Journal of Public Health Research

Journal of Public Health Research
2024, Vol. 13(1), 1–8

DOI: 10.1177/22799036241231786
journals.sagepub.com/home/phj
Sage

Machine learning based predictive model of Type 2 diabetes complications using Malaysian National Diabetes Registry: A study protocol

Mohamad Zulfikrie Abas[1], Ken Li[2], Noran Naqiah Hairi[1], Wan Yuen Choo[1] and Kim Sui Wan[3]

Abstract

Background: The prevalence of diabetes in Malaysia is increasing, and identifying patients with higher risk of complications is crucial for effective management. The use of machine learning (ML) to develop prediction models has been shown to outperform non-ML models. This study aims to develop predictive models for Type 2 Diabetes (T2D) complications in Malaysia using ML techniques.

Design and methods: This 10-year retrospective cohort study uses clinical audit datasets from Malaysian National Diabetes Registry from 2011 to 2021. T2D patients who received treatment in public health clinics in the southern region of Malaysia with at least two data points in 10 years are included. Patients with diabetes complications at baseline are excluded to ensure temporality between predictors and the target variable. Appropriate methods are used to address issues related to data cleaning, missing data imputation, data splitting, feature selection, and class imbalance. The study uses 7 ML algorithms, including logistic regression, support vector machine, *k*-nearest neighbours, decision tree, random forest, extreme gradient boosting, and light gradient boosting machine, to develop predictive models for four target variables: nephropathy, retinopathy, ischaemic heart disease, and stroke. Hyperparameter tuning is performed for each algorithm. The model training is performed using a stratified *k*-fold cross-validation technique. The best model for each algorithm is evaluated on a hold-out dataset using multiple metrics.

Expected impact of the study on public health: The prediction model may be a valuable tool for diabetes management and secondary prevention by enabling earlier interventions and optimal resource allocation, leading to better health outcomes.

Keywords

Type 2 diabetes, machine learning, predictive models, diabetes complications, diabetes registry

Date received: 20 April 2023; accepted: 24 January 2024

Significance for public health

The development of prediction models based on data from the Malaysian National Diabetes Registry for Type 2 Diabetes complications is significant in addressing the rising prevalence of diabetes in Malaysia. Diabetes affects millions of individuals and is associated with severe complications. With the development of prediction models, healthcare providers may identify patients with a higher risk of developing complications earlier, allowing for earlier interventions and better health outcomes. This model

[1]University of Malaya, Kuala Lumpur, Malaysia
[2]University College London, London, UK
[3]Institute of Public Health, Ministry of Health Malaysia, Selangor, Malaysia

Corresponding authors:
Mohamad Zulfikrie Abas, University of Malaya, Kuala Lumpur 50603, Malaysia.
Email: m_zulfikrie@yahoo.com

Wan Yuen Choo, Social and Preventive Medicine Department, Faculty of Medicine, University of Malaya, Kuala Lumpur 50603, Malaysia.
Email: ccwy@ummc.edu.my

may also assist in the optimal allocation of resources by identifying patients who require intensive monitoring and treatment, thereby enabling policymakers to allocate resources more efficiently. Machine learning algorithms enable the analysis of large and intricate datasets, thereby providing an opportunity to understand the population better. The insights gained from this study could also have broader implications for other low- and middle-income countries with similar demographics and healthcare systems, thus providing a valuable tool for diabetes management and secondary prevention.

Introduction

Diabetes is a global health concern, with a projected increase in prevalence in the coming years.[1] Malaysia has a higher prevalence of diabetes than the global average, attributed to factors such as an ageing population, urbanisation, sedentary behaviour, and unhealthy eating habits.[2] Type 2 diabetes (T2D) accounts for over 90% of all diabetes cases.[1] The increasing prevalence of T2D may lead to more individuals living with diabetes complications, which presents a challenge for the healthcare system to offer optimal care to all patients.[3] As resources are always limited, risk stratification of T2D patients is essential. Accurate risk stratification can be achieved with a good prediction model based on local data.[4] Risk prediction models have significant potential to guide decisions and allocate resources effectively, leading to improved patient outcomes.

A prediction model is a tool that uses multiple predictors to predict specific outcomes and can be developed using regression models, machine learning (ML), and pattern recognition.[4] A total of 87 ML models were evaluated in a systematic review.[5] The results suggest that ML models generally outperformed non-ML models in predicting diabetes complications in T2D patients.[5] ML algorithms are well-suited for handling complex and large datasets compared to traditional statistical methods. As ML utilises big data to make predictions, the availability of large datasets due to the digitalisation of various records has made them increasingly popular. When integrated into healthcare, it may revolutionise healthcare delivery to become more precise and targeted.[6] The Malaysian National Diabetes Registry (MNDR) stores data from millions of diabetes patients nationwide, which can be leveraged to better understand the disease within the local context.[3]

Many studies have been conducted in various countries to develop prediction models for disease management using ML methods, and these methods are expected to become mainstream.[6] However, there is a lack of studies on developing prediction models for diabetes complications in Malaysia, with previous studies using only a small sample size, limiting generalizability.[7,8] Therefore, this study aims to use ML techniques to develop predictive models for T2D complications using data acquired from the MNDR. By leveraging the ML technique and the large dataset, this study may overcome the limitations of non-ML models in capturing the complexity of the data, and small sample sizes encountered in previous studies, especially in Malaysia. This protocol aims to provide transparency to model building and ensure replicability of the model.

Research objective and hypothesis

The T2D complications of interest in this study are nephropathy, retinopathy, ischaemic heart disease and stroke. This study aims to develop a prediction model for each T2D complication mentioned previously using data acquired from the MNDR. We would like to evaluate the performance of such models, which can subsequently help us decide whether these models are sufficiently good to be implemented in real practice.

The null hypothesis is that there is no significant difference in the performance of the various machine learning models developed in this study for predicting each T2D complication. To address the null hypothesis, we develop separate prediction models for each T2D complication using different machine learning algorithms and compare the performance of these models using multiple evaluation metrics, including accuracy, precision, recall, F1 score, and receiver operating characteristic area under the curve (ROC-AUC).

Materials and methods

Study design

This is a 10-year retrospective open cohort study from 2011 to 2021. The study population is all patients with T2D who received diabetes treatment in any of the 172 public health clinics in the southern region of Malaysia. This region was chosen because it has the highest prevalence of diabetes in the country.[2] Only patients with at least two data points in the 10 years are included. Patients who already had diabetes complications intended for the analysis at baseline are excluded to ensure temporality between the predictors and target variables. For instance, when developing a prediction model for nephropathy, patients who had nephropathy at baseline are excluded from the analysis.

Data source

This study uses secondary data extracted from the MNDR. The MNDR consists of the 'registry' and the 'clinical audit' datasets.[3] The registry contains 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. Clinical audit datasets from 2011 to 2021 are obtained from the Ministry of Health and linked at the patient level using unique national registration identification card numbers. In this way, a patient who was audited at least twice in the 11 years will have at least two data points, thus forming a longitudinal dataset. This method of data linkage has been commonly applied in several previous studies.[9,10]

Study variables

The outcome variable in this study is the four diabetes-related complications: nephropathy, retinopathy, ischaemic heart disease, and stroke. Complications were recorded in the registry based on the clinical diagnoses by physicians. These variables are presented in categorical form.

The predictor variables include patients' sociodemographic information, clinical parameters, and medical history. Sociodemographic information includes patients' age, gender, and ethnicity. Clinical parameters include information on body mass index, waist circumference, blood pressure, fasting blood sugar, glycosylated haemoglobin A1c (HbA1c), cholesterol level, serum creatinine, and proteinuria. The patient's medical history includes smoking status, comorbidities, and prescribed medication information. There is also information on follow-up duration, which represents the time between predictor recording and outcome measurement. The predictors consist of numerical and categorical data.

Model-building strategy

This study adapts model-building strategies from previous studies to ensure reliable and robust results in developing the predictive model.[5,11] This modelling process aims to classify patients into 'T2D complication present' or 'T2D complication absent' groups for each complication mentioned above.

Generating input data. The patient-level linked dataset contains patients' data at two or more timepoints. When developing models for each diabetes complication, we remove ineligible patients from the dataset based on specific exclusion criteria for each complication. This will result in four datasets corresponding to the four diabetes complications intended to be studied. For patients who did not develop diabetes complications throughout the follow-up period, the year at the final timepoint available is used as the time for outcome measurement. For patients who developed diabetes complications within the follow-up period, the year of diagnosis is used as the time of outcome measure. For patients with more than two timepoints within the study period, each timepoint before the time for the outcome measure is considered an independent data sample and treated as an independent instance. Subsequent steps are performed using this newly generated dataset. Figure 1 summarises the process of generating a dataset for the input data using IHD as the target variable. A similar approach will be taken to develop the prediction model for the other three complications by replacing the target variable with the intended complications using their respective dataset.

Data cleaning. Data cleaning involves removing redundant or irrelevant records, followed by identifying and treating implausible values as missing. Missing values are examined, and the entire instances or features may be removed based on their importance. Categorical variables are encoded as ML algorithms require numerical data.

Missing data imputation. Handling missing values is critical for training classifiers because many ML algorithms cannot process incomplete data. To address the issue, four different methods are thoroughly tested.[11,12] One approach involves mean and mod substitution, which involves fillings in missing values in numerical features with their mean, and categorical features with their mod.[11] However, this method can lead to biased outcomes that may not accurately reflect the actual situation.

Another method is *k*-nearest neighbours (*k*-NN) imputation. This method uses a predefined distance metric to find the nearest neighbours of a missing value.[11,12] For each missing feature, the values from the *k* nearest neighbours with available feature values will be used to impute the missing feature. The values of the neighbouring features will be uniformly averaged. If a sample has multiple missing features, the neighbours of each feature can be different.

The third approach is MissForest, which iteratively utilises random forests to impute missing values.[11,12] The algorithm first selects the feature with the least missing values (candidate column). Missing values in the candidate column will be filled with the mean of that column. The candidate column will then be fitted to a random forest model and treated as the output, with the other columns in the dataset treated as inputs to the model. The random forest model will be trained, and the missing values of the candidate column will be imputed using the prediction from the fitted random forest. These steps will be repeated for all other columns in the dataset.

The final method being tested is multivariate imputation by chained equation (MICE). The algorithm replaces the missing values in each feature with the mean of that feature.[12] In each iteration, the algorithm predicts the missing values of each feature using the observed values of the other features and regresses the feature on all the other features in the dataset. This process is repeated for each feature with missing values until convergence, with each feature being updated in turn.

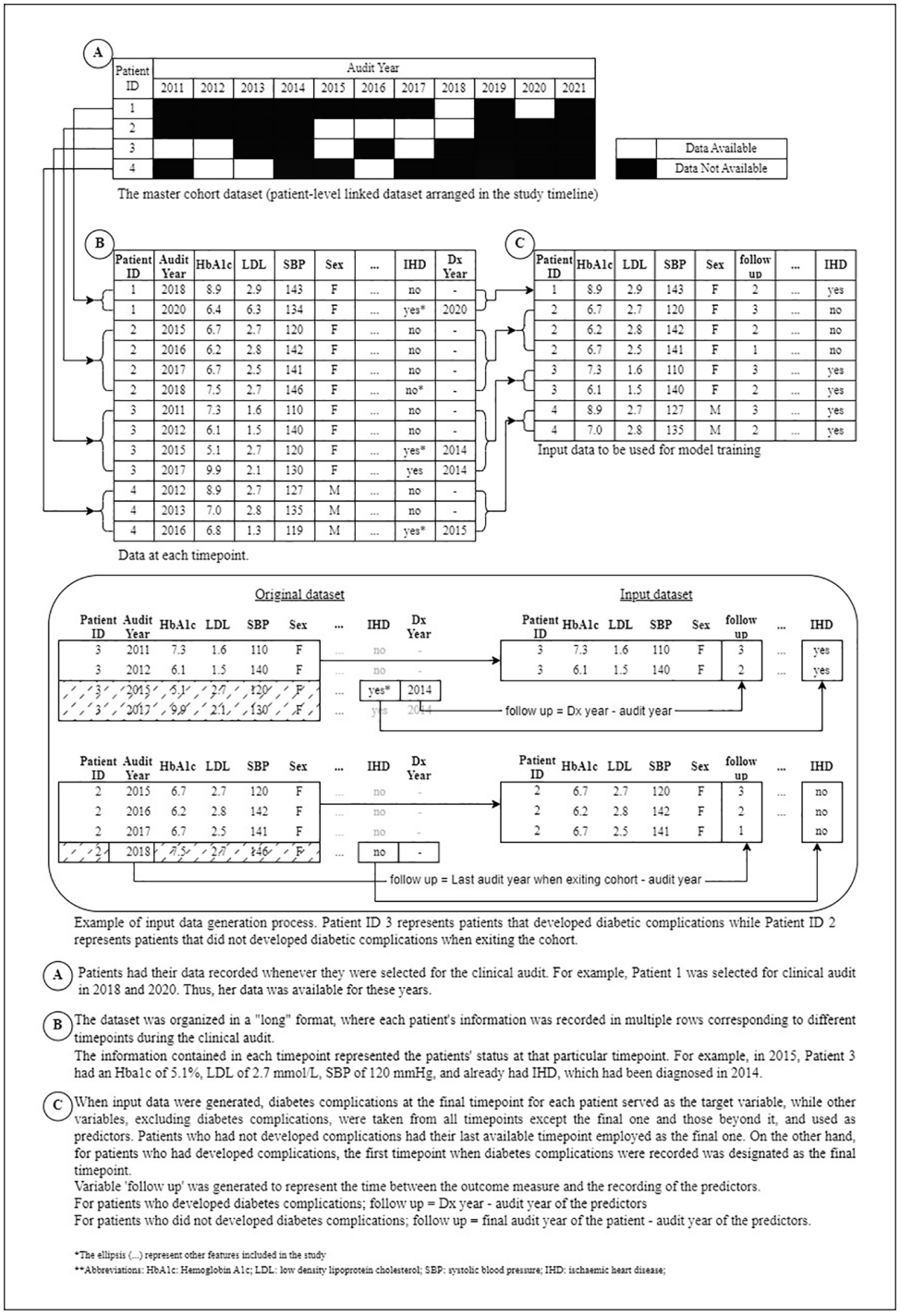

Figure 1. Illustration of the dataset generation for input data using IHD as the target variable.

To compare the four methods' performances, we calculate each method's root mean square error (RMSE). First, we simulate the missing value problem by selecting subset of the dataset with no missing values (Subset 1). We then calculate the percentage of missing values in the original dataset and randomly drop values from each column in Subset 1 based on the percentage of missing to create an artificial dataset with missing values (Subset 2). The distribution of the variables in Subset 2 is compared to the original dataset to ensure the results will not be biased. Finally, we use the four methods to impute the missing values in Subset 2 and evaluate their performance using the RMSE by comparing the imputed Subset 2 and complete datasets (Subset 1). The algorithm with the lowest RMSE is used to impute missing data in the original dataset.

Dataset splitting. After imputing the missing values, the dataset is split into two groups, stratified by the outcome variable. A larger set comprising 80% of the instances is used to train and test the model, whereas the remaining 20% is used as a hold-out set to validate the model's performance.

Feature selection. Next, feature selection for our model is performed. Feature selection is the process of selecting a subset of relevant features for use in machine learning model construction. Feature selection can be performed in three main ways: filter, wrapper, and embedded methods.[13] This study uses the filter method for feature selection. Filtre methods involve selecting features based on statistical measures, such as correlation with the target variable or variance, without using a machine learning algorithm. Pearson's correlation is used for numerical variables, and Cramer's V is used for categorical variables to determine the correlation between each feature. Highly correlated features may indicate redundancy and collinearity. Analysis of variance (ANOVA) and chi-square tests are used to assess the relationship between the features and target variable. ANOVA is used for numerical features, whereas the chi-square test is used for categorical features. Features without significant association (p-value > 0.25) with the target variable are removed. Feature selection offers several benefits, including faster training, reduced model complexity, improved accuracy, and reduced overfitting.

Managing class imbalance. Imbalanced classification is a problem for predictive modelling as most machine learning algorithms assume an equal number of examples for each class, leading to poor predictive performance for the minority class.[13] This is a challenge because the minority class is often more important, and the problem is more sensitive to classification errors for the minority class than the majority class. Severe class imbalance is expected in our dataset as the prevalence of diabetes complications among T2D patients ranges from less than 2% to around 20%, depending on the type of complications.[3]

Under-sampling and over-sampling are two common approaches for handling imbalanced classifications.[13] The synthetic minority oversampling technique (SMOTE) is a popular over-sampling method that generates synthetic samples for the minority class, resulting in a more diverse and representative dataset with a low risk of overfitting. This can improve the generalizability and accuracy of the ML models. We use this method to address the class imbalance problem in our datasets.

Data normalization. The final step before training the model is to perform data normalisation.[14] Data normalisation is a technique used to bring the numerical variables in a dataset to a common scale without distorting the differences in their ranges. The goal is to assign equal weights to each variable so that no variable dominates the model's performance based solely on its size.

Machine learning algorithms. In this study, 7 ML algorithms are used to predict diabetes-related complications. The algorithms were chosen based on previous studies' predictive models, problem types, and the performance of improved models.[5] The problem type in this case is classification, and improved model performance refers to algorithms such as Extreme Gradient-Boost (XGBoost) and Light Gradient-Bosting Machine (LightGBM), which are improved versions of decision tree.[14,15]

The first algorithm is logistic regression, which is based on calculating the probability of the target class by using a linear equation with intercept and slope coefficients for the features.[14] Despite the assumption of linearity between the dependent and independent variables not always being correct, the simplicity and effectiveness of logistic regression make it a good algorithm to test in this study.

The second algorithm is the Support Vector Machine (SVM), which is a supervised learning algorithm that can be used for both regression and classification problems.[14] SVM tries to find a line (hyperplane) that separates the data points into different categories. The line is chosen such that the distance between the line and the closest point from each category is maximised, making the classification more accurate.

The k-nearest neighbours (k-NN) algorithm is a classification method that estimates the likelihood of a data point belonging to a particular group based on the group of its k-nearest neighbours.[14] The algorithm calculates the distance between the data points and uses the k-nearest neighbours to determine the class of the new data point. The value of k can be selected based on the data, and it determines the number of neighbours to consider when making a classification decision.

The decision tree (DT) is a classification method that uses nodes and branches to make decisions from the dataset.[14] DT is useful for understanding the relationships between variables and making clear and easy-to-follow predictions. However, this method may be unstable and prone to overfitting. The criteria for selecting the attributes to split the data in each node are based on entropy and information gain, which are measurements of disorder and uncertainty, respectively.

Ensemble algorithms combine the results of multiple base models to improve the overall predictive performance of a single model.[14] This study uses the Random Forest (RF) ensemble algorithm, which creates multiple decision trees based on different random subsets of features and then aggregates the predictions of each tree to make the final prediction.

Boosting is an ensemble algorithm that improves the performance of multiple weak algorithms by adjusting the weights of the observations from previous classifications.

In this study, two boosting algorithms, Extreme Gradient Boosting (XGBoost) and its variation Light Gradient Boosting Machine (LightGBM), are used.[14,15]

Model training and hyperparameter tuning. Each model is trained using stratified k-fold cross-validation (SCV) with only the training dataset. The SCV method divides the training dataset into k equal folds, where $k-1$ folds are used for model training, and the remaining fold is used to validate the model's performance.[14] SCV provides an advantage over simple k-fold cross-validation by ensuring that the distribution of classes is preserved in each fold, thus preventing bias. In this study, we selected a value of $k=10$ because the k-value of 10 has been proven to provide a low bias and a modest variance while still being at an acceptable computational cost.[13]

Hyperparameter tuning is important because it may improve an algorithm's performance by finding the optimal hyperparameter values.[14] Hyperparameters are parameters that are set by the user before applying a machine-learning algorithm to a dataset to control the learning process. Examples of hyperparameters include the learning rate in gradient descent optimisation, the number of trees in RF, and the regularisation parameter in logistic regression. By tuning these hyperparameters, we can improve the model's performance and potentially avoid overfitting or underfitting. The hyperparameter is optimised using either 'grid search' or 'random search', depending on the available computational resources. As the algorithms determine the 'best performance' based on a single metric, the ROC- AUC score is used to identify the optimal hyperparameter configuration for each ML algorithm. This is because the ROC- AUC score considers the trade-off between true positive and false positive rates and summarises the model's overall performance, making it useful for comparing different hyperparameter settings.

Model performance evaluation. After the best model for each ML algorithm has been trained using their respective optimum hyperparameter setting, the performance of these models is tested on the hold-out dataset. For each target variable, we compare the performance of each ML model by assessing their accuracy, precision, recall, specificity, and F1 score at different probability threshold levels. By evaluating these metrics, we can determine the best model developed for predicting diabetes complications. We will also present relevant confusion matrix to better illustrate the performance of the models.

Feature elimination. To achieve a parsimonious model, we assess each model's feature importance using the recursive feature elimination technique (RFE).[13] RFE ranks the features based on their importance to the model. Then, we eliminate the least important feature and restart the model training with the new set of features to see if there is any improvement in the model performance. This process will be stopped if the model performance is not improved. Figure 2 summarises the model-building strategy. The entire process is repeated for each diabetes complication studied.

Sensitivity analysis. It should be noted that using SMOTE to address the class imbalance and imputation of missing values potentially introduces bias.[13] Thus, we perform a sensitivity analysis of the models by retraining them without using SMOTE. For models that can handle missing data, such as XGBoost, we also train the model without imputing the missing value.

Discussion

This study uses a registry-based dataset, which contains large real-world data, to train the ML model to predict diabetes complications in Malaysia. Real-world data offers better examples for ML models to learn from, making them more accurate and adaptable.[16] A large dataset provides more diverse examples and can capture rare events, leading to improved accuracy and reliability of the model.[4] In addition, this study uses clinical audit datasets, where the samples were randomly selected from all T2D patients receiving treatment in public health clinics based on a predefined protocol.[3] Compared to other previous studies done in Malaysia, which used samples randomly selected from a clinic,[7] or following a cohort of patients from two tertiary hospitals,[8] our study utilises a method that can better represent T2D patients in Malaysian population.

However, a registry-based study is restricted to variables already captured in the registry, which may limit the ability to include additional important factors.[9] This study also focuses on traditional ML models over deep learning algorithms for their simplicity, interpretability, less computational resources and robustness to noise and outliers in the data.[17] Unfortunately, this may limit the ability to fully harness the potential of machine learning acquired through deep learning algorithms in capturing complex nonlinear relationships for predicting diabetes complications. While the study excludes patients with complications at baseline to establish a temporal relationship between predictors and T2D complications, it is crucial to acknowledge that observational studies cannot establish causality, as predictive models only identify associations and predict outcomes without providing direct evidence of causation.

Conclusion

This protocol presents a practical methodology for developing a tailored predictive model for T2D complications in Malaysia. The use of machine learning techniques on a large real-world clinical dataset enhances the potential for a robust and accurate predictive model. This documented

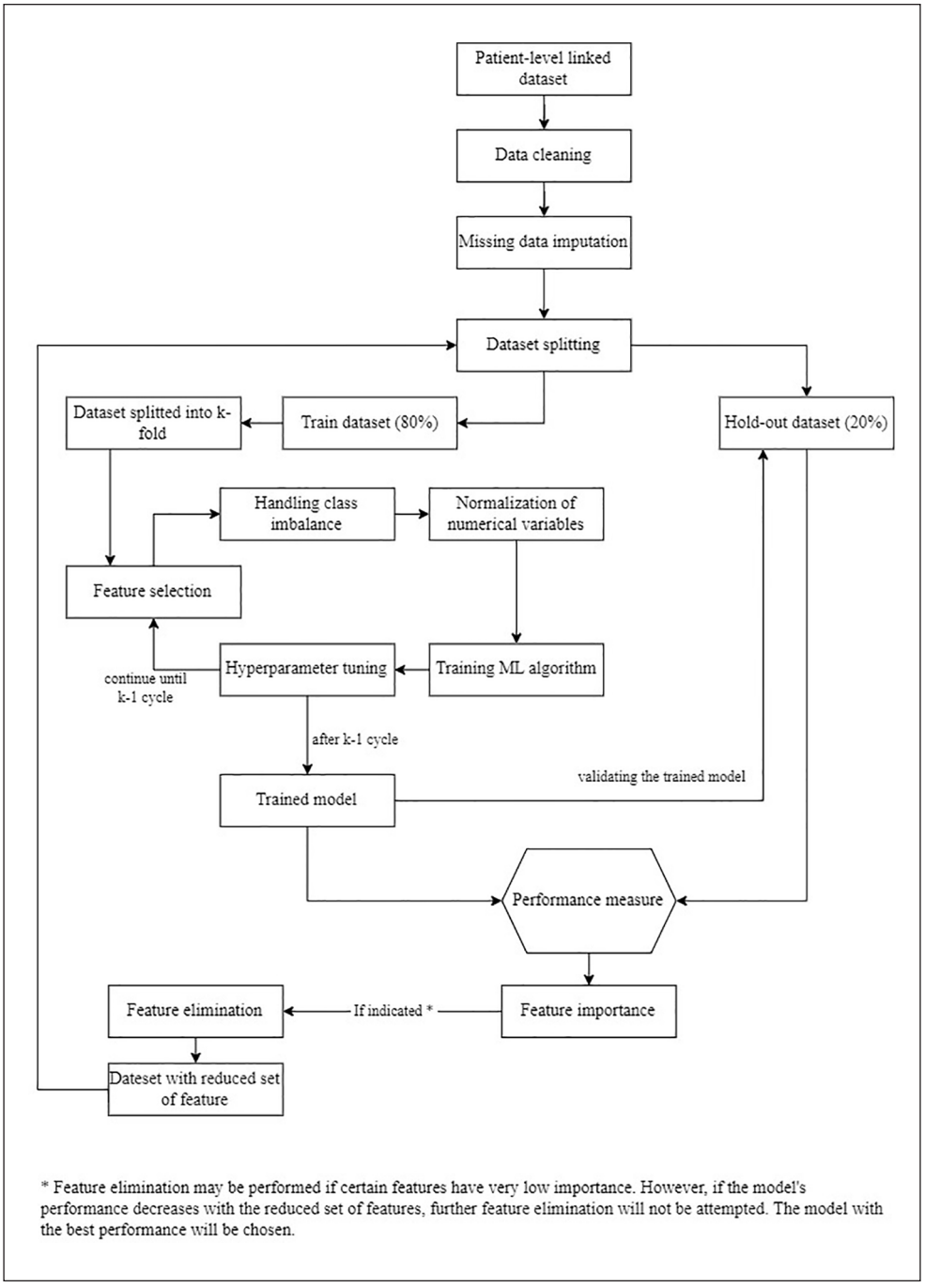

Figure 2. Summary of model-building strategy.

protocol provides guidance for future researchers to replicate or build similar models. The resulting predictive model has the potential to assist healthcare professionals in making informed decisions for effective diabetes management, thereby leading to improved patient outcomes. As this is the first few studies in a Malaysian setting that utilise big data to develop prediction models for T2D complications, the performance of the developed ML models can serve as a benchmark for further models that will be developed in the future.

Author contributions

Abas MZ initiated the study, developed the protocol, prepared and wrote the manuscript. Other authors provided input on the development of the protocol and reviewed the manuscript. All authors have read and approved the final protocol.

Declaration of conflicting interests

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.

Ethics approval

The protocol for this study has been approved by the Medical Research and Ethics Committee, Ministry of Health Malaysia (NMRR ID- 22-00928-MMB (IIR))

ORCID iD


Mohamad Zulfikrie Abas https://orcid.org/0000-0001-9170-2099

References

1. International Diabetes Federation. *IDF diabetes atlas*. 10th ed. Brussels: International Diabetes Federation, 2021.
2. Institute for Public Health. *National Health and Morbidity Survey (NHMS). 2019: Vol. I: NCDs – non-communicable diseases: risk factors and other health problems*. Selangor: Ministry of Health Malaysia, 2020.
3. Ministry of Health Malaysia. *National diabetes registry report 2013–2019*. Putrajaya: Ministry of Health Malaysia, 2020.
4. Grant SW, Collins GS and Nashef SAM. Statistical Primer: developing and validating a risk prediction model.. *Eur J Cardio-Thorac Surg* 2018; 54: 203–208.
5. 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: 474–489.
6. Cichosz SL, Johansen MD and Hejlesen O. Toward Big Data analytics: review of predictive models in management of diabetes and its complications. *J Diabetes Sci Technol* 2015; 10: 27–34.
7. Khairudin Z, Abdul Razak NA, Abd Rahman HA, et al. Prediction of diabetic retinopathy among type II diabetic patients using data mining techniques. *Malays J Comput* 2020; 5: 572–586.
8. Sim R, Chong CW, Loganadan NK, et al. Comparison of a chronic kidney disease predictive model for Type 2 diabetes mellitus in Malaysia using Cox regression versus machine learning approach. *Clin Kidney J* 2023; 16: 549–559.
9. Wan KS, Hairi NN, Mustapha F, et al. Five-year LDL-cholesterol trend and its predictors among Type 2 diabetes patients in an upper-middle-income country: a retrospective open cohort study. *PeerJ* 2022; 10: e13816.
10. Kim SW, Hairi NN, Mustapha FI, et al. Poorer attainment of hemoglobin A1C, blood pressure and LDL-cholesterol goals among younger adults with Type 2 diabetes. *Sains Malays* 2021; 50: 3631–3645.
11. Jian Y, Pasquier M, Sagahyroon A, et al. A machine learning approach to predicting diabetes complications. *Healthcare* 2021; 9: 1712.
12. Van Buuren S. *Flexible imputation of missing data*. New York: CRC Press, 2018.
13. Kuhn M and Johnson K. *Applied predictive modeling*. New York: Springer, 2013.
14. Géron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. California: O'Reilly Media, Inc, 2022.
15. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017; 3149 – 3157.
16. Liu F and Panagiotakos D. Correction: real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol* 2023; 23: 109.
17. Janiesch C, Zschech P and Heinrich K. Machine learning and deep learning. *Electronic Mark* 2021; 31: 685–695.