Machine Learning Models for Predicting Type 2 Diabetes Complications in Malaysia

Mohamad Zulfikrie Abas, M.P.H.[1], Kezhi Li, Ph.D.[2], Wan Yuen Choo, Ph.D.[1], Kim Sui Wan, Dr.P.H.[3], Noran Naqiah Hairi, Ph.D.[1]

Author Affiliations:[1]Department of Social and Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, [2]Institute of Health Informatics, University College London, London, [3]Institute of Public Health, National Institute of Health, Selangor

Mohamad Zulfikrie Abas
Social and Preventive Medicine Department, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia
+6012-7376131
m_zulfikrie@yahoo.com

Kezhi Li
Institute of Health Informatics, University College London, 222 Euston Road, London, United Kingdom
+4420 3549 5969
ken.li@ucl.ac.uk

Wan Yuen Choo
Social and Preventive Medicine Department, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia
+603-79674756
ccwy@um.edu.my

Kim Sui Wan
Institute for Public Health, National Institutes of Health, Blok B5 & B6, Kompleks NIH,No1, Jalan Setia Murni U13/52,Seksyen U13 Bandar Setia Alam,40170 Shah Alam, Selangor.
+603 3362 7800
kimsui@moh.gov.my

Noran Naqiah Hairi
Social and Preventive Medicine Department, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia
+603-79674756
noran@um.edu.my

Corresponding Authors:

Noran Naqiah Hairi, Social and Preventive Medicine Department, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia; email address noran@um.edu.my

Mohamad Zulfikrie Abas
Social and Preventive Medicine Department, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia; email address m_zulfikrie@yahoo.com

Acknowledgement: We would like to thank the Director General of Health Malaysia for his permission to publish this article.

Author Contribution: M.Z.A. conceptualized the research, performed data analysis, and drafted the manuscript. K.L. provided guidance on machine learning model development and data analysis. K.S.W. contributed to the study design, provided critical revisions, and guided the interpretation of results. W.Y.C and N.N.H. supervised the overall project, provided essential input throughout the study, and contributed to the final review of the manuscript. All authors read and approved the final version of the manuscript for submission.

Statements and declarations

Ethical Approval: This study was approved by the Malaysian Medical Research and Ethics Committee with registration number NMRR ID- 22-00928-MMB (IIR) and received permission from the respective State Health Department.

Consent to Participate: Not applicable

Consent for Publication: Not applicable

Declaration of Conflicting Interest: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Data Availability Statement: The data used in this study belong to the Ministry of Health Malaysia. Access to the data requires an official request and approval from the relevant authorities. Due to confidentiality and ethical considerations, the dataset is not publicly available.

Abstract

This study aimed to develop machine learning (ML) models to predict diabetic complications in patients with Type 2 diabetes (T2D) in Malaysia. Data from the Malaysian National Diabetes Registry and Death Register were used to develop predictive models for five complications: all-cause mortality, retinopathy, nephropathy, ischemic heart disease (IHD), and cerebrovascular disease (CeVD). Accurate predictions may enable targeted preventive intervention and optimal disease management. The cohort comprised 90,933 T2D patients treated at public health clinics in southern Malaysia from 2011 to 2021. Seven ML algorithms were tested, with the Light Gradient Boosting Machine (LGBM) demonstrating the best performance. LGBM models achieved ROC-AUC scores of 0.84 for all-cause mortality, 0.71 for retinopathy, 0.71 for nephropathy, 0.66 for IHD, and 0.74 for CeVD. These findings support integrating ML models, particularly LGBM, into clinical practice for predicting diabetes complications. Further optimization and validation are necessary to enhance applicability across diverse populations.

Keywords: Diabetes complications, Diabetes registry, Machine learning, Predictive models, Type 2 Diabetes

What We Already Know:

- Despite current management, T2D still increases the risk of complications, highlighting the need for better prevention strategies.
- Machine learning (ML) models outperform traditional methods in predicting complex medical outcomes with large, high-dimensional data.
- Limited research on diabetes complications prediction in Malaysian cohorts, often with small sample sizes, restricts the generalizability of findings.

What This Article Adds:

- This study supports integrating ML models into clinical practice to target interventions and slow T2D complication progression.
- The Light Gradient Boosting Machine (LGBM) model outperformed other algorithms tested, showing good performance for four out of five complications tested.
- Using a large dataset of 90,933 T2D patients from Malaysia, this study provides stronger, locally relevant evidence for predicting complications.

Introduction

Diabetes is a major health concern in Malaysia. The prevalence of the disease in the country exceeds the global average and is on an upward trajectory.[1,2] Patients with diabetes are at a higher risk of developing various microvascular and macrovascular complications.[3-5] The increasing prevalence of diabetes may result in a higher number of individuals experiencing diabetes-related complications, posing a challenge for healthcare systems in providing optimal care.[6] Type 2 Diabetes (T2D) accounts for over 90% of all diabetes cases.[1] Risk stratification of T2D patients is crucial for targeted intervention as it facilitates optimal resource allocation, and this approach has demonstrated greater efficiency than population-wide intervention.[7]

The identification of T2D patients at the greatest risk of developing complications remains a complex task despite the known risk factors due to the intricate and dynamic interplay between these factors. Accurate risk stratification can be achieved through the development of a robust prediction model based on local data.[8] Current evidence suggests that machine learning (ML) models generally outperformed non-ML models in predicting diabetes complications in T2D patients, as ML algorithms are well-suited for handling complex and large datasets compared to traditional statistical methods.[9] This is attributed to traditional statistical methods typically having a linear structure, while ML allows for modelling complex relationships between predictors and outcomes. Consequently, ML may effectively utilize high-dimensional data to construct prediction models.[10-14] As ML algorithms leverage big data for predictions, their popularity has increased due to the digitalization of diverse records.

Many studies worldwide have focused on creating prediction models for disease management using ML methods, which are expected to become mainstream.[15] However,

there is a paucity of research on the development of prediction models for diabetes-related complications in Malaysia. Previous studies in Malaysia often utilized small sample sizes, which limits generalizability.[16,17] Therefore, this study aimed to employ ML techniques to construct predictive models for T2D complications using big data acquired from the Malaysian National Diabetes Registry (MNDR) supplemented with data from the Malaysian Death Register (MDR). The T2D complications of interest in this study were all-cause mortality, retinopathy, nephropathy, ischaemic heart disease (IHD) and cerebrovascular disease (CeVD).

Methodology

A detailed protocol for the model-building strategy adopted in this study has been published.[18] This report was structured according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement for the development of prediction models in medicine.

Study design and data source

This was an eleven-year retrospective open cohort study from 2011 to 2021, where the longitudinal dataset used for this study was first extracted and analyzed specifically for this research. The longitudinal dataset was formed by merging eleven datasets in the MNDR and a death register of the patients recorded in 2021 from the MDR. The MNDR consists of the 'registry' and the 'clinical audit' datasets. The registry contains general information on all patients with diabetes who received treatment in public health clinics, while the clinical audit dataset is a subset of patients' registries that are randomly selected yearly for auditing clinical variables.[19] Eleven clinical audit datasets from 2011 to 2021 were merged, and the data

coming from similar patients were linked based on their national registration identity card numbers. Subsequently, this dataset was merged with the data from MDR to form a master cohort dataset.

Participants

The study included all T2D patients treated at 172 public health clinics in the southern region of Malaysia. This region was chosen for its high diabetes prevalence and the highest proportion of diabetes patients registered in the MNDR.[6] Moreover, the demographic characteristics of T2D patients in the southern region are comparable to the national average.[6] Patients were included if they had at least two clinical audits between 2011 and 2021, as the longitudinal study required data from at least two separate time points. Patients with other types of diabetes, such as Type 1 Diabetes, congenital diabetes, monogenic diabetes, and maturity-onset diabetes of the young, were excluded. Additionally, patients who already had the studied diabetes complications at baseline were excluded from specific analyses to ensure temporality between the predictors and target variables. For example, in analyses conducted for nephropathy, patients who already had nephropathy at baseline were excluded. Patients with missing information about the studied diabetes complications were also excluded from specific analyses to avoid introducing bias through imputed target variables. For instance, in analyses conducted for nephropathy, patients with missing information about their nephropathy status were excluded.

Predictors and outcomes variables

The study included various sociodemographic data, clinical parameters, and medical history as the predictors in the models.[18] The outcomes predicted by the model were all-cause

mortality, retinopathy, nephropathy, IHD, and CeVD. These outcomes were assessed based on clinical diagnoses recorded in the MNDR. The timing of these assessments varied, with outcomes recorded as they occurred during routine follow-ups.

Sample Size

While no universally accepted sample size calculation exists for ML algorithms[20], a common rule suggests having at least ten times as many data instances as there are data features[21]. With 50 features analyzed, the minimum required sample size was 500.

Missing Data

The dataset had a complex pattern of missing data, requiring imputation. Of the 48 variables in the dataset, 25 contained missing values, ranging from 0.02% to 22.3% (Supplementary Figure 1). A simulation compared four methods: mean/mode substitution, k-nearest neighbors (k-NN), MissForest, and multivariate imputation by chained equations (MICE). A complete subset of the dataset was used to create an artificial dataset with missing values matching the original dataset's pattern. The performance of each method was evaluated using root mean square error (RMSE). MissForest, which had the lowest RMSE, was selected to input all missing values into the full dataset (Supplementary Table 1). This method can impute continuous and categorical data, including complex interactions and non-linear relations.[22]

Analysis Methods

From the master cohort dataset, information was extracted to generate five datasets which corresponded to the five diabetic complications intended to be analyzed in this study. Each dataset was used to developed prediction model for their respective complications.

For the development of each prediction model, the respective dataset was randomly split into 80% training and 20% validation using a stratified approach based on the target variable to maintain class balance. Feature selection was conducted using the filter method for simplicity and computational efficiency. Correlations between predictors were assessed using Pearson's correlation for numerical variables and Cramer's V for categorical variables. Highly correlated predictors were removed to avoid multicollinearity issues. Mutual information with the target variable was also considered, and features with mutual information below 0.001 were removed.

The class imbalance was managed using the synthetic minority oversampling technique (SMOTE), which generated synthetic samples for the minority class, ensuring a more balanced dataset. Data normalization was applied to bring numerical variables to a common scale, preventing any single variable from disproportionately influencing the model.

Seven ML algorithms were tested: logistic regression (LR), support vector machine (SVM), k-nearest neighbors (kNN), decision tree (DT), random forest (RF), Extreme Gradient Boosting (XGB), and Light Gradient-Boosting Machine (LGBM). Stratified k-fold cross-validation (k=10) was used to maintain balanced class distribution in each fold, enhancing the robustness of model evaluation. Hyperparameter tuning through grid search was performed to identify the optimal settings for each algorithm, with the highest Receiver Operating Curve – Area Under the Curve (ROC-AUC) score guiding the selection. The 10-fold cross-validation was used within the training set to tune hyperparameters and optimize

model performance. The separate 20% validation set was reserved for final model evaluation on unseen data, ensuring an unbiased assessment of generalizability.

The best models for each complication were further evaluated for accuracy, sensitivity, and specificity across various decision thresholds, providing a more comprehensive understanding of the models' performance. A decision threshold in a binary classification model is the probability value that determines how predictions are classified. If a model's predicted probability for a given case exceeds the threshold, it is classified as a positive case; otherwise, it is classified as negative. Typically, the threshold is set at 0.5 (default threshold), but it can be adjusted based on the desired balance between false positives and false negatives. In medical applications, where the consequences of misclassification can vary, adjusting the threshold allows for optimizing sensitivity and specificity depending on clinical priorities.

This study specifically presented accuracy, sensitivity, and specificity at the optimal threshold (decision threshold where Youden's J statistic is highest) to provide a balanced view of model performance, as this threshold best distinguishes between true positives and false positives. Youden's J statistic is a widely used measure in diagnostic test evaluation to assess the effectiveness of a medical test in distinguishing between diseased and non-diseased individuals.[23] It is calculated as:

$$J=\text{Sensitivity}+\text{Specificity}-1$$

The value ranges from 0 to 1, where 0 indicates no discriminatory power (the test gives the same proportion of positive results for groups with and without the disease), and 1 indicates perfect sensitivity and specificity. A higher Youden's J value suggests a better trade-off between sensitivity and specificity, making it a useful metric for determining the optimal

decision threshold in classification models. The feature importance of these models was also assessed to ensure the interpretability of the models.

Results

Participants

The master cohort dataset comprised 90,933 T2D patients from southern Malaysia, resulting in a total of 288,308 instances available for the development of prediction models. Information from this dataset was extracted and used to form five datasets corresponding to the five selected diabetes complications being studied. The cohort predominantly featured female participants (60.89%) and was mostly Malay (66.03%), followed by Chinese (19.89%), Indian (13.59%), and other ethnicities (0.49%). Such distribution reflects the gender and ethnic distribution of patients with diabetes in Peninsular Malaysia.[24] Overall, the whole cohort was followed up for a median of eight years. At baseline, the median age of the cohort was 59 years old, with median diabetes duration of four years. Summary of the number of patients and instances extracted for each dataset is available in Supplementary Figure 2.

Regarding all-cause mortality, 17.79% (n=16,180) of the cohort developed complications, with an incidence rate of 23.6 per 1000 person-years. For retinopathy, 4.23% (n=3,537) of 83,602 patients experienced complications, with an incidence rate of 10.2 per 1000 person-years. Nephropathy affected 5.74% (n=4,642) of 80,876 patients, with an incidence rate of 14.0 per 1000 person-years. For IHD, 1.34% (n=1,131) of 84,383 patients developed complications, with an incidence rate of 3.2 per 1000 person-years. Lastly, for CeVD, 0.32% (n=282) of 88,491 patients faced complications, with an incidence rate of 0.8

per 1000 person-years. Table 1 summarizes the incidence rate of the diabetic complications found in this study.

“INSERT TABLE 1 HERE”

Prediction models

The results of the models for all-cause mortality prediction showed that the XGB and LGBM models performed the best, with ROC-AUC scores of 0.84, while the other models scored between 0.78 and 0.79. In predicting retinopathy, the XGB and LGBM models had the highest ROC-AUC scores, at 0.69 and 0.71, respectively, while the SVM model had the lowest score of 0.52. The assessment of nephropathy prediction capabilities also showed a similar pattern, where LGBM marginally led with a score of 0.71, followed closely by the XGB model at 0.70, and the SVM model remaining the least effective, with a score of 0.53. For IHD prediction, both the XGB and LGBM models had the highest performance with a score of 0.66, in contrast to the SVM's lowest score of 0.51. Finally, in CeVD prediction, the LGBM model had the best performance with a score of 0.74, followed by the XGB model at 0.71, while the DT model exhibited the least competence, scoring 0.50.

“INSERT TABLE 2 HERE”

Overall, the XGB and LGBM models consistently demonstrated superior performance across all five selected diabetic complications. In this study, the LGBM models showed the best performance in predicting the five selected diabetic complications, followed closely by the XGB model. Among the five models, four of them (excluding the IHD model) showed good performance with ROC-AUC score of at least 0.70. Table 2 summarizes the ROC-AUC score for all models.

Evaluation of the best models – the LGBM models

As the best-performing model for each complication, the performance of the LGBM models was further evaluated. The ROC curve for each LGBM model, along with their respective ROC-AUC score, optimal threshold and the best Youden's Index is available in Supplementary Figure 3. Figure 1 depicts their accuracy, sensitivity, and specificity at various decision thresholds, along with their corresponding accuracy, sensitivity, and specificity values at the optimal threshold.

Among the LGBM models, the all-cause mortality model showed the best performance with a sensitivity of 0.76, specificity of 0.74, and a Youden's index of 0.51 at the optimal threshold of 0.18. The retinopathy model had an optimal threshold of 0.04 with a sensitivity of 0.63, a specificity of 0.67, and a Youden's index of 0.32. The nephropathy model performed better than retinopathy at an optimal threshold of 0.04, with a sensitivity of 0.70, specificity of 0.62, and a Youden's index of 0.32. The IHD model had an optimal threshold of 0.01 with a sensitivity of 0.57, a specificity of 0.68, and a Youden's index of 0.25. The CeVD model had an optimal threshold of 0.002 with a sensitivity of 0.64, a specificity of 0.72, and a Youden's index of 0.37.

"INSERT FIGURE 1 HERE"

In addition to their performance, the feature importance of the LGBM models was also evaluated for a better understanding of the models. This analysis demonstrates a consistent pattern of feature importance across the LGBM models for the five diabetes complications. Several features such as ethnicity, glycosylated haemoglobin A1c (hbA1c), blood pressure, LDL-cholesterol, and diabetes duration consistently emerge as key predictors, though their importance scores vary across models. Despite this variability, the overall trend

shows these features as dominant in predicting complications. Conversely, other features, such as medications and comorbidities, consistently exhibit lower importance scores, reinforcing their lesser predictive value (Supplementary Figure 4).

Discussion

The overall performance of the models shows that the LGBM and XGB models consistently outperformed simpler models like kNN, SVM, LR, and DT, highlighting the strength of ensemble methods in capturing complex patterns in medical data.[9,25] LGBM achieved the highest ROC-AUC scores across all five complications, with four models showing good performance (ROC-AUC $\geq$ 0.7). Despite the similar performance between the LGBM and XGB models, the LGBM models were much more efficient, training seven to 13 times faster than XGB (Supplementary Table 2). Simpler models, particularly DT and SVM, showed lower performance, likely due to their limited ability to model complex, non-linear relationships and high-dimensional data interactions, with SVM models being computationally expensive and time-consuming to train.[26]

In comparison to findings from a systematic review,[9] the performance of ML models in this study demonstrated relatively lower ROC-AUC scores, particularly for microvascular outcomes like retinopathy and nephropathy. The review reported that neural networks and decision trees achieved mean ROC-AUCs of 0.87 and 0.86, respectively, while this study observed the highest ROC- AUC for LightGBM, with 0.70 for retinopathy and 0.71 for nephropathy. Similarly, for macrovascular outcomes such as ischemic heart disease, the review found that ensemble methods had a mean ROC-AUC of 0.70, which is slightly higher than the results in this study, where random forest and LightGBM reached ROC-AUCs of

0.66. These differences may be attributed to variations in model complexity, dataset size, or the features used in the analyses.

The LGBM, which is the best-performing model, was further evaluated using accuracy, sensitivity, and specificity across various decision thresholds. Due to class imbalance, the models showed high specificity but low sensitivity at the default threshold. A more balanced accuracy and sensitivity were observed at the optimal threshold. At the optimal threshold, the models demonstrated acceptable levels of accuracy, sensitivity, and specificity, indicating that they are reliable tools for prediction. In practice, the choice of decision threshold depends on the desired balance between the consequences of false positives and false negatives. All models have relatively low optimal thresholds, suggesting that such low points might be considered the starting point for deciding on the desired decision threshold to be used in real-world practice.

While most models did not achieve excellent ROC-AUC values (>0.8), they remain clinically useful for risk stratification among diabetes patients. Since all individuals with diabetes are inherently at risk of complications, the cost of false positives is relatively low. Therefore, a higher sensitivity may be preferable, which can be achieved by lowering the decision threshold. The acceptable trade-off in specificity should be determined based on the available resources for screening in each healthcare setting, ensuring feasibility in real-world practice.

In practice, training ML models require relatively high computational resources. However, this demand is primarily limited to the training phase. Once trained, these models are computationally efficient for inference (making predictions), ensuring practical feasibility for real-world implementation. Additionally, newer algorithms like LGBM are specifically designed to optimize computational efficiency, significantly reducing resource demands

while maintaining strong predictive performance. As shown in Supplementary Table 2, the LGBM model required significantly less train time than most other algorithms, highlighting its efficiency.

In this study, feature importance scores reveal consistent trends across diabetic complications, highlighting key predictors such as ethnicity, hbA1c, blood pressure, LDL-cholesterol, and duration of diabetes as crucial for predicting complications, aligning with existing evidence.[27] Conversely, predictors such as prescribed medications rank lower, suggesting that they have a less direct impact on outcomes. This may be due to adherence issues and registry limitations, as prescription data do not reflect actual medication use or capture dosage adjustments over time. Such a consistent pattern across the five models supports shared pathways in diabetes complications and aligns with current scientific understanding, enhancing model validity.[27] Understanding these feature importance scores helps to address the black box issue in ML by improving model interpretability, thereby increasing confidence among healthcare professionals in using these models for patient care.

One of the key strengths of this study was the use of real-world, big data from the MNDR, which represents a comprehensive dataset of patients in Malaysia. The routine nature of MNDR data collection ensures that the methodology of this study can be replicated in the future, facilitating direct comparisons and practical applications of the developed prediction models. Furthermore, this study represents one of the first efforts in Malaysia to develop ML models for predicting diabetic complications by utilizing advanced algorithms such as XGB and LGBM. These models are based on data from primary healthcare settings, ensuring their relevance to the local context while offering greater accuracy and robustness than traditional methods.

Several limitations should be considered when interpreting these findings. One major limitation of this study stems from the constraints of using secondary data, which only includes variables already collected in the registry. This limits the ability to consider other important factors, such as family history, physical inactivity, and health literacy, which could be relevant to diabetes complications. Additionally, the study did not employ deep learning methods, which may restrict the potential of the model to capture more complex associations. While deep learning can offer higher accuracy, concerns over interpretability and the high computational resources required led to its exclusion from this study.

Another potential limitation is that the dataset is geographically limited to southern Malaysia. While it includes a diverse population with varying socioeconomic backgrounds, urban and rural distributions, and major ethnic groups, the model's generalizability to other regions remains uncertain and would require external validation. Additionally, although efforts were made to capture real-world clinical diversity, potential biases in model predictions cannot be fully ruled out, highlighting the need for ongoing evaluation to ensure fairness in different healthcare settings.

Findings from this study provide evidence to support the integration of ML models, particularly LGBM, into clinical practice as supportive tools to complement clinical judgement, enhance risk stratification and provide personalized care for patients with diabetes. To ensure robustness and applicability, further optimization, validation, and fine-tuning of these models across diverse populations is necessary. Future studies may consider comparative studies with deep learning models to identify the most effective predictive tools, as these models may capture complex patterns that simpler models might overlook; however, balancing accuracy and interpretability should always be prioritized. In addition, using a more comprehensive national dataset would improve model generalization and accuracy, making it applicable across different regions and populations.

Conclusion

In predicting diabetic complications, the LGBM models demonstrated superior performance compared to other ML algorithms, including LR, kNN, SVM, DT, and RF. The LGBM models achieved good performance in all diabetic complications with an ROC-AUC of at least 0.7, except for IHD. The performance of the XGB models was comparable to that of the LGBM models but required a much longer training time. The feature importance analysis highlights that features such as age at first diagnosis of T2D, T2D duration, ethnicity, BP, and HbA1c are crucial predictors of various diabetic complications, aligning with the existing medical literature. This emphasizes the potential of LGBM models in medical predictive analytics, particularly in diabetes management, while also highlighting the importance of feature selection and the need for careful consideration of model interpretability and computational efficiency in healthcare applications.

Ethical Approval: Approved Institutional Review Board Name: Medical Research and Ethics Committee, Ministry of Health Malaysia IRB reference Number: NMRR ID- 22-00928-MMB (IIR)

References

1. International Diabetes Federation. *IDF Diabetes Atlas*. 10th Edition ed. 2021. Accessed 14/2/2023. https://www.diabetesatlas.org

2. Institute for Public Health. *National Health and Morbidity Survey (NHMS) 2019: Vol. I: NCDs – Non-Communicable Diseases: Risk Factors and other Health Problems*. 2020. http://www.iku.gov.my/nhms-2019

3. Viigimaa M, Sachinidis A, Toumpourleka M, Koutsampasopoulos K, Alliksoo S, Titma T. Macrovascular Complications of Type 2 Diabetes Mellitus. *Curr Vasc Pharmacol*. 2020;18(2):110-116. doi:10.2174/1570161117666190405165151

4. Zimmerman RS. Diabetes mellitus: management of microvascular and macrovascular complications. *Cleveland Clinic: Centers for Continuing Education*. 2016. https://www.clevelandclinicmeded.com/medicalpubs/diseasemanagement/endocrinology/diabetes-mellitus/

5. Emerging Risk Factors Collaboration. Diabetes mellitus, fasting glucose, and risk of cause-specific death. *N Engl J MedNew.* 2011;364(9):829-841.

6. Ministry of Health Malaysia. *National Diabetes Registry Report 2013-2019*. 2020. www.moh.gov.my

7. Zulman DM, Vijan S, Omenn GS, Hayward RA. The relative merits of population-based and targeted prevention strategies. *Milbank Q*. Dec 2008;86(4):557-80. doi:10.1111/j.1468-0009.2008.00534.x

8. Grant SW, Collins GS, Nashef SAM. Statistical Primer: developing and validating a risk prediction model†. *Eur J Cardiothorac Surg*. 2018;54(2):203-208. doi:10.1093/ejcts/ezy180

9. Tan KR, Seng JJB, Kwan YH, et al. Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review. *J Diabetes Sci Technol*. 2023;17(2):474-489.

10. Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol*. Mar 2018;12(2):295-302. doi:10.1177/1932296817706375

11. Dworzynski P, Aasbrenn M, Rostgaard K, et al. Nationwide prediction of type 2 diabetes comorbidities. *Sci Rep*. Feb 4 2020;10(1):1776. doi:10.1038/s41598-020-58601-7

12. Ljubic B, Hai AA, Stanojevic M, et al. Predicting complications of diabetes mellitus using advanced machine learning algorithms. *J Am Med Inform Assoc*. Jul 1 2020;27(9):1343-1351. doi:10.1093/jamia/ocaa120

13. Segar MW, Vaduganathan M, Patel KV, et al. Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *Diabetes Care*. Dec 2019;42(12):2298-2306. doi:10.2337/dc19-0587

14. Song X, Waitman LR, Yu AS, Robbins DC, Hu Y, Liu M. Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients Using a Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study. *JMIR Med Inform*. Jan 31 2020;8(1):e15510. doi:10.2196/15510

15. Cichosz SL, Johansen MD, Hejlesen O. Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications. *J Diabetes Sci Technol*. Oct 14 2015;10(1):27-34. doi:10.1177/1932296815611680

16. Khairudin Z, Razak NAA, Abd Rahman HA, Kamaruddin N, Abd Aziz NA. Prediction of Diabetic Retinopathy Among Type II Diabetic Patients Using Data Mining Techniques. *Malays J Comput*. 2020;5(2):572-586.

17. Sim R, Chong CW, Loganadan NK, Adam NL, Hussein Z, Lee SWH. 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*. 2022;16(3):549-559. doi:10.1093/ckj/sfac252

18. Abas MZ, Li K, Hairi NN, Choo WY, Wan KS. Machine learning based predictive model of Type 2 diabetes complications using Malaysian National Diabetes Registry: A study protocol. *J Public Health Res*. 2024;13(1):22799036241231786. doi:10.1177/22799036241231786

19. Feisul MI ASE. *National Diabetes Registry Report, Volume 1, 2009-2012.* Vol. 1. 2013. http://www.moh.gov.my

20. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. 2020;368:l6927. doi:10.1136/bmj.l6927

21. Falconer N, Abdel-Hafez A, Scott IA, Marxen S, Canaris S, Barras M. Systematic review of machine learning models for personalised dosing of heparin. *Br J Clin Pharmacol*. 2021;87(11):4124-4139. doi:https://doi.org/10.1111/bcp.14852

22. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118. doi:10.1093/bioinformatics/btr597

23. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its Associated Cutoff Point. *Biom J*. 2005;47(4):458-472. doi:https://doi.org/10.1002/bimj.200410135

24. Ministry of Health Malaysia. *National Diabetes Registry Report 2023*. 2023. www.moh.gov.my

25. Kee OT, Harun H, Mustafa N, et al. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. *Cardiovasc Diabetol*. 2023;22(1):1-10.

26. Géron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."; 2022.

27. Elhefnawy ME, Ghadzi SMS, Noor Harun S. Predictors Associated with Type 2 Diabetes Mellitus Complications over Time: A Literature Review. *J Vasc Dis*. 2022;1(1):13-23.

Table 1: The incidence rate of diabetic complications

	No of patients	**Complications developed**		**Diabetes duration[a] (years)**	**Incidence rate[b]**
		Yes	**No**		
All-cause mortality	90933	16180 (17.79%)	74753 (82.21%)	13 (10, 17)	23.6
Retinopathy	83602	3537 (4.23%)	80065 (95.77%)	10 (7, 13)	10.2
Nephropathy	80876	4642 (5.74%)	76234 (94.26%)	10 (7, 13)	14.0
IHD	84383	1131 (1.34%)	83252 (98.66%)	10 (7, 13)	3.2
CeVD	88491	282 (0.32%)	88209 (99.68%)	10 (7, 14)	0.8

[a] Duration of diabetes when patients exit the cohort

[b] Incidence rate per 1000 patient-years

Complications presented in n (%); T2D duration presented in median (Q1, Q3)

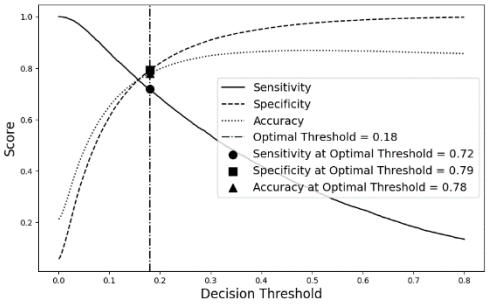
Table 2: Models' performance for each diabetes complication evaluated on their respective hold-out datasets.

	k-NN	SVM	LR	DT	RF	XGB	LGBM
All-cause mortality	0.78	0.65	0.78	0.78	0.79	0.84	0.84
Retinopathy	0.64	0.52	0.62	0.58	0.61	0.69	0.71
Nephropathy	0.65	0.53	0.64	0.59	0.65	0.70	0.71
IHD	0.63	0.51	0.63	0.56	0.6	0.66	0.66
CeVD	0.67	0.61	0.66	0.50	0.65	0.71	0.74

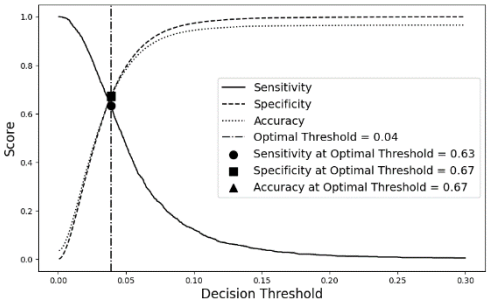
The performance was measured using ROC-AUC score

kNN: k-Nearest Neighbors; SVM: Support Vector Machine; LR: Logistic Regression; DT: Decision Tree; RF: Random Forest; XGB: Extreme Gradient Boosting; LGBM: Light Gradient Boosting Machine.

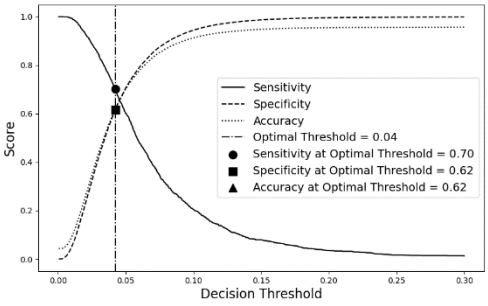
All-cause mortality



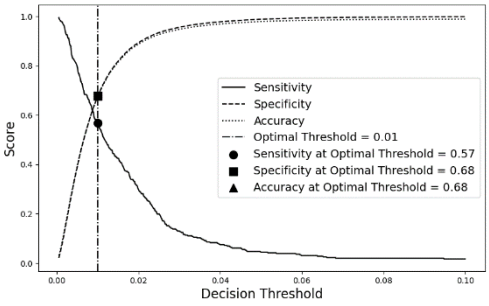
Retinopathy



Nephropathy



Ischaemic heart disease



Cerebrovascular disease

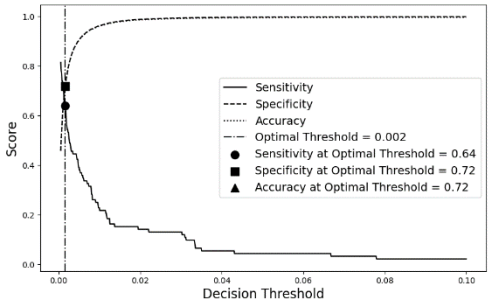


Figure 1: Accuracy, specificity, and sensitivity of the LGBM models for each diabetes complication at various decision thresholds with their corresponding value at the optimal threshold.