhou J, Wong WT, Liu T, Wu WK, Wong IC, et al. Glycemic and lipid variability for predicting complications and mortality in diabetes mellitus using machine learning. BMC Endocr Disord. 2021;21(1):94.
5. Nghiem N, Wilson N, Krebs J, Tran T. Predicting the risk of diabetes complications using machine learning and social administrative data in a country with ethnic inequities in health: Aotearoa New Zealand. BMC Med Inform Decis Mak. 2024;24(1):274.
6. Nicolucci A, Romeo L, Bernardini M, Vespasiani M, Rossi MC, Petrelli M, et al. Prediction of complications of type 2 diabetes: A machine learning approach. Diabetes Res Clin Pract. 2022;190:110013.
7. Hosseini Sarkhosh SM, Esteghamati A, Hemmatabadi M, Daraei M. Predicting diabetic nephropathy in type 2 diabetic patients using machine learning algorithms. J Diabetes Metab Disord. 2022;21(2):1433-41.
8. Mora T, Roche D, Rodríguez-Sánchez B. Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms. Diabetes Res Clin Pract. 2023;204:110910.
9. Tan KR, Seng JJ, Kwan YH, Chen YJ, Zainudin SB, Loh DH, et al. Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review. J Diabetes Sci Technol. 2023;17(2):474-89.
10. Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol. 2023;22(1):259.
11. Nagaraj SB, Sidorenkov G, van Boven JF, Denig P. Predicting short-and long-term glycated haemoglobin response after insulin initiation in patients with type 2 diabetes mellitus using machine-learning algorithms. Diabetes Obes Metab. 2019;21(12):2704-11.
12. Maniruzzaman M, Islam MM, Rahman MJ, Hasan MA, Shin J. Risk prediction of diabetic nephropathy using machine learning techniques: A pilot study with secondary data. Diabetes Metab Syndr. 2021;15(5):102263.
13. Nabrdalik K, Kwiendacz H, Drożdż K, Irlik K, Hendel M, Wijata AM, et al. Machine learning predicts cardiovascular events in patients with diabetes: the silesia diabetes-heart project. Curr Probl Cardiol. 2023;48(7):101694.
14. Yin JM, Li Y, Xue JT, Zong GW, Fang ZZ, Zou L. Explainable machine learning-based prediction model for diabetic nephropathy. J Diabetes Res. 2024;2024(1):8857453.
15. Zhu Y, Zhang Y, Yang M, Tang N, Liu L, Wu J, et al. Machine learning-based predictive modeling of diabetic nephropathy in type 2 diabetes using integrated biomarkers: a single-center retrospective study. Diabetes Metab Syndr Obes. 2024;17:1987-97.
16. Schallmoser S, Zueger T, Kraus M, Saar-Tsechansky M, Stettler C, Feuerriegel S. Machine learning for predicting micro-and macrovascular complications in individuals with prediabetes or diabetes: retrospective cohort study. J Med Internet Res. 2023;25:e42181.
17. Jian Y, Pasquier M, Sagahyroon A, Aloul F. A machine learning approach to predicting diabetes complications. In: Healthcare. 2021;9(12):1712.
18. Slieker RC, Münch M, Donnelly LA, Bouland GA, Dragan I, Kuznetsov D, et al. An omics-based machine learning approach to predict diabetes progression: a RHAPSODY study. Diabetologia. 2024;67(5):885-94.
19. Sang H, Lee H, Lee M, Park J, Kim S, Woo HG, et al. Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts. Sci Rep. 2024;14(1):14966.
20. McDaniel CC, Lo-Ciganic WH, Huang J, Chou C. A machine learning model to predict therapeutic inertia in type 2 diabetes using electronic health record data. J Endocrinol Invest. 2024;47(6):1419-33.
21. Colmenares-Mejia CC, García-Suaza AF, Rodríguez-Lesmes P, Lochmuller C, Atehortúa SC, Camacho-Cogollo JE, et al. Predicting diabetes mellitus metabolic goals and chronic complications transitions—analysis based on natural language processing and machine learning models. PLoS One. 2025;20(4):e0321258.
22. Musacchio N, Zilich R, Masi D, Baccetti F, Nreu B, Giorda CB, et al. A transparent machine learning algorithm uncovers HbA1c patterns associated with therapeutic inertia in patients with type 2 diabetes and failure of metformin monotherapy. Int J Med Inform. 2024;190:105550.
23. Chandra G, Lavikainen P, Siirtola P, Tamminen S, Ihalapathirana A, Laatikainen T, et al. Explainable prediction of long-term glycated hemoglobin response change in Finnish patients with type 2 diabetes following drug initiation using evidence-based machine learning approaches. Clin Epidemiol. 2025;17:225-40.
24. Aagaard A, Röttger R, Johnson EK, Olsen KR. Comparing the predictive performance of diabetes complications using administrative health data and clinical data. Sci Rep. 2025;15(1):33035.

25. Sim R, Chong CW, Loganadan NK, Adam NL, Hussein Z, Lee SW. Comparison of a chronic kidney disease predictive model for type 2 diabetes mellitus in Malaysia using Cox regression versus machine learning approach. Clin Kidney J. 2023;16(3):549-59.
26. Negreiros AB, Ory MG. Navigating uncertain outcomes: Returning genomic results in children with developmental delays. Asian J Ethics Health Med. 2024;4:20-7. doi:10.51847/grOfZd8oyo
27. Karatas KS. First episode psychotic disorder and COVID-19: A case study. Bull Pioneer Res Med Clin Sci. 2024;4(1):19-23. doi:10.51847/VP5xOKglSX
28. Figueroa-Valverde L, Marcela R, Alvarez-Ramirez M, Lopez-Ramos M, Mateu-Armand V, Emilio A. Statistical data from 1979 to 2022 on prostate cancer in populations of Northern and Central Mexico. Bull Pioneer Res Med Clin Sci. 2024;4(1):24-30. doi:10.51847/snclnafVdg
29. Belfiore CI, Galofaro V, Cotroneo D, Lopis A, Tringali I, Denaro V, et al. Studying the effect of mindfulness, dissociative experiences, and feelings of loneliness in predicting the tendency to use substances in nurses. J Integr Nurs Palliat Care. 2024;5:1-7. doi:10.51847/LASijYayRi
30. Wolderslund M, Kofoed P, Ammentorp J. Investigating the effectiveness of communication skills training on nurses' self-efficacy and quality of care. J Integr Nurs Palliat Care. 2024;5:14-20. doi:10.51847/55M0sHLo3Z
31. Younus M, Munna MT, Alam MM, Allayear SM, Ara SJ. Prediction model for prevalence of type-2 diabetes mellitus complications using machine learning approach. In: Data Management and Analysis: Case Studies in Education, Healthcare and Beyond. Cham: Springer Int Publ; 2019. pp. 103-16.
32. Zhou D, Shao L, Yang L, Chen Y, Zhang Y, Yue F, et al. A machine learning model for predicting diabetic nephropathy based on TG/Cys-C ratio and five clinical indicators. Diabetes Metab Syndr Obes. 2025;18:955-67.
33. Garg S, Kitchen R, Gupta R, Pearson E. Applications of AI in predicting drug responses for type 2 diabetes. JMIR Diabetes. 2025;10(1):e66831.
34. Kęska M, Suchy W. Cardiovascular risk and systemic inflammation in rheumatoid arthritis: A comparative analysis with psoriatic arthritis. J Med Sci Interdiscip Res. 2024;4(2):30-40. doi:10.51847/PvcqitKMgB
35. Noor H, Sabău D, Coţe A, Mihetiu AF, Pirvut V, Mălinescu B, et al. Advancements in esophageal stricture treatment: The role of stents in benign and malignant conditions. J Med Sci Interdiscip Res. 2024;4(2):47-52. doi:10.51847/LtuxAzRl0M
36. Schneider TL, Krüger BE. Breast cancer-specific mortality in stage IV patients with small tumors: Insights from a population-based cohort. Arch Int J Cancer Allied Sci. 2025;5(2):1-12. doi:10.51847/b9vFcweAVg
37. Abdullah NA, Zulkifli MI, Mohamed AS. Refinement of the 8th AJCC staging system for medullary thyroid cancer: Integrating tumor size and lymph node characteristics with SEER and multicenter validation. Arch Int J Cancer Allied Sci. 2025;5(2):34-43. doi:10.51847/R1sIaONOms
38. Lee MJ, Ferreira J. COVID-19 and children as an afterthought: Establishing an ethical framework for pandemic policy that includes children. Asian J Ethics Health Med. 2024;4:1-19. doi:10.51847/haLKYCQorD
39. Petridis PD, Kristo AS, Sikalidis AK, Kitsas IK. A review on trending machine learning techniques for type 2 diabetes mellitus management. In: Informatics. 2024;11(4):70.
40. Joungtrakul J, Smith ID. Exploring the path from organizational justice to organizational citizenship behavior: Job commitment as a mediator. Ann Organ Cult Leadersh Extern Engagem J. 2025;6:31-5. doi:10.51847/DBvez9u8O9
41. Kebe IA, Kahl C, Liu Y. The role of transformational leadership in enhancing employee performance: A study of the Vietnamese banking industry. Ann Organ Cult Leadersh Extern Engagem J. 2025;6:21-30. doi:10.51847/g7jtt7Qgxk
42. Rypel J, Kubacka P, Mykała-Cieśla J, Pająk J, Bulska-Będkowska W, Chudek J. Case presentation of breast adenoid cystic carcinoma. Asian J Curr Res Clin Cancer. 2024;4(1):18-24. doi:10.51847/6eOqq2KFjp
43. Osluf ASH, Shoukeer M, Almarzoog NA. Case report on persistent fetal vasculature accompanied by congenital hydrocephalus. Asian J Curr Res Clin Cancer. 2024;4(1):25-30. doi:10.51847/0gjOEudJNr
44. Jin LW, Tahir NAM, Islahudin F, Chuen LS. Exploring treatment adherence and quality of life among patients with transfusion-dependent thalassemia. Ann Pharm Pract Pharmacother. 2024;4:8-16. doi:10.51847/B8R85qakUv
45. Csep AN, Voiţă-Mekereş F, Tudoran C, Manole F. Understanding and managing polypharmacy in the aging population. Ann Pharm Pract Pharmacother. 2024;4:17-23. doi:10.51847/VdKr0egSln
46. Clark A, Foster H. Network pharmacology integration and experimental verification to elucidate the molecular mechanisms of triptolide in treating membranous nephropathy. Pharm Sci Drug Des. 2025;5:33-47. doi:10.51847/X9UVmVSJ4E
47. Njoroge E, Odhiambo S. Elucidating the therapeutic mechanisms of Agrimonia pilosa Ledeb. extract for acute myocardial infarction via network pharmacology and experimental validation. Pharm Sci Drug Des. 2025;5:48-63. doi:10.51847/eZOWCUj80m
48. Raza S, Khan A, Mehmood F, Farooq U. Nationwide implementation of essential pharmacogenomic testing in the Netherlands: A decision-analytic model of lives saved and cost-effectiveness. Spec J Pharmacogn Phytochem Biotechnol. 2025;5:39-49. doi:10.51847/PUWEymkYkk
49. Musa K, Noor O, Ibrahim M, Saleh A. A validated whole-body PBPK model of dextromethorphan and its metabolites for genotype-based prediction of CYP2D6 phenotype and urinary metabolic ratio. Spec J Pharmacogn Phytochem Biotechnol. 2025;5:50-76. doi:10.51847/xbESBJHHcx

50. Ghiga I, Pitchforth E, Lundborg CS, Machowska A. Bacterial infections and antibiotic resistance in Romanian children: Insights from a hospital-based study. Interdiscip Res Med Sci Spec. 2024;4(2):1-8. doi:10.51847/pISlxaQJVu
51. Kounatidis D, Dalamaga M, Grivakou E, Karampela I, Koufopoulos P, Dalopoulos V, et al. Evaluation of blood-aqueous barrier permeability in response to tetracycline antibiotics under normal and pathological conditions. Interdiscip Res Med Sci Spec. 2024;4(2):9-17. doi:10.51847/wu4fOEjgDv
52. Petronis Z, Golubevas R, Rokicki JP, Guzeviciene V, Sakavicius D, Lukosiunas A. A systematic review and meta-analysis on trigeminal neuralgia linked to neurovascular compression using MRI analysis. J Curr Res Oral Surg. 2025;5:17-24. doi:10.51847/sptZWIrWeo
53. Yu M, Ma Y, Han F, Gao X. Effectiveness of mandibular advancement splint in treating obstructive sleep apnea: A systematic review. J Curr Res Oral Surg. 2025;5:25-32. doi:10.51847/AInSXrD9rc
54. Jagsi R, Lee J, Roselin D, Ira K, Williams J. Do U.S. medical schools follow medical associations' recommendations on paid parental leave for faculty? Ann Pharm Educ Saf Public Health Advocacy. 2025;5:1-11. doi:10.51847/r117In8wdi
55. Wong Y, Lin S, Cheng H, Hsieh T, Hsiue T, Chung H, et al. Understanding the impact of medical humanities on internship training and performance. Ann Pharm Educ Saf Public Health Advocacy. 2025;5:12-21. doi:10.51847/Z1fogzPksy