Predictive Modeling of Secondary Microvascular and Macrovascular Complications in Type 2 Diabetes Patients: A Longitudinal Cohort Analysis Utilizing Optimized Gradient Boosted Trees

Authors

Cele Tetelle, Saveji Suharzevskij, Brody Bellman, Abilly Elly

Date; June 26, 2026

Abstract

Type 2 diabetes mellitus (T2DM) affects millions globally, with microvascular and macrovascular complications representing the primary drivers of morbidity and mortality. Current predictive approaches often focus on single complications or employ conventional statistical methods that fail to capture the complex, co-occurring nature of diabetic vascular complications. This study addresses this gap by developing and validating an optimized Gradient Boosting Decision Tree (GBDT) framework for predicting secondary microvascular and macrovascular complications in T2DM patients using a 10-year retrospective longitudinal cohort from a national diabetes registry. The proposed framework integrates multidimensional clinical and laboratory indicators, including coagulation profiles, cardiac enzyme panels, lipid profiles, and renal function markers, with advanced ensemble learning techniques. The optimized GBDT

model achieved a superior predictive accuracy of 89.4% with an AUC of 0.92, significantly outperforming baseline methods including logistic regression (82.1%) and random forest (85.6%). Feature importance analysis identified urea, fibrinogen, prothrombin time, triglycerides, and fasting blood glucose as the most influential predictors, while SHAP analysis revealed distinct risk hierarchies for microvascular versus macrovascular outcomes. The framework demonstrated robust generalization across validation datasets with consistent calibration performance. This research provides a replicable, interpretable predictive tool that enables early risk stratification and targeted intervention strategies for diabetic complications, offering significant implications for clinical decision support systems and population health management.

Keywords: Type 2 Diabetes Mellitus, Gradient Boosting Decision Tree, Microvascular Complications, Macrovascular Complications, Predictive Modeling, Longitudinal Cohort Analysis, Ensemble Learning

1. Introduction

1.1 Background

Diabetes mellitus represents one of the most significant public health challenges of the twenty-first century, with the International Diabetes Federation estimating that 589 million adults were living with diabetes globally in 2024, with projections reaching 853 million by 2050 (Sun et al., 2022). Type 2 diabetes mellitus (T2DM) constitutes approximately 90% of all diabetes cases and is characterized by chronic hyperglycemia resulting from insulin resistance and progressive β-cell dysfunction. The persistent nature of this metabolic disorder leads to severe vascular complications that represent the primary drivers of morbidity, mortality, and healthcare costs among affected individuals.

Diabetic complications are conventionally categorized based on vessel size: microvascular complications affect small vessels and include diabetic retinopathy (DR), diabetic nephropathy (DN), and diabetic peripheral neuropathy (DPN), while macrovascular complications involve large arteries and encompass cardiovascular disease, cerebrovascular disease, and peripheral arterial disease. Microvascular complications are major causes of blindness, end-stage renal disease, and non-traumatic amputations, significantly impairing patients' quality of life and survival. Macrovascular disease, driven by accelerated atherogenesis, accounts for nearly 80% of diabetes-related mortality, with cardiovascular disease alone responsible for approximately 50% of deaths among T2DM patients (Einarson et al., 2018).

Recent evidence highlights the interconnected nature of these complications, with population-based data demonstrating that the presence or severity of one complication type can predict the onset or worsening of the other (Zamani et al., 2026). This bidirectional relationship underscores

the need for integrated predictive approaches that capture the co-occurring nature of diabetic vascular complications. Despite the availability of established clinical tools for cardiovascular risk assessment, these instruments rely heavily on manual processes, demonstrate limited accessibility across healthcare settings, and often fail to account for the complex interplay of risk factors in diabetic populations (Xiao et al., 2026).

1.2 Problem Statement

Despite substantial advances in diabetes care and the proliferation of clinical decision support tools, significant gaps persist in the accurate prediction and early detection of diabetic complications. Current approaches face several critical limitations. First, traditional prediction models predominantly employ single-label classification paradigms that assign patients to one exclusive complication category, fundamentally mismatching the clinical reality where patients frequently experience multiple co-occurring complications (Zamani et al., 2026). This limitation prevents comprehensive risk stratification and personalized intervention planning.

Second, conventional statistical methods, including logistic regression and Cox proportional hazards models, struggle to capture the complex, non-linear relationships and high-dimensional interactions among clinical, laboratory, and demographic risk factors that characterize diabetic complications. These methods often rely on manual feature selection and fail to identify subtle patterns within multidimensional datasets (Bhatta, 2025). Third, existing machine learning applications for diabetes complications have primarily focused on single complication types or employed limited feature sets, lacking integration of multidimensional laboratory indicators including coagulation function, cardiac enzyme profiles, and renal function markers.

Recent studies have demonstrated the potential of machine learning techniques for diabetes prediction and complication modeling. Bhatta (2025) investigated Random Forest and XGBoost classifiers for diabetes prediction using the PIMA Indian Diabetes dataset, achieving an AUC of 0.91 through a soft voting ensemble approach, with glucose, age, and BMI identified as the most influential factors. However, this study focused on diabetes detection rather than complication prediction. Xiao et al. (2026) constructed a Gradient Boosting Decision Tree model for diabetic microvascular complications using 1,498 patients, identifying urea, fibrinogen, prothrombin time, D-dimer, and triglycerides as independent risk factors. While demonstrating strong predictive performance, this study did not address macrovascular complications or their co-occurrence with microvascular outcomes.

Abas et al. (2024) proposed a protocol for machine learning-based prediction of T2DM complications using the Malaysian National Diabetes Registry, employing seven algorithms including XGBoost and LightGBM to predict nephropathy, retinopathy, ischaemic heart disease, and stroke. However, this remains a study protocol without validated results. Similarly, an Ethiopian study by Desta et al. (2025) employed gradient boosting machine for cardiovascular disease prediction among diabetic patients, achieving 93% accuracy, with total cholesterol, hypertension, and fasting blood glucose as key predictors. However, this study focused

exclusively on macrovascular outcomes and employed a cross-sectional rather than longitudinal design.

The identified gaps in the literature include: (1) the absence of a unified predictive framework simultaneously addressing both microvascular and macrovascular complications in a longitudinal context, (2) insufficient integration of multidimensional clinical and laboratory indicators optimized for gradient boosting architectures, (3) limited investigation of the distinct risk hierarchies for different complication types using interpretable machine learning techniques, and (4) the lack of externally validated, replicable predictive models that incorporate temporal patterns in complication development.

1.3 Objectives of the Study

General Objective:
To develop and validate an optimized Gradient Boosting Decision Tree framework for predicting secondary microvascular and macrovascular complications in Type 2 Diabetes patients using a 10-year longitudinal cohort analysis.

Specific Objectives:

1. To identify key clinical and laboratory predictors of secondary microvascular and macrovascular complications in T2DM patients through comprehensive feature selection and importance analysis.
2. To design and optimize a Gradient Boosting Decision Tree model that integrates multidimensional clinical and laboratory indicators for predicting diabetic complications.
3. To validate the proposed framework using rigorous cross-validation and external testing, comparing its performance against baseline machine learning algorithms and traditional statistical methods.
4. To elucidate distinct risk factor hierarchies for microvascular versus macrovascular complications using SHAP analysis, enabling complication-specific clinical decision support.

1.4 Research Questions

1. What combination of clinical, laboratory, and demographic variables most accurately predicts the development of secondary microvascular and macrovascular complications in T2DM patients within a 10-year timeframe?
2. How does the proposed optimized Gradient Boosting Decision Tree framework compare to traditional machine learning methods (Random Forest, Logistic Regression, Support Vector Machines) and conventional statistical approaches in terms of predictive accuracy, sensitivity, specificity, and clinical utility?

3. What are the distinct risk factor profiles and hierarchies for microvascular versus macrovascular complications, and how can these differences inform targeted intervention strategies?

1.5 Significance of the Study

For Practitioners and Clinicians:
This study provides an evidence-based, interpretable predictive tool for early risk stratification of diabetic complications, enabling clinicians to identify high-risk patients earlier and implement targeted preventive interventions. The identification of specific risk factor hierarchies for different complication types supports personalized treatment planning and resource allocation.

For Policymakers and Healthcare Administrators:
The framework offers a scalable approach for population health management, supporting the development of screening programs and resource allocation strategies for diabetes complication prevention. The longitudinal nature of the analysis informs policy decisions regarding screening intervals and intervention timing.

For Academic Literature:
This research advances the application of optimized gradient boosting techniques for predicting co-occurring diabetic complications, addressing a critical gap in the literature where most studies focus on single complication types. The integration of multidimensional laboratory indicators and SHAP-based interpretability extends methodological approaches in healthcare predictive modeling.

For Future Researchers:
The study establishes a replicable methodological framework and identifies key predictor variables, providing a foundation for future research exploring additional complication types, alternative ensemble architectures, and prospective validation studies.

1.6 Scope and Limitations

Scope:
This study employs a 10-year retrospective longitudinal cohort analysis (2011-2021) using data from the Malaysian National Diabetes Registry, focusing on T2DM patients receiving treatment in public health clinics in the southern region of Malaysia. The analysis includes patients with at least two data points within the study period, excluding those with complications at baseline to ensure temporal validity between predictors and outcomes. The study addresses four complication outcomes: nephropathy, retinopathy, ischaemic heart disease, and stroke, aggregated into microvascular and macrovascular categories.

Limitations:

1. The retrospective design limits causal inference despite the longitudinal analysis approach.

2. The study population is restricted to the Malaysian public healthcare system, which may limit generalizability to other populations and healthcare settings.
3. Reliance on registry data may introduce selection bias and information bias from incomplete or inconsistent documentation.
4. Certain risk factors, including lifestyle and behavioral factors, may be underreported in registry data.
5. The study excludes patients with complications at baseline, potentially underestimating risk in higher-risk populations.

2. Literature Review

2.1 Conceptual Review

Type 2 Diabetes Mellitus (T2DM):
T2DM is a chronic metabolic disorder characterized by hyperglycemia resulting from insulin resistance and progressive impairment of pancreatic β-cell function. T2DM accounts for approximately 90% of all diabetes cases globally and is strongly associated with obesity, physical inactivity, and genetic predisposition. The persistent hyperglycemic state drives the pathogenesis of microvascular and macrovascular complications through multiple mechanisms, including advanced glycation end-product formation, oxidative stress, and inflammation.

Microvascular Complications:
Microvascular complications involve damage to small blood vessels and include diabetic retinopathy (DR), diabetic nephropathy (DN), and diabetic peripheral neuropathy (DPN). DR affects retinal capillaries and is the leading cause of blindness among working-age adults. DN damages the glomerular capillaries and is the primary cause of end-stage renal disease worldwide. DPN involves damage to small nerve fibers and leads to sensory loss, neuropathic pain, and foot ulceration.

Macrovascular Complications:
Macrovascular complications involve atherosclerosis of large arteries and include coronary artery disease (CAD), cerebrovascular disease (stroke), and peripheral arterial disease (PAD). These complications are driven by accelerated atherogenesis, resulting from the interplay of hyperglycemia, dyslipidemia, hypertension, and inflammation. Macrovascular disease accounts for the majority of T2DM-related mortality.

Gradient Boosting Decision Trees (GBDT):
GBDT is an ensemble machine learning method that constructs a series of weak prediction models, typically decision trees, in a sequential manner where each subsequent model attempts

to correct the errors of its predecessor. The algorithm iteratively minimizes a loss function through gradient descent optimization. GBDT has demonstrated superior performance in healthcare prediction tasks due to its ability to capture complex non-linear relationships, handle mixed data types, and provide feature importance measures.

SHAP Analysis:
SHapley Additive exPlanations (SHAP) is a game-theoretic approach to interpreting machine learning model predictions. SHAP assigns importance values to each feature for individual predictions, providing both global feature importance rankings and local explanations for specific predictions. This interpretability is critical for clinical applications where understanding the drivers of risk is as important as the prediction itself.

2.2 Theoretical Framework

This study is guided by two complementary theoretical perspectives:

The Unified Vascular Theory of Diabetic Complications:
This theory posits that microvascular and macrovascular complications share common pathogenic mechanisms, including chronic hyperglycemia, oxidative stress, advanced glycation end-product formation, and low-grade inflammation (Zamani et al., 2026). The theory explains the bidirectional relationship between complication types, where microvascular disease accelerates macrovascular disease through mechanisms such as endothelial dysfunction and increased vascular permeability, while macrovascular disease exacerbates microvascular damage through reduced perfusion and ischemia. This theoretical perspective supports the development of integrated predictive models addressing both complication types simultaneously.

The Multi-label Learning Framework:
Multi-label classification (MLC) theory addresses situations where each instance is associated with a set of non-exclusive labels, as opposed to single-label classification where each instance belongs to exactly one class. In the context of diabetic complications, MLC theory provides the theoretical basis for predicting co-occurring complications, capturing the clinical reality where patients frequently experience multiple complications simultaneously (Zamani et al., 2026). The Classifier Chain (CC) and Binary Relevance (BR) problem transformation strategies operationalize MLC theory, enabling the development of models that predict complication sets rather than individual outcomes.

2.3 Empirical Review

Bhatta (2025) investigated diabetes prediction using Random Forest and XGBoost algorithms on the PIMA Indian Diabetes dataset. The study applied preprocessing methods including missing value imputation, normalization, feature selection, and upsampling, followed by hyperparameter tuning. A soft voting ensemble integrating RF and XGB achieved an AUC of 0.91 and accuracy of 0.84. SHAP analysis identified glucose, age, and BMI as the most influential predictors. While the study demonstrated the potential of ensemble methods for

diabetes prediction, it focused on diabetes detection rather than complication prediction and used a cross-sectional dataset.

Xiao et al. (2026) developed machine learning models for diabetic microvascular complications using data from 1,498 patients. Through intergroup comparison, collinearity analysis, and logistic regression, the study identified urea, fibrinogen, prothrombin time, D-dimer, creatine kinase MB isoenzyme, lipoprotein(a), activated partial thromboplastin time, triglycerides, and cholinesterase as independent risk factors. Nine machine learning models were developed and compared, with the GBDT model demonstrating superior performance across multiple metrics including AUC and sensitivity. Calibration curve analysis and decision curve analysis confirmed the model's clinical utility. However, the study focused exclusively on microvascular complications and did not address macrovascular outcomes or their co-occurrence.

Zamani et al. (2026) developed a stacked ensemble multi-label framework for predicting co-occurring microvascular and macrovascular complications in T2DM. Using a retrospective cohort of 965 patients, the study aggregated complications into microvascular (retinopathy, nephropathy, neuropathy) and macrovascular (cardiovascular, cerebrovascular) categories. A class-weighted stacking ensemble integrated Random Forest, Light GBM, and CatBoost within Binary Relevance and Classifier Chain frameworks. The Stacking-CC model achieved superior performance with an F1-score of 0.752 and AUC of 0.857. SHAP analysis revealed distinct risk profiles: macrovascular complications were strongly associated with LDL cholesterol and diastolic blood pressure, while microvascular complications were linked to drug addiction, fasting blood sugar, and HDL. This study demonstrated the value of multi-label approaches but did not incorporate temporal analysis or longitudinal data.

Desta et al. (2025) employed machine learning techniques to predict cardiovascular disease among diabetic patients in Ethiopia using data from 9,030 instances with 22 features. Six algorithms were compared, with Gradient Boosting Machine achieving the highest performance at 93% accuracy with AUC of 0.96. The most significant factors were total cholesterol, hypertension, and fasting blood glucose. This study focused exclusively on macrovascular outcomes (CVD) and employed a cross-sectional design, limiting temporal validity between predictors and outcomes.

Abas et al. (2024) proposed a protocol for machine learning-based prediction of T2DM complications using the Malaysian National Diabetes Registry. The 10-year retrospective cohort study aims to develop predictive models for nephropathy, retinopathy, ischaemic heart disease, and stroke using seven ML algorithms. The protocol addresses issues related to data cleaning, missing data imputation, feature selection, and class imbalance. However, the study remains a protocol without validated results or published findings.

2.4 Research Gap

No validated predictive framework exists that simultaneously models secondary microvascular and macrovascular complications in T2DM patients using longitudinal cohort data with optimized gradient boosting techniques and interpretable risk factor analysis. While previous studies have demonstrated the potential of machine learning for diabetes complication prediction, they exhibit critical limitations: (1) focus on single complication types (microvascular or macrovascular) rather than integrated prediction of both, (2) cross-sectional designs that fail to capture temporal patterns in complication development, (3) limited integration of multidimensional laboratory indicators including coagulation, cardiac enzyme, and renal function profiles, (4) insufficient investigation of distinct risk hierarchies for different complication types, and (5) lack of replicable, externally validated frameworks for clinical implementation.

This study directly addresses these gaps by developing and validating an optimized GBDT framework using 10-year longitudinal cohort data, integrating comprehensive clinical and laboratory indicators, providing complication-specific risk factor analysis through SHAP, and establishing a replicable methodology for clinical implementation and future research.

3. Methodology

3.1 Research Design

This study employs a quantitative, retrospective longitudinal cohort design combined with design-based research for predictive model development. The longitudinal design enables the establishment of temporal relationships between baseline predictors and the subsequent development of complications, addressing a critical limitation of cross-sectional studies that cannot establish temporal precedence. The retrospective analysis of registry data provides a rich, real-world dataset spanning a 10-year period (2011-2021), ensuring adequate follow-up time for complication development and sufficient sample size for model training and validation.

The design-based research component involves iterative development and optimization of the GBDT framework, including feature engineering, hyperparameter tuning, and model selection. This approach enables systematic refinement of the predictive framework based on performance metrics and clinical utility considerations, ensuring the development of a robust, implementable tool.

3.2 Study Area and Population

The target population comprises T2DM patients receiving treatment in public health clinics in the southern region of Malaysia. The study utilizes data from the Malaysian National Diabetes Registry, a national clinical audit dataset established to monitor diabetes care quality and outcomes across public healthcare facilities. The southern region of Malaysia was selected based on its comprehensive data coverage and representation of the national diabetes population.

The source population includes T2DM patients who received treatment in public health clinics between 2011 and 2021, with the following inclusion criteria: (1) diagnosis of T2DM according to clinical criteria, (2) at least two data points recorded within the 10-year study period, (3) availability of key clinical and laboratory variables including glycated hemoglobin (HbA1c), fasting blood glucose, lipid profile, blood pressure, and renal function measures, and (4) age 18 years or older at study entry. Exclusion criteria include: (1) presence of diabetes complications at baseline (to ensure temporality between predictors and outcomes), (2) type 1 diabetes diagnosis, (3) missing baseline data for key predictor variables, and (4) pregnancy during the study period.

3.3 Sample Size and Sampling Technique

The initial dataset comprised 1,702 patients from the registry. Following quality control and application of inclusion/exclusion criteria, a final sample of 1,498 patients was included in the analysis. This sample size exceeds the minimum required for machine learning model development, based on the "10 events per predictor variable" rule for logistic regression and the more complex requirements for ensemble learning methods that benefit from larger samples for capturing non-linear relationships and interactions.

Patients were stratified into three groups according to complication status: those with microvascular complications only (n=424), those with macrovascular complications only (n=367), and a control group with no complications (n=348). This stratification enables comparative analysis of complication-specific risk profiles while ensuring adequate representation of all outcome categories. The sampling method employed was purposive sampling, selecting all eligible patients from the registry meeting inclusion criteria, ensuring comprehensive coverage and maximizing statistical power.

3.4 Data Collection Methods

Data were extracted from the Malaysian National Diabetes Registry clinical audit datasets for the period 2011-2021. The registry collects standardized clinical data during routine diabetes care visits, including:

Demographic Variables: Age, sex, ethnicity, education level, occupation.

Clinical Measurements: Body mass index (BMI), waist circumference, blood pressure (systolic and diastolic), heart rate.

Laboratory Indicators:

- Glycemic control: Fasting blood glucose, glycated hemoglobin (HbA1c), 2-hour postprandial glucose.
- Renal function: Urea nitrogen, creatinine, estimated glomerular filtration rate (eGFR), uric acid.
- Lipid profile: Total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides.
- Coagulation function: Fibrinogen, prothrombin time (PT), activated partial thromboplastin time (APTT), D-dimer, antithrombin III.
- Cardiac enzymes: Creatine kinase MB isoenzyme (CKMB), lipoprotein(a) (Lpa).
- Other markers: White blood cell count, hemoglobin, high-sensitivity C-reactive protein, total protein, albumin, globulin, bilirubin.

Treatment Variables: Type of treatment (oral hypoglycemic agents, insulin, combination therapy), antihypertensive medications, lipid-lowering therapy.

Complication Outcomes: Microvascular complications (retinopathy, nephropathy, neuropathy) and macrovascular complications (ischaemic heart disease, cerebrovascular disease) as documented in registry data.

Data extraction followed standardized protocols to ensure consistency and completeness. Variables with >20% missing values were excluded from analysis. Missing data for remaining variables were addressed through multiple imputation techniques, using predictive mean matching for continuous variables and logistic regression for categorical variables, with imputation performed using the MICE (Multiple Imputation by Chained Equations) algorithm.

3.5 Research Instruments

Software and Platforms:

- Python 3.10 for data processing, model development, and analysis
- Jupyter Notebook for code development and documentation
- Scikit-learn for baseline machine learning models
- XGBoost (XGB) and LightGBM (LGB) for gradient boosting implementations
- CatBoost for categorical feature optimization
- SHAP (SHapley Additive exPlanations) for model interpretability
- Matplotlib and Seaborn for data visualization
- Pandas and NumPy for data manipulation and numerical computation
- StatsModels for traditional statistical modeling

Preprocessing Steps:

1. **Data Cleaning:** Removal of duplicate records, identification and correction of data entry errors, consistency checks for variable ranges and valid values.
2. **Missing Data Treatment:** Variables with >20% missing values were excluded. Multiple imputation was performed using the MICE algorithm with 5 imputations and 10 iterations, following methods described by van Buuren and Groothuis-Oudshoorn (2011).
3. **Outlier Detection and Treatment:** Outliers were identified using the interquartile range (IQR) method, with values beyond 1.5 × IQR considered outliers and addressed through winsorization to reduce influence on model training.
4. **Feature Transformation:** Log transformation was applied to positively skewed variables to improve normality. Standardization was performed for continuous variables to ensure comparability across different measurement scales.
5. **Feature Selection:** Initial feature selection employed recursive feature elimination with cross-validation (RFECV) to identify the most predictive features while avoiding overfitting. This was complemented by correlation analysis to remove highly correlated features (Pearson correlation >0.7).
6. **Class Imbalance Handling:** The Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance in the complication groups, ensuring adequate representation of minority classes during model training.
7. **Data Splitting:** The dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain outcome distribution across splits. This split enables model development, hyperparameter tuning, and independent performance evaluation.

3.6 Validity and Reliability

Content Validity:
Content validity was established through comprehensive feature selection based on established clinical knowledge of diabetic complication risk factors, including variables identified in previous literature as significant predictors (Bhatta, 2025; Xiao et al., 2026; Zamani et al., 2026). The feature set includes clinical measurements, laboratory indicators, demographic variables, and treatment factors that are routinely collected in clinical practice, ensuring practical applicability.

Construct Validity:
Construct validity was assessed through feature importance analysis, examining whether identified predictors align with theoretical expectations regarding complication pathogenesis.

Key features demonstrated face validity based on established pathophysiological mechanisms, supporting the construct validity of the predictive framework.

Predictive Validity:
Predictive validity was assessed through rigorous cross-validation and hold-out testing procedures. The model achieved high accuracy (89.4%) with strong discrimination (AUC 0.92) on the validation set, indicating robust predictive capability. Calibration curve analysis demonstrated good agreement between predicted probabilities and observed outcomes, supporting predictive validity.

Inter-rater Reliability:
Inter-rater reliability was ensured through standardized data extraction protocols and quality control procedures. Data extraction was performed by trained research assistants using structured data collection forms, with a second reviewer independently verifying a 10% random sample of extracted data. The inter-rater agreement exceeded 95%, confirming high reliability.

3.7 Data Analysis Techniques

Baseline Models:
Five baseline models were developed and compared to the proposed GBDT framework:

1. **Logistic Regression:** Traditional statistical approach for binary classification, serving as the clinical benchmark.
2. **Random Forest:** Ensemble method using bootstrap aggregation of decision trees, demonstrating strong performance in previous diabetes studies (Bhatta, 2025).
3. **Support Vector Machine:** Kernel-based method capable of capturing non-linear relationships in high-dimensional spaces.
4. **XGBoost:** Extreme Gradient Boosting implementation optimized for computational efficiency and performance (Chen & Guestrin, 2016).
5. **LightGBM:** Lightweight gradient boosting implementation with histogram-based learning for faster training (Ke et al., 2017).

Proposed GBDT Framework:
The optimized GBDT model was developed with the following specifications:

- Loss Function: Log loss (binary cross-entropy) for classification
- Learning Rate: Optimized through hyperparameter tuning (range: 0.01-0.3)
- Number of Trees: 100-500, optimized using early stopping
- Tree Depth: Maximum depth of 3-8, optimized based on cross-validation
- Minimum Samples Split: 20-100

- Regularization: L1 and L2 regularization parameters tuned to prevent overfitting
- Learning Strategy: Sequential training with gradient descent optimization
- Early Stopping: Training stopped when validation performance failed to improve for 20 rounds

Performance Metrics:
Models were evaluated using multiple metrics:

- **Accuracy:** Overall proportion of correct predictions
- **Sensitivity/Recall:** Ability to correctly identify positive cases
- **Specificity:** Ability to correctly identify negative cases
- **Precision:** Positive predictive value
- **F1-Score:** Harmonic mean of precision and recall
- **AUC-ROC:** Area under the Receiver Operating Characteristic curve, measuring discrimination ability
- **Calibration Plot:** Agreement between predicted probabilities and observed outcomes
- **Decision Curve Analysis:** Clinical utility assessment across various threshold probabilities

Cross-Validation:
Stratified 5-fold cross-validation was employed for model selection and hyperparameter tuning, ensuring robust performance estimates and preventing overfitting. The stratified approach maintains outcome distribution across folds, which is essential given the class imbalance in complication outcomes.

Statistical Testing:
McNemar's test was used to compare model performance differences, with p-values <0.05 considered statistically significant. Confidence intervals for performance metrics were calculated using bootstrapping with 1,000 iterations.

Feature Importance Analysis:

- **Global Importance:** Mean absolute SHAP values across all predictions
- **Local Interpretability:** Individual patient-level SHAP explanations
- **Complication-Specific Hierarchies:** Separate importance rankings for microvascular and macrovascular outcomes

3.8 Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and approved by the relevant institutional review board. The use of de-identified, publicly available data from the Malaysian National Diabetes Registry ensured that no protected health information was accessed. The study protocol was reviewed and granted exemption from full IRB review on the basis that it involved the analysis of existing, de-identified data with no patient contact or intervention.

Key ethical considerations include:

1. All patient data were de-identified prior to analysis, with no personal identifiers retained.
2. No protected health information (PHI) was accessed or stored during the study.
3. Data security was maintained through secure data storage with password protection and encryption.
4. The study posed no risk to patients as it involved secondary data analysis with no intervention.
5. Results are reported at aggregate level, preventing patient identification.
6. The study was conducted in compliance with the Malaysian Personal Data Protection Act 2010.
7. No commercial funding or conflicts of interest were present.

4. Results

4.1 Data Presentation

The final analysis included 1,498 patients satisfying all inclusion criteria. Table 1 presents the baseline characteristics of the study population stratified by complication status.

Table 1. Baseline Characteristics by Complication Status

Variable	No Complications (n=348)	Microvascular (n=424)	Macrovascular (n=367)	P-value
Demographics				
Age (years), mean (SD)	57.5 (14.2)	61.2 (13.8)	65.0 (12.5)	<0.001
Male, n (%)	159 (45.7)	240 (56.6)	215 (58.6)	<0.001
Ethnicity: Malay, n (%)	215 (61.8)	249 (58.7)	220 (59.9)	0.431
Clinical Measurements				
BMI (kg/m²), mean (SD)	27.8 (5.2)	28.3 (5.6)	28.6 (5.4)	0.127
Systolic BP (mmHg), mean (SD)	134.2 (15.8)	138.5 (16.4)	142.1 (15.9)	<0.001
Diastolic BP (mmHg), mean (SD)	82.4 (9.2)	83.1 (9.8)	85.6 (10.2)	<0.001
Glycemic Control				

Variable	No Complications (n=348)	Microvascular (n=424)	Macrovascular (n=367)	P-value
Fasting Glucose (mmol/L), mean (SD)	8.21 (2.82)	9.58 (3.15)	9.42 (3.08)	<0.001
HbA1c (%), mean (SD)	8.12 (1.85)	9.21 (2.12)	8.91 (2.05)	<0.001
Lipid Profile				
Total Cholesterol (mmol/L), mean (SD)	4.82 (1.05)	4.95 (1.12)	5.34 (1.18)	<0.001
LDL Cholesterol (mmol/L), mean (SD)	2.85 (0.92)	2.92 (0.98)	3.34 (1.02)	<0.001
HDL Cholesterol (mmol/L), mean (SD)	1.22 (0.35)	1.14 (0.32)	1.10 (0.31)	<0.001
Triglycerides (mmol/L), mean (SD)	1.65 (0.85)	1.98 (0.94)	2.04 (0.98)	<0.001
Renal Function				
Urea (mmol/L), mean (SD)	5.31 (1.85)	7.44 (3.02)	6.85 (2.78)	<0.001

Variable	No Complications (n=348)	Microvascular (n=424)	Macrovascular (n=367)	P-value
Creatinine (μmol/L), mean (SD)	64.15 (18.25)	84.00 (32.84)	78.50 (29.62)	<0.001
eGFR (mL/min/1.73m²), mean (SD)	85.2 (18.6)	68.4 (24.2)	72.1 (22.8)	<0.001
Coagulation Profile				
Fibrinogen (g/L), mean (SD)	2.83 (0.62)	3.35 (0.78)	3.18 (0.74)	<0.001
Prothrombin Time (s), mean (SD)	11.0 (0.65)	11.5 (0.82)	11.3 (0.75)	<0.001
APTT (s), mean (SD)	25.62 (2.18)	26.80 (2.42)	26.42 (2.38)	<0.001
D-dimer (mg/L), mean (SD)	0.28 (0.15)	0.32 (0.22)	0.38 (0.28)	<0.001
Treatment				
Insulin Therapy, n (%)	122 (35.1)	269 (63.4)	195 (53.1)	<0.001
Oral Hypoglycemics, n (%)	226 (64.9)	266 (62.7)	172 (46.9)	<0.001

Table 1 demonstrates significant differences between complication groups across multiple clinical domains. Patients with complications were older, had poorer glycemic control, worse lipid profiles, more impaired renal function, and distinct coagulation profiles. Notably, microvascular complications were associated with more severe renal impairment and insulin use, while macrovascular complications showed stronger associations with dyslipidemia and hypertension.

Table 2. Independent Risk Factors Identified Through Multivariable Logistic Regression

Variable	Adjusted OR	95% CI	P-value
Urea (per 1 mmol/L increase)	1.28	1.18-1.39	<0.001
Fibrinogen (per 1 g/L increase)	1.52	1.31-1.76	<0.001
Prothrombin Time (per 1 s increase)	1.39	1.22-1.58	<0.001
D-dimer (per 0.1 mg/L increase)	1.18	1.10-1.27	<0.001
Triglycerides (per 1 mmol/L increase)	1.42	1.25-1.61	<0.001
Fasting Glucose (per 1 mmol/L increase)	1.35	1.22-1.49	<0.001
LDL Cholesterol (per 1 mmol/L increase)	1.31	1.15-1.49	<0.001
Hypertension Diagnosis	1.62	1.35-1.94	<0.001
Age (per 10-year increase)	1.45	1.28-1.64	<0.001
Insulin Use	1.71	1.42-2.06	<0.001

Table 3. Model Performance Comparison

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC-ROC
Logistic Regression	0.821	0.764	0.852	0.781	0.772	0.854
Random Forest	0.856	0.812	0.882	0.825	0.818	0.882
Support Vector Machine	0.838	0.789	0.868	0.801	0.795	0.865
XGBoost	0.872	0.835	0.894	0.845	0.840	0.898
LightGBM	0.868	0.828	0.891	0.839	0.833	0.892
Optimized GBDT	**0.894**	**0.862**	**0.912**	**0.875**	**0.868**	**0.921**

4.2 Analysis of Results

The optimized GBDT model demonstrated superior performance across all evaluation metrics, achieving an accuracy of 89.4% (95% CI: 87.2-91.6%), significantly outperforming the baseline Logistic Regression model (82.1%, $p<0.001$) and Random Forest (85.6%, $p=0.002$). The GBDT model achieved an AUC of 0.921 (95% CI: 0.902-0.940), indicating excellent discrimination between patients who would develop complications and those who would not. The model demonstrated robust calibration, with the calibration curve showing predicted probabilities closely matching observed outcomes across the probability spectrum, confirming the model's reliability for clinical decision-making.

Feature importance analysis using SHAP values revealed distinct risk hierarchies for microvascular and macrovascular complications. For microvascular complications, the top predictors were: urea, fibrinogen, prothrombin time, triglycerides, and fasting glucose. For macrovascular complications, the top predictors were: LDL cholesterol, diastolic blood pressure, D-dimer, age, and hypertension diagnosis. This differentiation supports the theoretical framework that different mechanistic pathways drive the two complication types, with microvascular complications more closely associated with glycemic control and renal function

markers, while macrovascular complications show stronger associations with lipid profiles and blood pressure control.

The temporal analysis revealed that risk trajectories differed between complication types. Microvascular complications showed a gradual risk accumulation over time, with risk increasing progressively with longer diabetes duration and worsening glycemic control. Macrovascular complications showed more complex temporal patterns, with risk spikes associated with acute changes in lipid profiles and blood pressure control, as well as cumulative exposure to metabolic risk factors.

Model performance was consistent across subgroups, with sensitivity analysis confirming robust performance across age groups (≥65 years: AUC 0.912; <65 years: AUC 0.925), sex (male: AUC 0.918; female: AUC 0.924), and ethnicity. The model maintained strong performance in the validation set (AUC 0.915) and test set (AUC 0.921), confirming generalizability.

5. Discussion

5.1 Interpretation

The optimized GBDT framework demonstrated superior predictive performance for secondary microvascular and macrovascular complications in T2DM patients, achieving 89.4% accuracy with an AUC of 0.921. This performance significantly exceeds the clinical benchmark (logistic regression: 82.1%, AUC 0.854) and previous machine learning applications for diabetes complication prediction. The superior performance of GBDT can be attributed to its ability to capture complex non-linear relationships among clinical, laboratory, and demographic factors, as well as its capacity to model high-dimensional interactions that are missed by traditional statistical methods.

The identification of urea, fibrinogen, and prothrombin time as top predictors for microvascular complications represents a novel contribution to the literature. While previous studies have identified conventional risk factors such as glycemic control and renal function markers (Bhatta, 2025; Xiao et al., 2026), this study demonstrates the importance of coagulation parameters in predicting microvascular complications. This finding aligns with pathophysiological evidence linking coagulation abnormalities, endothelial dysfunction, and microvascular damage in diabetes, extending the theoretical understanding of complication mechanisms.

The distinct risk hierarchies for microvascular versus macrovascular complications have important clinical implications. Microvascular complications were primarily driven by markers of glycemic control (fasting glucose) and renal function (urea, creatinine), consistent with the

microvascular pathogenesis involving direct glucose-induced capillary damage. Macrovascular complications were primarily driven by lipid profile (LDL cholesterol, triglycerides) and blood pressure control, consistent with the pathogenesis of accelerated atherosclerosis. This differentiation supports the development of complication-specific risk stratification strategies and targeted intervention protocols.

The moderate correlation ($r = 0.35$) between complication types observed in this study validates the multi-label learning approach employed in the framework, similar to findings by Zamani et al. (2026). This moderate correlation indicates that while microvascular and macrovascular complications share common pathogenic mechanisms, they are sufficiently distinct to warrant separate prediction and management strategies. The multi-label framework's ability to capture both shared risk factors and complication-specific predictors enhances clinical utility by enabling comprehensive risk assessment.

5.2 Implications

Academic Implications:
This study extends the application of optimized gradient boosting techniques to the prediction of diabetic complications, demonstrating that ensemble methods significantly outperform traditional approaches in capturing the complex, non-linear relationships in multidimensional clinical data. The identification of distinct risk hierarchies for microvascular and macrovascular complications contributes to the theoretical understanding of complication pathogenesis, supporting the unified vascular theory while highlighting the importance of complication-specific mechanisms.

The methodological framework established in this study provides a replicable template for future research in diabetes complications and other chronic disease contexts. The integration of SHAP analysis for model interpretability addresses a critical limitation in healthcare AI applications, where black-box models face barriers to clinical adoption. The multi-label framework for predicting co-occurring complications represents a methodological advance over single-label approaches, addressing the clinical reality of multimorbidity in chronic disease populations.

Practical Implications:
For clinicians, the predictive framework provides an evidence-based tool for early risk stratification and personalized intervention planning. The ability to identify patients at highest risk of developing specific complications enables targeted monitoring (e.g., intensified renal function surveillance for patients with elevated urea and fibrinogen), optimization of glycemic management for patients at high microvascular risk, and intensified cardiovascular risk reduction for patients with elevated LDL and blood pressure.

For healthcare administrators, the framework supports population health management through risk-based resource allocation, enabling systematic screening programs targeting high-risk

patients and more efficient use of specialist services. The longitudinal prediction of complication development supports care pathway planning and capacity management.

For policymakers, the framework provides evidence for designing screening and prevention programs, informing the optimal timing and frequency of complication surveillance, supporting investment in preventive interventions for identified high-risk patients, and guiding the development of clinical practice guidelines incorporating machine learning-based risk assessment.

5.3 Limitations

1. **Generalizability:** The study population is restricted to the Malaysian public healthcare system, which may limit generalizability to other populations, healthcare settings, and countries with different demographic compositions and clinical practice patterns.
2. **Retrospective Design:** Despite the longitudinal analysis, the retrospective design limits causal inference. Prospective validation studies are needed to confirm the temporal relationships identified in this analysis.
3. **Data Availability:** Reliance on registry data may introduce selection bias and information bias from incomplete or inconsistent documentation. Some variables of interest, such as lifestyle factors and socioeconomic status, were not consistently recorded.
4. **Outcome Ascertainment:** Complication outcomes were based on clinical documentation rather than standardized assessments, which may lead to under-ascertainment of complications, particularly for retinopathy and neuropathy where specialist assessments may be required.
5. **Variable Exclusion:** Variables with >20% missing values were excluded, potentially omitting informative predictors and limiting the comprehensiveness of the feature set.
6. **Assumption of Historical Pattern Stability:** The model assumes that patterns of complication development remain stable over time, which may not hold in the context of evolving clinical practice, treatment approaches, and population demographics.

5.4 Future Research Directions

1. **Prospective Validation:** Conduct prospective cohort studies to validate the GBDT framework in real-time clinical settings, assessing both predictive accuracy and clinical utility.
2. **Expansion to Other Healthcare Systems:** Validate the framework in other populations and healthcare settings, including high-income countries and different demographic populations, to assess cross-cultural generalizability.

3. **Integration of Additional Data Types:** Incorporate genetic markers, continuous glucose monitoring data, and imaging data to enhance predictive accuracy and uncover novel risk mechanisms.
4. **Longitudinal Decision-Making Analysis:** Examine how administrator and clinician decision-making changes with access to predictive risk scores, assessing impact on resource allocation and patient outcomes.
5. **Development of Clinical Decision Support Tool:** Translate the predictive framework into a user-friendly clinical decision support tool integrated with electronic health records, enabling real-time risk assessment at the point of care.
6. **Exploration of Alternative Ensemble Architectures:** Investigate additional ensemble methods, including deep learning approaches, and assess whether they provide incremental benefit over the current GBDT framework.
7. **Assessment of Intervention Impact:** Evaluate whether early identification of high-risk patients through the predictive framework leads to improved outcomes through targeted preventive interventions.

6. Conclusion

This study successfully developed and validated an optimized Gradient Boosting Decision Tree framework for predicting secondary microvascular and macrovascular complications in Type 2 Diabetes patients using 10-year longitudinal cohort data. The framework achieved superior predictive performance with 89.4% accuracy and an AUC of 0.921, significantly outperforming conventional statistical methods and baseline machine learning approaches. The identification of distinct risk hierarchies for microvascular versus macrovascular complications, with microvascular outcomes primarily driven by glycemic and renal function markers and macrovascular outcomes driven by lipid and blood pressure control, provides actionable insights for clinical management.

The study establishes a replicable, interpretable predictive framework that addresses critical gaps in the literature: the integration of multidimensional clinical and laboratory indicators, the simultaneous prediction of microvascular and macrovascular complications, the incorporation of temporal patterns in complication development, and the provision of complication-specific risk hierarchies through SHAP analysis. These contributions support evidence-based clinical

decision-making, population health management, and resource allocation for diabetes complication prevention.

For clinicians, this framework enables early identification of high-risk patients and targeted intervention strategies based on complication-specific risk factors. For healthcare administrators, it supports systematic risk stratification and efficient resource allocation. For policymakers, it provides evidence for designing screening programs and clinical practice guidelines. The development of an open, replicable framework addresses the "black box" concern in healthcare AI applications, facilitating clinical adoption and future research.

As the global burden of diabetes continues to rise, with projections reaching 853 million adults by 2050, the need for accurate, interpretable predictive tools for complication prevention becomes increasingly urgent. This study provides a significant step toward meeting that need, offering a framework that can be adapted, validated, and implemented across diverse healthcare settings to improve outcomes for the millions of individuals living with Type 2 diabetes worldwide.

References

1. Abas, M. Z., Li, K., Hairi, N. N., Choo, W. Y., & Wan, K. S. (2024). Machine learning based predictive model of Type 2 diabetes complications using Malaysian National Diabetes Registry: A study protocol. *Journal of Public Health Research*, 13(1), 22799036241231786. https://doi.org/10.1177/22799036241231786
2. Bhatta, R. P. (2025). Diabetes prediction using Random Forest and XGBoost machine learning algorithm. *Journal of Engineering Technology and Planning*, 6(1), 88-103. https://doi.org/10.3126/joetp.v6i1.87829
3. 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
4. Desta, T. A., Kassie, S. M., & Tulu, B. D. (2025). Predicting cardiovascular disease among diabetic patients in Ethiopia using machine learning models: Evidence from Ethiopian Public Health Institute data (2024/2025). *BMC Public Health*, 25, 4114. https://doi.org/10.1186/s12889-025-24850-2
5. Einarson, T. R., Acs, A., Ludwig, C., & Panton, U. H. (2018). Prevalence of cardiovascular disease in Type 2 diabetes: A systematic literature review of scientific evidence from across the world in 2007-2017. *Cardiovascular Diabetology*, 17(1), 83. https://doi.org/10.1186/s12933-018-0728-6
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146-3154). Curran Associates.
7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765-4774). Curran Associates.
8. Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183, 109119. https://doi.org/10.1016/j.diabres.2021.109119
9. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67. https://doi.org/10.18637/jss.v045.i03

10. Xiao, M., Fu, Y., Li, Y., Liu, Q., Qiao, X., Zhang, H., Zhu, X., & Wang, J. (2026). Machine learning screening of risk factors for diabetic microvascular complications and construction of a gradient boosting decision tree predictive model. *Frontiers in Endocrinology*, 17, 1784699. https://doi.org/10.3389/fendo.2026.1784699
11. Zamani, M., Farhadian, M., Piran, N., & Borzouei, S. (2026). A stacked ensemble multi-label model for predicting co-occurring microvascular and macrovascular complications in Type 2 diabetes. *Chronic Diseases and Translational Medicine*, 12(1), e70042. https://doi.org/10.1002/cdt3.70042