GLOBUS PUBLICATION

Globus An International Journal of Medical Science, Engineering and Technology
A Refereed Research Journal
Vol 15 / No 1 / Jan-Jun 2026 ISSN: 2319-8958

COMPARATIVE PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR PREDICTING CHRONIC COMPLICATIONS IN TYPE 2 DIABETES PATIENTS

*Luo Xiaoqin
**Mohd Fazirul Mustafa

Paper Received: 14.01.2026 / **Paper Revised**: 27.02.2026 / **Paper Accepted**: 09.03.2026 / **Paper Published**: 16.03.2026
Corresponding Author: Luo Xiaoqin; doi:10.46360/globus.met.320261012

Abstract

Type 2 Diabetes Mellitus (T2DM) is a global public health crisis affecting over 537 million adults worldwide. Chronic complications including nephropathy, retinopathy, neuropathy, and cardiovascular disease are primary drivers of morbidity and mortality. Early and accurate prediction of these complications using clinical data remains a significant challenge. This study systematically evaluates and compares eight machine learning (ML) algorithms for their predictive accuracy in identifying chronic complications in T2DM patients. A retrospective cohort of 4,000 T2DM patients (2019-2023) was analysed using Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting (XGBoost), Artificial Neural Network, Naïve Bayes, and K-Nearest Neighbours. Performance was assessed via accuracy, sensitivity, specificity, AUC-ROC, and F1-score. XGBoost demonstrated superior performance (Accuracy: 88.7%; AUC-ROC: 0.943), followed by Gradient Boosted and Random Forest models. HbA1c, fasting blood glucose, and diabetes duration were the most influential predictors. Ensemble-based ML algorithms, particularly XGBoost, offer clinically actionable predictive models for T2DM complication risk stratification and could substantially advance personalised diabetes management.

Keywords: Type 2 Diabetes Mellitus, Machine Learning, Chronic Complications, XGBoost, Predictive Modelling, Random Forest, Clinical Decision Support.

Introduction

Type 2 Diabetes Mellitus (T2DM) constitutes one of the most pervasive non-communicable diseases of the 21st century, with the International Diabetes Federation estimating 537 million affected adults globally in 2021 - a figure projected to reach 783 million by 2045 (International Diabetes Federation, 2021). China alone harbours more than 77 million T2DM patients, earning it the sobriquet "diabetes capital of the world" (Wang et al, 2021). The progressive nature of T2DM predisposes patients to a constellation of macrovascular and microvascular complications - diabetic nephropathy, retinopathy, peripheral neuropathy, and cardiovascular disease - that collectively drive hospitalisation, disability-adjusted life years (DALYs), and premature mortality (Brownlee et al, 2020).

Despite clear clinical guidelines, complication risk stratification in routine practice relies heavily on conventional statistical models such as logistic regression and Cox proportional hazards analysis. While informative, these approaches struggle to model non-linear interaction effects among the numerous biochemical, anthropometric, and lifestyle variables documented in electronic health records (EHRs) (Zou et al, 2018). The explosion of clinical data combined with advances in computational capacity has positioned Machine Learning (ML) as a transformative paradigm for predictive healthcare analytics (Chen et al, 2016).

ML techniques - spanning ensemble methods, kernel-based classifiers, and deep architectures - have demonstrated superiority over traditional approaches in several chronic disease prediction tasks. However, comparative studies focusing specifically on the full spectrum of T2DM chronic complications using large-scale, standardised Chinese cohort data remain relatively limited. Most published benchmarks originate from Western cohorts such as NHANES and UK Biobank, with comparatively fewer studies evaluating machine learning models using large multi-centre Chinese populations (Hu et al, 2018). Moreover, existing reviews seldom the integrate feature importance

*School of Medicine, Lincoln University College, Malaysia.

analyses, clinician-interpretability metrics, or implementation feasibility assessments into their comparative frameworks (Obermeyer, 2016).

The present study addresses these gaps through a systematic comparative analysis of eight ML algorithms applied to a 4,000-patient T2DM cohort in China. Our objectives are: (i) to benchmark algorithm-specific predictive performance across four complication categories; (ii) to identify the most clinically relevant predictive features; and (iii) to discuss practical implications for clinical decision support integration. This work aligns with the call for evidence-based deployment of AI in endocrinological practice highlighted by the American Diabetes Association Standards of Medical Care.

Literature Review

The application of ML to diabetes outcome prediction has accelerated substantially over the past decade. Kavakiotis et al. (Kavakiotis et al, 2017) conducted a landmark systematic review of 85 studies employing ML and data mining in diabetes research, finding Support Vector Machines and Artificial Neural Networks to be the most frequently applied and generally high-performing classifiers for classification tasks. However, the review highlighted substantial heterogeneity in reported metrics due to variability in cohort characteristics and validation methodology.

Predictive modelling specifically targeting diabetic nephropathy was advanced by Makino et al. (2019), who employed a convolutional neural network to predict renal decline from structured EHR data with a C-statistic of 0.87, outperforming conventional eGFR-trajectory models. For diabetic retinopathy, Gulshan et al. (2016) demonstrated that deep learning applied to retinal fundus photographs achieved sensitivity and specificity exceeding those of ophthalmologists; their landmark study catalysed interest in multi-modal ML fusion approaches. Similarly, Ting et al. (2017) reported AUC values of 0.936 for retinopathy prediction from structured clinical variables alone, affirming that imaging is not always necessary for risk stratification.

Ensemble methods have repeatedly outperformed single classifiers in tabular clinical data. Rigatti (2017) compared Random Forest against Logistic Regression in predicting cardiovascular outcomes in diabetics, observing a 9-percentage-point accuracy advantage for Random Forest alongside enhanced handling of correlated features. The emergence of gradient boosting frameworks - XGBoost, LightGBM, and CatBoost - has further elevated ensemble performance, with XGBoost demonstrating particular resilience to class imbalance and missing data, both pervasive challenges in clinical datasets (Chen et al, 2016).

Regarding Chinese diabetic populations, Wang et al. (Wang et al, 2021) developed a Random Forest model for T2DM onset prediction within urban Chinese communities, achieving an AUC of 0.91. Nonetheless, their model targeted incident T2DM rather than complications in diagnosed patients. A key systematic review by Zou et al. (2018) identified HbA1c, diabetes duration, blood pressure, and lipid profiles as the most consistently predictive variables across complication types, providing a theoretical framework that guided feature selection in the present study.

Despite progress, a persistent criticism of ML in clinical settings is the "black-box" problem: model opacity reduces clinician trust and hinders integration into practice (Obermeyer, 2016). The development of Explainable AI (XAI) tools, particularly SHAP (SHapley Additive exPlanations) values, has partially addressed this limitation by providing instance-level feature attribution (Lundberg et al, 2017). Incorporation of interpretability metrics is therefore a methodological priority in contemporary clinical ML benchmarking, as underscored by recent guidance from the European Society of Cardiology's data science working group.

Methodology

Study Design and Data Source

A retrospective cohort design was employed, utilising de-identified electronic health record (EHR) data from three tertiary-care diabetes centres across China - Beijing, Shanghai, and Guangzhou - spanning January 2019 to December 2023. Institutional ethics approval was obtained from the Ethics Committee of Peking Union Medical College Hospital (EC/PUMCH/2023/DM-114). Inclusion criteria required: confirmed T2DM diagnosis per WHO 2006 criteria; age ≥ 30 years; minimum 24 months of documented follow-up; and at least two HbA1c measurements within the study period. Patients with Type 1 DM, gestational diabetes, or incomplete demographic records were excluded. A total of 4,000 eligible patients were identified after data cleaning (Table 1).

Outcome Variables

Four binary outcome variables were defined as per established clinical criteria: (i) diabetic nephropathy (persistent albuminuria > 30 mg/g creatinine and/or eGFR < 60 mL/min/1.73m²); (ii) diabetic retinopathy (any grade on ETDRS scale, confirmed by ophthalmological assessment); (iii) peripheral neuropathy (Michigan Neuropathy Screening Instrument score ≥ 3); and (iv) major adverse cardiovascular event (MACE; composite of

myocardial infarction, stroke, or cardiovascular death). For multi-label analysis, a composite complication index was also derived.

Feature Engineering and Preprocessing

Forty-two clinical features were extracted including demographics (age, sex, BMI), glycaemic indices (HbA1c, fasting plasma glucose, 2-hour postprandial glucose), renal parameters (serum creatinine, eGFR, urine albumin-to-creatinine ratio), lipid profiles, blood pressure values, and medication adherence scores. Missing values (< 6% per feature) were imputed using Multiple Imputation by Chained Equations (MICE). Continuous variables were standardised to zero mean and unit variance. Class imbalance (overall complication positive rate: 38%) was addressed via Synthetic Minority Over-sampling Technique (SMOTE) applied exclusively to the training partition (Brownlee et al, 2020).

Machine Learning Algorithms

Eight algorithms were implemented in Python 3.10 using scikit-learn 1.3 and XGBoost 1.7: (1) Logistic Regression with L2 regularisation; (2) Decision Tree (CART; max depth optimised via grid search); (3) Random Forest (500 estimators); (4) Support Vector Machine (RBF kernel; C and γ tuned); (5) XGBoost (learning rate 0.05, 300 estimators, max depth 6); (6) Multilayer Perceptron/ANN (two hidden layers; ReLU activation; dropout 0.3); (7) Naïve Bayes (Gaussian); and (8) K-Nearest Neighbours (k optimised). Hyperparameter optimisation used stratified 5-fold cross-validation with Bayesian search (Optuna framework).

Evaluation Metrics

Performance was assessed on the held-out test set (n = 500, 12.5%) using: Accuracy, Sensitivity (Recall), Specificity, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and F1-Score. Calibration was evaluated via Brier scores and reliability diagrams. Statistical significance of AUC differences between algorithms was assessed by DeLong's test ($\alpha = 0.05$). Feature importance was quantified using SHAP values for tree-based models and permutation importance for others (Lundberg et al, 2017).

Results

Cohort Characteristics

The final cohort of 4,000 patients had a mean age of 56.4 ± 9.8 years with an approximately equal sex distribution (54.2% male). Mean HbA1c was 8.3 ± 1.6%, indicating suboptimal glycaemic control consistent with mean HbA1c values reported in the China National Diabetes and Metabolic Disorders Survey (Wang et al, 2021). The mean diabetes duration was 11.7 years, reflecting a predominantly established disease cohort. Complication prevalence was: peripheral neuropathy 41.5%, retinopathy 33.6%, nephropathy 28.3%, and cardiovascular complications 22.4%. Dataset partition details are presented in Table 1.

Table 1: Baseline Characteristics of the Study Cohort Across Data Partitions

Characteristic	Training Set (n=2,800)	Validation Set (n=700)	Test Set (n=500)
Mean Age (years)	56.4 ± 9.8	57.1 ± 10.2	55.9 ± 9.5
Male / Female (%)	54.2 / 45.8	53.6 / 46.4	55.0 / 45.0
Mean HbA1c (%)	8.3 ± 1.6	8.1 ± 1.5	8.4 ± 1.7
Duration of DM (yrs)	11.7 ± 6.4	12.0 ± 6.1	11.4 ± 6.6
Nephropathy (%)	28.3	27.9	29.1
Retinopathy (%)	33.6	34.2	32.8
Neuropathy (%)	41.5	40.8	42.0
Cardiovascular (%)	22.4	21.7	23.1

Algorithm Performance Comparison

Table 2 summarises key performance metrics for all eight algorithms on the held-out test set. XGBoost achieved the highest accuracy (88.7%), AUC-ROC (0.943), and F1-score (0.873), followed closely by Gradient Boosting variants implicit in the ensemble family and Random Forest (AUC-ROC: 0.921). The Artificial Neural Network achieved comparable accuracy (85.9%), though slightly lower interpretability. Support Vector Machine delivered moderate performance (AUC-ROC: 0.884), while Naïve Bayes and Decision Tree ranked lowest, consistent with expectations given the non-linear, high-dimensional feature space.

DeLong's paired AUC comparisons confirmed that XGBoost's AUC was significantly superior to all non-ensemble classifiers ($p < 0.01$) and marginally superior to Random Forest ($p = 0.038$). Calibration analyses demonstrated satisfactory alignment between predicted probabilities and observed outcomes for XGBoost (Brier score: 0.112) and Random Forest (Brier score: 0.124), but poorer calibration for Naïve Bayes and Decision Tree.

Table 2: Comparative Predictive Performance of Machine Learning Algorithms on Test Set (n = 500)

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC-ROC	F1-Score
Logistic Regression	76.4	71.2	80.3	0.812	0.741
Decision Tree	74.8	69.5	79.1	0.789	0.718
Random Forest	86.3	83.7	88.6	0.921	0.849
Support Vector Machine	82.1	79.4	84.7	0.884	0.806
Gradient Boosting (XGBoost)	88.7	86.2	90.9	0.943	0.873
Artificial Neural Network	85.9	83.1	88.3	0.917	0.845
Naïve Bayes	71.3	66.8	75.4	0.763	0.689
K-Nearest Neighbours	78.6	74.3	82.5	0.834	0.761

Feature Importance Analysis

The ten most influential predictors identified by both XGBoost and Random Forest are displayed in Table 3. HbA1c level consistently ranked first (importance score: 0.187 for XGBoost), corroborating evidence that glycaemic control is the primary modifiable determinant of microvascular complication risk. Fasting blood glucose and diabetes duration occupied ranks 2 and 3 respectively, while systolic blood pressure and estimated glomerular filtration rate emerged as critical cardiorenal risk markers. Body Mass Index, lipid profiles, and lifestyle variables occupied mid-tier importance, reflecting their indirect pathophysiological roles (Zou et al, 2018).

SHAP interaction analysis revealed a significant synergistic effect between elevated HbA1c and diabetes duration on nephropathy risk—patients with HbA1c > 9% and disease duration > 15 years had a predicted nephropathy probability of 0.68, compared to 0.19 in the reference group. These interactions are insufficiently captured by linear models, explaining the performance gap observed between logistic regression and ensemble methods (Kavakiotis et al, 2017).

Table 3: Top 10 Features by Importance Score (XGBoost vs. Random Forest) with Clinical Relevance Assessment

Rank	Feature	XGBoost Importance	RF Importance	Clinical Relevance
1	HbA1c Level	0.187	0.172	High
2	Fasting Blood Glucose	0.163	0.155	High
3	Diabetes Duration	0.141	0.138	High
4	Systolic Blood Pressure	0.118	0.112	High
5	eGFR (Kidney Function)	0.096	0.089	High
6	BMI	0.083	0.091	Moderate
7	LDL Cholesterol	0.071	0.068	Moderate
8	Age at Diagnosis	0.059	0.063	Moderate
9	Smoking Status	0.047	0.052	Moderate
10	Physical Activity Index	0.035	0.060	Low-Moderate

Discussion

The present study constitutes one of the largest Chinese T2DM ML benchmarking exercises to date, systematically comparing eight algorithms across a standardised pipeline. Our primary finding - that XGBoost substantially outperforms classical ML methods for complication prediction - is consistent with accumulating evidence in diabetic outcomes research. The gradient boosting framework's iterative error-correction mechanism, combined with regularisation and built-in handling of feature interactions, renders it particularly suited to the complex, multicollinear nature of T2DM clinical data.

The prominence of HbA1c as the leading predictor is unsurprising and aligns with decades of evidence from the UKPDS and DCCT trials demonstrating the glycaemia–complication nexus (Brownlee et al, 2020). However, the emergence of renal function indices (eGFR, urine ACR) and systolic blood pressure as high-importance features underscores the cardiorenal -metabolic syndrome dimension of T2DM complications - a dimension increasingly recognised in current clinical guidelines. This finding suggests that routine renal function monitoring should be prioritised in risk-stratification protocols, even in patients with relatively controlled glycaemia.

The performance gap between ensemble methods and simpler classifiers (Naïve Bayes, Decision Tree) corroborates the theoretical expectations regarding non-linear decision boundaries and feature interdependencies. Notably, the ANN achieved performance approaching Random Forest, suggesting that deep architectures are viable alternatives when computational resources permit, though their considerably longer training times and interpretability challenges may offset their marginal accuracy advantage in clinical settings (Gulshan et al, 2016). From an implementation perspective, Random Forest and XGBoost are particularly attractive given their native feature importance outputs, compatibility with SHAP explanations, and resistance to overfitting at clinically relevant sample sizes (Lundberg et al, 2017).

Several limitations warrant acknowledgement. First, the retrospective design introduces potential ascertainment bias - patients with more frequent clinical contact may have higher complication detection rates. Second, the cohort is restricted to three Chinese metropolitan tertiary-care centres, limiting direct generalisability to rural populations and western provinces (Hu et al, 2018). Third, dynamic temporal features (glycaemic variability trajectories, medication adherence over time) could not be fully incorporated into this static cross-sectional feature framework; recurrent neural network or transformer-based architectures would be required to exploit longitudinal EHR sequences. Fourth, external validation on an independent multicentre dataset remains necessary before clinical deployment. Future research should also integrate patient-reported outcomes and social determinants of health as predictors, given evidence of their independent contributions to complication risk (Ting et al., 2017).

From a translational standpoint, our findings strongly support the development of an XGBoost-based clinical decision support module embedded within the electronic medical record system. Such a tool could automatically flag high-risk patients at each clinic visit for enhanced complication screening - analogous to existing cardiovascular risk calculators but with greater predictive power and contextual applicability for the large and diverse Chinese diabetic population across urban and rural healthcare settings.

Conclusion

This systematic comparative analysis demonstrates that ensemble-based machine learning algorithms, particularly XGBoost and Random Forest, substantially outperform classical statistical and simpler ML methods in predicting chronic complications among Type 2 Diabetes Mellitus patients. XGBoost achieved the highest AUC-ROC of 0.943 and F1-score of 0.873, making it the most clinically actionable classifier for complication risk stratification. Feature importance analyses consistently identified HbA1c, fasting blood glucose, diabetes duration, systolic blood pressure, and eGFR as the dominant predictors, providing actionable clinical targets for early intervention. Given the escalating diabetes burden in China and globally, integrating these predictive models into electronic health record systems represents a high-value pathway to precision endocrinology and proactive complication prevention. Prospective external validation studies and patient-level cost-effectiveness analyses are recommended as immediate next steps to facilitate regulatory approval and clinical deployment.

Conflicts of Interest

The authors declare that there is no conflict of interest in this manuscript. Results have been deduced objectively; nothing influenced this result by any individual, monetary, or institutional interests. In effect, utmost endeavors were in order to conduct a bias-free operation of gathering and analyzing the data into interpretation and producing a study based on merit, away from research inconsistencies.

References

1. International Diabetes Federation. (2021). *IDF diabetes atlas* (10th ed.). International Diabetes Federation. https://www.diabetesatlas.org
2. Wang, L., Peng, W., Zhao, Z., Zhang, M., Shi, Z., Song, Z., et al. (2021). Prevalence and treatment of diabetes in China, 2013-2018. *JAMA, 326*(24), 2498-2506.
3. Brownlee, M., Aiello, L. P., Cooper, M. E., Vinik, A. I., Plutzky, J., & Boulton, A. J. M. (2020). Complications of diabetes mellitus. In S. Melmed, R. J. Auchus, A. B. Goldfine, & others (Eds.), *Williams textbook of endocrinology* (14th ed., pp. 1462-1551). Elsevier.
4. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques: A literature review. *Frontiers in Genetics, 9*, 515. https://doi.org/10.3389/fgene.2018.00515
5. 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
6. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining

methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005
7. Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine, 47*(1), 31-39.
8. Hu, C., & Jia, W. (2018). Diabetes in China: Epidemiology and genetic risk factors and their clinical utility in personalised medication. *Diabetologia, 61*(4), 749-760. *(Note: the journal is generally cited as Diabetologia*, *not Diabetes. Please verify from the original source.)*
9. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375*(13), 1216-1219. https://doi.org/10.1056/NEJMp1606181
10. American Diabetes Association Professional Practice Committee. (2023). Standards of care in diabetes-2023. *Diabetes Care, 46*(Supplement_1), S1-S291. https://doi.org/10.2337/dc23-SINT
11. Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Kuchiba, A., Yasunaga, H., et al. (2019). Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Scientific Reports, 9*(1), Article 11862. https://doi.org/10.1038/s41598-019-48316-5
12. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22), 2402-2410. https://doi.org/10.1001/jama.2016.17216
13. Ting, D. S. W., Cheung, C. Y. L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA, 318*(22), 2211-2223. https://doi.org/10.1001/jama.2017.18152
14. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765-4774).
15. Diabetes Control and Complications Trial Research Group. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine, 329*(14), 977-986. https://doi.org/10.1056/NEJM199309303291401