Development and validation of a type 2 diabetes machine learning classification model for clinical decision support framework

Victor Glanz[1,*], Vladimir Dudenkov [1,+], Alexey Velikorodny[1,+]

[1]BioME, Paris, 75001, France
[*] v.glanz@biome-science.net
[+]these authors contributed equally to this work

ABSTRACT

Undiagnosed type 2 diabetes is very common and represents a significant challenge for all national healthcare systems. Although diagnostic criteria and laboratory screening procedures are well-established, clinical tests have limitations, and in many cases diagnosis confirmation and more precise interpretation of the tests results are required. Machine learning methods, when applied to clinical outcomes risk prediction, demonstrate great effectiveness as they recognize specific patterns in data dynamics and thus can be used for identification of at-risk cases where diabetes and complications can be delayed or even prevented. We developed and validated a machine learning classification model for type 2 diabetes that possesses several important advantages over conventional methods (FINDRISC, ADA risk score), including mean values of 0.959, 0.92 and 0.89 for AUC, specificity and sensitivity, respectively. The study results potentially have major clinical implication, as the model is able to exchange data with electronic health records. Thus, it can be employed in clinical decision support framework together with other diabetes, cardiovascular disease models and models for related conditions.

Introduction

Diabetes remains one of the most rapidly developing health emergencies around the world despite continuous scientific progress uncovering molecular basis of its pathogenesis. Globally, more than 537 million people are affected by diabetes, and the situation is expected to get even worse, as 46% more cases are predicted to emerge by 2045, according to International Diabetes Federation [1]. Being a complex systemic disorder, diabetes can be divided into four general categories, as proposed by American Diabetes Association (ADA): 1) type 1 diabetes (T1D), absolute insulin deficiency resulting from autoimmune destruction of β-cells; 2) type 2 diabetes (T2D), loss of insulin secretion by β-cells and tissue sensitivity to insulin resulting from inflammation, metabolic stress and genetic predisposition to certain abnormalities of metabolism; 3) gestational diabetes (GA), pregnancy-specific type, not evident before gestation; 4) other specific types [2]. Hyperglycemia is a clinical manifestation of various underlying mechanisms of diabetes development. Diagnostic procedures consist of laboratory measurements of plasma glucose (fasting or oral glucose tolerance test) and glycated hemoglobin. Current diagnostic cut-off values, established by ADA, are above 126 mg/dL (7.0 mmol/L) for fasting glucose and ≥ 6.5% for glycated hemoglobin.

Missed undiagnosed cases of diabetes and inability to efficiently detect high-risk individuals early are the main issues in clinic, that scientific research is yet to resolve. To address this problem, it seems reasonable to use machine learning and deep learning methods that have created opportunities to improve medical data processing, which is confirmed by numerous studies [3, 4]. Such methods as support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) were successfully used for predicting the outcomes of hypertension [5]. Neural networks and deep learning techniques can be applied for electronic health records (EHR) analysis [6]. The joint application of deep learning models and bootstrap aggregation was able to improve the quality of information extraction systems from cancer pathology reports [7]. In managing diabetes, machine learning (ML) methods can facilitate both diagnosis and prognosis. In comprehensive reviews by Kodama et al. and Silva et al. some of the most clinically relevant ML diabetes prediction models were evaluated in terms of performance metrics [8, 9]. Prognostic only ML diabetes models had mean area under the curve (AUC) of 0.88, while pooled prognostic and diagnostic low-bias-algorithms showed 0.830 c-index. In contrast, several traditional diabetes risk estimation methods (FINDRISC, AUSDRISC, etc.) showed mean AUC of 0.79 [8].

Type 2 diabetes is the focus of our work due to several reasons. First, T2D affects significantly more individuals than any other type, representing 90-95% of all cases. Worth noting that T2D is no longer considered as 'disease of the adults', nor does T1D occur specifically at young age. Second, the rate of undiagnosed T2D cases is alarmingly high due to increase of plasma glucose levels being gradual, often too slow for a person to notice any symptoms. Finally, although T2D screening laboratory tests and risk prediction algorithms are established and available in clinic, proper time interval between screenings is not defined. Current ADA guidelines recommend a minimum 3-year interval, but it is admitted that the timing can and should be personalized [2]. Personalized screening intervals based on diabetes risk can be critical for timely intervention aimed at delaying or preventing the onset of T2D. We believe, our work addresses these issues by combining real-world patient-generated data and sophisticated machine learning algorithms in a prediction model that serves as a medical decision support and health monitoring tool (a digital health framework, which contains a number of interconnected risk models, is currently being developed). In this study, we present development and validation of a Type 2 diabetes machine learning classification algorithm, based on electronic health records data.

Results

Data description

The source of data for the study were EHR containing ultrasound markers, laboratory measurement results and information about clinical outcomes (ICD-10 codes determined by medical specialists). We used records from 04.09.2018 to 01.01.2022 extracted from a private anonymized dataset that included 157615 patients. After primary filtration, the following patients were excluded: T1D and other non-T2D types of diabetes, pregnant women, and patients under 18 years of age. The remaining data underwent outlier detection and feature selection procedures. Only information regarding patients with significant (relevant) features was obtained. Then we selected measurement results gathered 12 months before the diagnosis (for each patient in 'diabetes' cohort) and measurement results gathered 12 months before the examination by an endocrinologist (for non-diabetics). Table rows that did not contain information about fasting plasma glucose level were excluded. The described data filtration process is shown in Figure 1. Thus, the final dataset for model derivation was a table containing 40039 records corresponding to 30550 patients (1108 diabetics and 29442 non-diabetics), with each patient having one or several records in a 5-year timeline. Dataset characteristics corresponding to 49 features are presented in Table 1. Individuals with diabetes were characterized by greater median age (61 years vs. 41 years for 'healthy' individuals), plasma triglycerides level (1.81 mmol/L vs. 1.19 mmol/L), fasting plasma glucose level (7.41 mmol/L vs. 5.04 mmol/L). T2D was also associated with hypertension (median systolic blood pressure was 142.25 mm Hg). Moreover, there was a difference in several sonography-derived measurements of liver and pancreas between two cohorts.

Model development and evaluation

The dataset described above was used for the evaluation of several models. A detailed description of the models is given in Methods (see Machine learning algorithms). Each model was trained and evaluated using training and test subsets, respectively. To obtain more reliable results, we assigned three train/test splits and calculated performance metrics in each of them. The results are shown in Table 2. Application of the DES classifier was declined because it did not provide any metric improvements compared to the base bagging model. ROC curves of the considered classifiers for the first train/test split are shown in Figure 2. Figure 2 demonstrates that the stacked generalization model shows the best performance results. The stacked generalization model was further improved at the development phase and validated (see below).

Interpretation

A post hoc importance analysis has been performed by applying the SHAP method. The beeswarm plot and the bar chart of average absolute SHAP values are shown in Figure 3 A and B. To facilitate interpretation, absolute SHAP values were normalized to add up to 100 in order to simplify interpretation procedures. Based on these importance values additional rules were set to employ the model in medical practice. It was proposed to make five most important features mandatory and to exclude eleven least important ones. This decision makes practical use of the model flexible in case of insufficient number of non-mandatory predictors (none of them are used in extreme cases), which might be often seen in reality, for example, during initial visits to a clinic, i.e., when EHR system would contain a limited number of patient-specific measurements.

External validation

Data from National Health and Nutrition Examination Survey (NHANES) was used for external validation of the diabetes model (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020). 2017-March 2020 data considered herein was representative of data accumulation period for the present study. Only features present in training dataset were selected from NHANES. Similarly to the filtration process described above, pregnant women and patients under 18 years old were excluded. The presence of fasting plasma glucose level measurements was mandatory for inclusion in the dataset. The external validation dataset characteristics are shown in Table 3. Achieved performance metrics for the external dataset were as follows: 0.899 AUC, 17.896 EER, 0.627 F1 score, 0.764 macro F1 score, 0.809 sensitivity, 0.850 specificity, 0.512 precision, 0.958 NPV.

Discussion

Machine learning can be applied to a disease prediction in two different ways: disease (condition) identification and future occurrence prediction (prognosis). We developed and validated a model capable of accurate classification of patient-generated data and identification of diabetes-specific patterns in clinical biomarkers values with 0.92 specificity and 0.89 sensitivity (AUC 0.959, precision 0.246, NPV 0.996; see Table 2). Although several diabetes risk scores exist, most of them use virtually the same non-biochemical predictors, such as chronological age, medical history, body mass index. These risk scores demonstrate overall moderate diagnostic accuracy, but are undeniably easy to use in clinical setting, as there is no need to perform most of the laboratory tests [10]. An example of such diabetes risk models is ADA diabetes risk score, one of the most well-known clinically validated models. Reported performance metrics for ADA diabetes risk score applied to a NHANES dataset were as follows: 0.79 sensitivity, 0.67 specificity, 0.10 positive predictive value, 0.99 negative predictive value, 0.83 AUC. Comparison of NHANES dataset-derived performance metrics of ADA score and our model proved the latter to be superior in most aspects (see Results: External validation). Another widely used diabetes prediction model is The Finnish Diabetes Risk Score (FINDRISC). When applied to a European (Spanish) population, it showed the following diagnostic metrics for detecting new cases of diabetes based on a fasting plasma glucose criterion: 0.68 AUC, 0.64 sensitivity, 0.64 specificity, 0.076 positive predictive value, 0.976 negative predictive value [11].

The ML algorithm proposed in this work can be viewed as a more promising method of detecting undiagnosed diabetes. Due to low positive predictive value of ADA diabetes risk score and FINDRISC, these models are more suitable for populations with higher ratio of undiagnosed diabetes, rather than general population. Trained on general population-derived real-world data and having relatively high positive predictive value, our model is free of this limitation.

Fasting plasma glucose level serves in our model as one of the outcome predictors and as an outcome-defining feature (see Methods). Some studies avoid doing this because of possible bias [12]. However, multiple evidence suggests that glucose and glycated hemoglobin (HbA1c), being two main diagnostic criteria for diabetes, are not as explicit in diagnostic value and need to be interpreted correctly in the context of patient's medical history, previous glycemic status, complaints, etc. In clinical setting, patients are usually screened for diabetes with a single fasting plasma glucose test, yet other conditions, such as acute pancreatitis, may be the cause of elevated glucose levels [13]. The phenomenon of stress hyperglycemia, also known as transient hyperglycemia, can be observed in patients without pre-existing diabetes. The association with stress hyperglycemia is common for ischemic stroke, myocardial infarction and other critical conditions. In some cases, diabetes screening procedure itself may be the cause of stress-induced hyperglycemia [14]. Therefore, hyperglycemia at admission does not necessarily correspond to diabetes diagnosis. The baseline data utilized in this study supports this point view, as it is shown in Table 4.

The odds ratio (OR) calculated from this table equals 74.05 with a 95% confidence interval [64.73, 84.72]. Thus, we confirm the existence of a significant association between high glucose levels and T2D. However, the first row of the contingency table shows that accurate diagnostics of diabetes based solely on fasting glucose levels is not possible.

Similarly, diagnosing diabetes by HbA1c alone may result in many cases remaining undiagnosed, as glycated hemoglobin has low sensitivity as a diagnostic biomarker. The sensitivity value of glycated hemoglobin-based diabetes diagnosis was reported to be only 41% [15]. On the other hand, combination of glucose and glycated hemoglobin yields much better results, diagnosing approximately 85% of type 2 diabetes cases [15]. Use of two different diagnostic tests to detect diabetes, however, is not obligatory and is recommended by ADA when diagnosis confirmation is needed.

The ML model, a primary result of this study, is able to exchange and store data in EHRs, which greatly facilitates its potential practical use in clinical setting. Current ADA guidelines summarize crucial elements of an effective diabetes care framework as the Chronic Care Model (CCM), with decision support being one of core components of the CCM (see [2] for

detailed description of the CCM). The model presented in this work can serve as a part of diabetes care framework - a clinical decision support tool in form of an EHR-based ML-powered screening method for type 2 diabetes in populations of ambulatory patients (symptomatic or non-symptomatic) without confirmed diabetes.

Machine learning as a method can be considered as a more appropriate technique for diagnosing and classification of diabetes, as some individuals present with features characteristic of both most prevalent diabetes types at the time of diagnosis, which complicates classification and prolongs time period of choosing appropriate treatment strategy. In view of insufficient understanding of molecular mechanisms of T2D development machine learning algorithms became very promising solutions, because unknown patterns in data dynamics are recognized and processed. However, ML-based predictions inevitably face the challenge of dealing with imbalance in data used for model training. Class imbalance in real-world clinical diabetes data is demonstrated by the incidence rate of diabetes being much lower than that of diabetes-free outcomes., In this study we did not use the accuracy metric for model performance evaluation because it can be misleading in the imbalanced dataset. A comprehensive set of performance metrics was employed instead.

The study has some limitations. First, as misdiagnosis is common when it comes to classifying patients as having either type 1 or type 2 diabetes, and due to absence of genetic data available in electronic health records, there is a minor chance of type 1 diabetes cases present in our training dataset, despite exclusion of ICD codes for T1D. Second, only continuous variables were used as diabetes risk features. Alcohol and tobacco consumption, sleep duration and other potentially relevant information, which require a questionnaire, was not included in model development due to current data unavailability. Future versions of our model will consider these risk factors. It should be noted, however, that contribution of socioeconomic and demographic characteristics of populations to type 2 diabetes risk is yet to be defined [16].

In conclusion, we developed and validated a machine learning classification model for type 2 diabetes that demonstrates several important advantages over established methods and possesses a major clinical use potential as a part of clinical decision support framework.

Methods

Model development workflow

EHR data went through sequential stages of collection, feature selection, and outlier detection. The resulting dataset was split into test and training subsets. We used a stratified split of 70% and 30% for training and testing datasets, respectively. With the stratified split strategy, the ratio of diabetics to non-diabetic individuals in splits remains the same as in full population. Since the splitting was random, it was repeated 3 times to ensure the reliability of the results. Because of missing values in the data, it was necessary to apply data imputation methods [17]. Both multivariate and univariate approaches to missing-values imputation were considered in this work. Data imputation step was combined with the hyperparameter optimization procedure. A specific imputation method was selected individually for each machine learning model. A standard procedure of removing mean values and normalization to unit variance was then applied.

To achieve the best result, we combined several machine learning approaches. The following algorithms were selected:

- XGBoost
- Multilayer perceptron
- Ridge classifier
- Ridge classifier with polynomial features
- Bootstrap aggregating
- Dynamic ensemble selection

In addition, a stacked generalization was used as a combination of several machine learning models with logistic regression as a meta-estimator. Final modeling results were evaluated and interpreted both by biomedical experts and statistical methods. The resulted workflow is shown in the following diagram (Figure 4). The research workflow was designed in

accordance with the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement [18].

Outlier detection

Prior experience with clinical biomarker data had shown that automatic methods do not provide sufficient quality for outlier detection. Thus, cleaning of the data from outliers was carried out using the intervals of acceptable values specified by laboratory manuals and results of literature analysis.

Feature selection

Feature selection is the process of dimensionality reduction of the feature space that helps to improve computation time and to better understand the data [19, 20]. In addition, a possible reduction of features number is necessary for real medical application of the model. The feature importance is the central concept of feature selection and refers to a score assigned for all the input features. A higher score means that the feature will have a larger effect on the ML model that is being used to predict a target variable. State-of-the-art methods such as mutual information, Gini index, and ANOVA were used in the study [21-24]. Extremely randomized trees were employed as an intrinsic tree-based method [25]. It is an ensemble learning technique aggregating the outcomes of many decision trees to generate the final prediction. In case of tree-based method, the importance of a feature is computed as the normalized total reduction of the criterion brought by that feature (impurity-based importance). Because of the stochastic nature of the algorithm, feature importance values might vary from calculation to calculation. Therefore, the final value was calculated as an average of 10 runs. To generalize the results of all methods, a total importance of features was computed. The selected methods were applied to the feature space, resulting in M×Nmatrix, where M is the number of features, N is the number of methods. The values in each row were normalized and added up. The final values represented generalized feature importance.

Data imputation and hyperparameter optimization

Optuna framework was utilized for hyperparameter optimization [26]. Tree-structured Parzen Estimator (TPE) was chosen as a sampling method [27]. TPE is a single-objective Bayesian algorithm that uses tree-structured adaptive Parzen estimators as a surrogate [28]. For each model data imputation step was combined with the hyperparameter tuning process (assuming that each model can have its optimal data imputation algorithm). In addition, some models, such as XGBoost, do not require such procedures at all. Therefore, data imputation type and imputation method parameters were considered as additional hyperparameters. Data imputation types are listed below. Both univariate and multivariate imputation methods were employed.

- K-nearest neighbor (KNN). A missing sample is imputed by finding the closest samples in the training set and averages these nearby samples to fill in the value. Configuration of KNN imputation involves selecting the distance measure and the number of neighbors for each prediction.
- Multiple Imputation by Chained Equations (MICE). MICE imputes missing values through an iterative series of predictive models [29]. In each iteration, each specified feature in the dataset is imputed using other features. Iterations must be carried out until convergence is achieved.
- Univariate imputation using the feature statistics. This method assumes that the missing values are replaced with some statistical measure of the entire feature column (median, mode or mean value). The choice of a particular statistic is a configurable parameter of this method.

Machine learning algorithms

XGBoost

Extreme Gradient Boosting (XGBoost) is a state-of-the-art algorithm that is designed to be parallelizable, computationally efficient, and robust to overfitting. It is based on the idea of gradient boosting [30]. Gradient boosting combines the predictions of weak estimators by dint of additive training strategies. Classification and regression trees (CART) are used as weak estimators in XGBoost. The output of each CART is combined using the following additive function:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

where KK is the number of trees, ff_{kk} is a function in the functional space of all possible CARTs. XGBoost learning is an iterative process. At every iteration new CART tries to rectify the weakness in the prediction of previous one. $yy\hat{}_{u}$ for iteration t is given as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

where $yy\hat{}_{u}^{(tt)}$ is the prediction for iteration tt, $ff_{tt}(xx_{ii})$ is an additive function for iteration tt. XGBoost introduces the idea of regularization to reduce the model complexity. The objective with regularization is:

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + r(f_t)$$

where ll is the loss function, rr is the regularization term. The regularization term is defined as:

$$r(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega^2$$

where $\gamma\gamma$ and $\lambda\lambda$ are parameters for controlling the degree of regularization, $\gamma\gamma$ is the number of leaves, $\omega\omega$ is the score on each leaf. Regularization and tree-specific parameters, the number of boosting iterations, and sampling restrictions were selected by hyperparameters tuning procedure.

Multilayer perceptron

A multilayer perceptron (MLP) is a fully connected artificial neural network [31]. MLP consists of layers of neurons: one input layer, one or more hidden layers and one output layer. Each neuron receives inputs from other neurons from the previous layer and provides outputs to the neurons in the following layer. The presence of at least one hidden layer allows MLP to approximate any continuous function. The connection strength between neurons is controlled by weights. The output of a given neuron is the linear weighted sum of the inputs from the previous layer, which is passed through an activation function:

$$y_j = \phi\left(\sum_{i=1}^{n}\left(w_{ji}x_i + w_{j0}\right)\right)$$

where yy_{jj} is an output of j-th neuron, xx_{ii} is an input from the i-th neuron of previous layer, ww_{jjii} is a weight, $\phi\phi$ is an activation function.

Weights are sought in the training process by optimizing a predefined loss function. We used binary cross entropy loss function:

$$Loss = \frac{-1}{N}\sum_{i=1}^{N}\left(y_i log\hat{y}_i + (1 - y_i)log(1 - \hat{y}_i)\right)$$

where $yy\hat{}_{u}$ is the i-th value in the model output, yy_{ii} is the corresponding target value, N is the number of values in the model output, N=1 in our case.

Dropout can be used as a method of the neural network regularization [32]. It improves the generalization performance by repeating the procedure of temporarily removing neurons. The choice of neurons to dropout is random. The strength of dropout can be regulated by dropout rate i.e., by the probability that the neuron will be removed.

The dropout rate, activation function, number of hidden layers, and number of neurons in layers were the main hyperparameters for tuning present MLP model.

Ridge classifier

Ridge regression improves the classical linear regression model by imposing L2 regularization element [33]. The ridge coefficients minimize a penalized residual sum of squares:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j\, x_{ij}\right)^2 + \alpha \sum_{j=0}^{p} w_j^2$$

where $\alpha\alpha$ is a regularization strength. Larger values specify stronger regularization. We identified regularization strength as one of the most important for the hyperparameter optimization stage. The ridge classifier, based on the ridge regression method, can be applied to the binary classification problem. It converts the target variable ("diabetic" and "non-diabetic" in our case) into [-1, 1] and applies the regression method. The highest value in prediction is accepted as a target class.

Ridge classifier with polynomial features

The polynomial extension of the ridge classifier was considered as a separate model. We expanded the feature space by polynomial combinations of some base features with degrees less than or equal to the specified degree. Thereby, the specified degree and set of features to expand acted as parameters to hyperparameter optimization. Described polynomial extension of the ridge classifier may be considered as a type of feature engineering. This extension allows to identify nonlinear patterns [34].

Bootstrap aggregating

Bootstrap aggregation (bagging) is an ensemble learning technique that uses bootstrap sampling [35]. Bagging fits base estimators on different bootstrap samples of the training dataset and combines them using the average of the predictions or another statistic. Bagging is a classical machine learning model and is used frequently for improving the accuracy of arbitrary base estimators [36, 37]. The type of base estimators and their number were significant parameters of the algorithm, hence their inclusion in the hyperparameter optimization stage. Bootstrap parameters, such as samples limit and features limit for splitting, were also employed.

Dynamic ensemble selection

Dynamic ensemble selection strategies refer to ensemble learning techniques that dynamically choose a subset of models to make a prediction based on the specific input [38]. The subset is defined according to the model competence, estimated over some local region of the target instance. This local region is called Region of Competence (RoC). Clustering and distancebased methods were used [39, 40] to define RoC. The competence of the base models is estimated using the samples in the RoC according to predefined criteria. Several DES algorithms were used in this study:

- K-Nearest Oracles Eliminate (KNORA-E)
- K-Nearest Oracles Union (KNORA-U)
- Dynamic Ensemble Selection-Kullback-Leibler divergence (DES-KL)
- Randomized Reference Classifier (RRC)
- Meta-learning for dynamic ensemble selection (META-DES)

A description of the listed algorithms is given in [41-45]. The algorithm type was considered as a hyperparameter for optimization, META-DES was eventually selected. An optimized and trained bagging ensemble was used as a pool of DES base models.

Stacked generalization

Stacked generalization is a technique that combines multiple models to reduce their biases [46]. Various models (level-0 models) make classification or regression predictions, and their outputs form a new set of data [47]. The new set of data is treated as input to another learning model, (level-1 model), being trained through cross-validation to resolve it. Previously described models were utilized as level-0 models, optimized and trained individually. Three classical models were considered as level-1 candidates: naive Bayes classifier, logistic regression, and decision tree classifier. Finally, logistic regression was chosen by the hyperparameter optimization procedure.

Evaluation metrics

Accuracy metric was declined because it can be misleading in the imbalanced dataset. A comprehensive set of performance metrics derived from ROC analysis was employed instead:

- Area under the receiver operating characteristic curve (AUC)
- Equal error rate (EER)

- F1 score
- Macro F1 score
- True positive rate (sensitivity)
- True negative rate (specificity)
- Positive predictive value (precision)
- Negative predictive value (NPV)

The cut-off point for ROC curves was found by the maximization of the Youden's index:

$$J = \max_{c}(\text{sensitivity}(c)+\text{specificity}(c) - 1)$$

where JJ is the Youden's index and cc is the cut-off point.

Results interpretation

Post-processing the results, was achieved by application of Shapley additive explanations (SHAP) method. SHAP explains predictions of the arbitrary model using the game theoretic approach. SHAP computes Shapley values, and features in a dataset are considered as players in a coalition. Shapley value explanation is represented as an additive feature attribution method [48]:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

where ll is the explanation model, zz' is the coalition vector, MM is the maximum coalition size and $\phi\phi_{jj}$ is the Shapley value for a feature j. Using Shapley values, we can obtain an a posteriori feature importance, which is evaluated as an average of the absolute Shapley values per feature [49]:

$$I_j = \frac{1}{n}\sum_{i=1}^{n}\left|\phi_j^{(i)}\right|$$

Ethical considerations

Ethical review and approval are not relevant for this study, because it uses existing data. Informed Consent Statement: not applicable.

Data availability statement

Abbreviated version of the training dataset and processed full external validation dataset are available at https://github.com/biome-science/srep-t2d-data-availability. Web-demo of the diabetes model can be accessed at https://biome-science.com/demo/model/?id=type2diabet. Full data used in model development is available on reasonable request from the corresponding author.

References

1. International Diabetes Federation. *IDF Diabetes Atlas*. (International Diabetes Federation, 2021).
2. American Diabetes Association. Professional Practice Committee: Standards of Medical Care in Diabetes—2022. *Diabetes Care* **45** xx–xxxx (2022).
3. Rahmani, A. M. *et al.* Machine learning (Ml) in medicine: Review, applications, and challenges. *Mathematics* **9**, (2021).
4. Egger, J. *et al.* Medical deep learning—A systematic meta-review. *Comput. Methods Programs Biomed.* **221**, 106874 (2022).
5. Chang, W. *et al.* A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics* **9**, 178 (2019).
6. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Heal. Informatics* **22**, 1589–1604 (2018).
7. Yoon, H. J. *et al.* Accelerated training of bootstrap aggregation-based deep information extraction systems from cancer pathology reports. *J. Biomed. Inform.* **110**, 103564 (2020).

8. Kodama, S. *et al.* Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. *J. Diabetes Investig.* **13**, 900–908 (2022).
9. Silva, K. De *et al.* Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *Int. J. Med. Inform.* **143**, 104268 (2020).
10. Brown, N., Critchley, J., Bogowicz, P., Mayige, M. & Unwin, N. Risk scores based on self-reported or available clinical data to detect undiagnosed Type 2 Diabetes: A systematic review. *Diabetes Res. Clin. Pract.* **98**, 369–385 (2012).
11. Salinero-Fort, M. A. *et al.* Performance of the Finnish diabetes risk score and a simplified Finnish diabetes risk score in a community-based, cross-sectional programme for screening of undiagnosed type 2 diabetes mellitus and dysglycaemia in madrid, Spain: The SPREDIA-2 study. *PLoS One* **11**, 1–17 (2016).
12. Moon, S., Jang, J. Y., Kim, Y. & Oh, C. M. Development and validation of a new diabetes index for the risk classification of present and new-onset diabetes: multicohort study. *Sci. Rep.* **11**, 15748 (2021).
13. Yang, X. *et al.* Stress Hyperglycemia Is Independently Associated with Persistent Organ Failure in Acute Pancreatitis. *Dig. Dis. Sci.* **67**, 1879–1889 (2022).
14. Peer, N., Balakrishna, Y. & Durao, S. Screening for type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews* vol. 2020 (2020).
15. Bergman, M. *et al.* Review of methods for detecting glycemic disorders. *Diabetes Res. Clin. Pract.* **165**, 108233 (2020).
16. Skyler, J. S. *et al.* Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes* **66**, 241–255 (2017).
17. Saar-Tsechansky, M. & Provost, F. Handling missing values when applying classification models. *J. Mach. Learn. Res.* **8**, 1625–1657 (2007).
18. Moons, K. G. M. *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **162**, W1–W73 (2015).
19. Ladha, L. & Deepa, T. Feature Selection Methods And Algorithms. *Int. J. Comput. Sci. Eng.* **3**, 1787–1797 (2011).
20. Kira, K. & Rendell, L. A. Feature selection problem: traditional methods and a new algorithm. in *Proceedings Tenth National Conference on Artificial Intelligence* vol. 2 129–134 (1992).
21. Ross, B. C. Mutual information between discrete and continuous data sets. *PLoS One* **9**, (2014).
22. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* **69**, 16 (2004).
23. Kim, H.-Y. Analysis of variance (ANOVA) comparing means of more than two groups. *Restor. Dent. Endod.* **39**, 74 (2014).
24. Shang, W. *et al.* A novel feature selection algorithm for text categorization. *Expert Syst. Appl.* **33**, 1–5 (2007).
25. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
26. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2623–2631 (Association for Computing Machinery, 2019). doi:10.1145/3292500.3330701.
27. Bergstra, J., Bardenet, R., Bengio, Y. & Kegl, B. Algorithms for Hyper-Parameter Optimization. *Adv. Neural Inf. Process. Syst.* 2546–2554 (2011).
28. Ozaki, Y., Tanigaki, Y., Watanabe, S. & Onishi, M. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. in *GECCO 2020 - Proceedings of the 2020 Genetic and Evolutionary Computation Conference* 533–541 (Association for Computing Machinery, 2020). doi:10.1145/3377930.3389817.
29. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
30. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
31. Haykin, S. *Neural Networks: A Comprehensive Foundation*. (Prentice Hall, 1998).
32. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
33. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer Series in Statistics).
34. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. (Chapman & Hall/CRC Data Science Series).
35. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
36. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
37. Hall, P., Park, B. U. & Samworth, R. J. Choice of neighbor order in nearest-neighbor classification. *Ann. Stat.* **36**, 2135–2152 (2008).

38. Britto, A. S., Sabourin, R. & Oliveira, L. E. S. Dynamic selection of classifiers - A comprehensive review. *Pattern Recognit.* **47**, 3665–3680 (2014).
39. Cruz, R. M. O., Oliveira, D. V. R., Cavalcanti, G. D. C. & Sabourin, R. FIRE-DES++: Enhanced online pruning of base classifiers for dynamic ensemble selection. *Pattern Recognit.* **85**, 149–160 (2019).
40. Souza, M. A., Cavalcanti, G. D. C., Cruz, R. M. O. & Sabourin, R. On evaluating the online local pool generation method for imbalance learning. in *Proceedings of the International Joint Conference on Neural Networks* vols. 2019-July 1– 8 (2019).
41. Ko, A. H. R., Sabourin, R. & Britto, A. S. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognit.* **41**, 1718–1731 (2008).
42. Woloszynski, T., Kurzynski, M., Podsiadlo, P. & Stachowiak, G. W. A measure of competence based on random classification for dynamic ensemble selection. *Inf. Fusion* **13**, 207–213 (2012).
43. Woloszynski, T. & Kurzynski, M. A probabilistic model of classifier competence for dynamic ensemble selection. *Pattern Recognit.* **44**, 2656–2668 (2011).
44. Cruz, R. M. O., Sabourin, R., Cavalcanti, G. D. C. & Ing Ren, T. META-DES: A dynamic ensemble selection framework using meta-learning. *Pattern Recognit.* **48**, 1925–1935 (2015).
45. Cruz, R. M. O., Sabourin, R. & Cavalcanti, G. D. C. META-DES.H: A Dynamic Ensemble Selection technique using meta-learning and a dynamic weighting approach. in *Proceedings of the International Joint Conference on Neural Networks* vols. 2015-Septe 8 (2015).
46. Wolpert, D. H. Stacked generalization. *Neural Networks* **5**, 241–259 (1992).
47. Breiman, L. Stacked regressions. *Mach. Learn.* **24**, 49–64 (1996).
48. Shapley, L. S. A VALUE FOR n-PERSON GAMES. in *Classics in Game Theory* vol. 2 69–79 (2020).
49. Molnar, C. Interpretable Machine Learning. https://christophm. github.io/interpretable-ml-book/ (2019)

Author contributions statement

A.V. and V.G. conceived the experiment, V.D. and V.G. conducted the experiment, all authors analyzed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

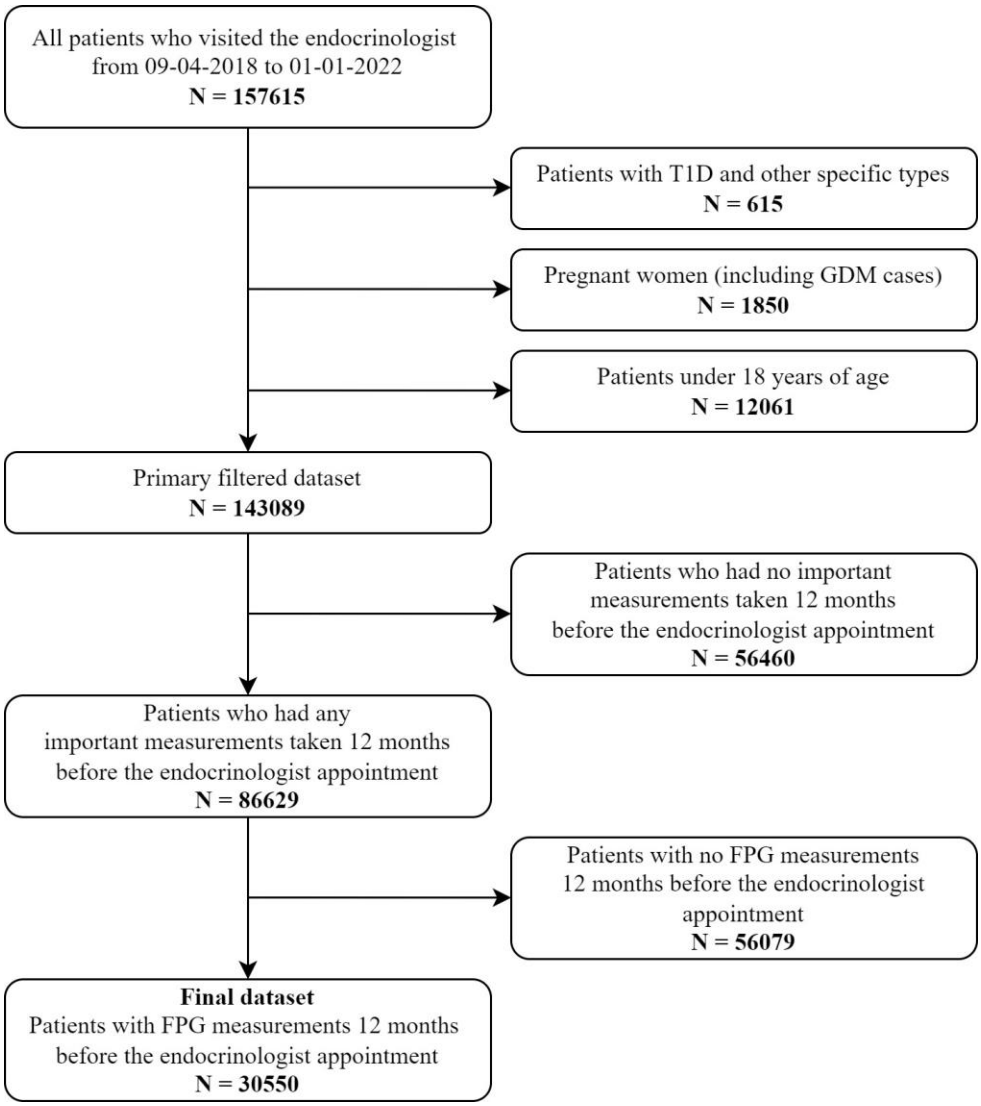

Figure 1. Data selection chart for the main dataset.

		Missing	Overall	Diabetic	Healthy	P-Value	SMD
n			40039	1230	38809		
Age, years, median [Q1, Q3]		0	42.00 [31.00,57.00]	61.00 [53.00,69.00]	41.00 [31.00,57.00]	<0.001	-1.156
Glucose (measured with blood glucose meter), mmol/L, median [Q1, Q3]		36904	5.80 [5.10,6.80]	8.00 [6.80,10.80]	5.70 [5.00,6.30]	<0.001	-1.090
Glucose (fasting), mmol/L, median [Q1, Q3]		0	5.06 [4.68,5.51]	7.41 [6.21,9.59]	5.04 [4.67,5.46]	<0.001	-1.350
Sex, median [Q1, Q3]	0	0	11566 (28.89)	407 (33.09)	11159 (28.75)	0.001	0.094
	1		28473 (71.11)	823 (66.91)	27650 (71.25)		
SBP, mm Hg, median [Q1, Q3]		15498	127.00 [117.00,140.00]	142.25 [130.00,157.00]	126.00 [116.00,140.00]	<0.001	-0.803
Glucose in urine, mmol/L, median [Q1, Q3]		21245	0.00 [0.00,0.00]	0.00 [0.00,3.00]	0.00 [0.00,0.00]	<0.001	-0.592
DBP, mm Hg, median [Q1, Q3]		15486	80.00 [74.00,89.00]	86.00 [78.00,94.00]	80.00 [74.00,88.00]	<0.001	-0.444

Alpha angle of electrical heart axis, degree, median [Q1, Q3]		26196	44.00 [12.00,65.00]	13.50 [-5.00,36.00]	45.00 [14.00,66.00]	<0.001	0.718
ESR, mm/hour, median [Q1, Q3]		5662	7.00 [4.00,12.00]	13.00 [7.00,24.00]	7.00 [4.00,12.00]	<0.001	-0.570
Blood oxygen saturation, %, median [Q1, Q3]		20788	98.00 [97.00,99.00]	98.00 [97.00,98.00]	98.00 [97.00,99.00]	<0.001	0.316
Liver right lobe thickness, mm, median [Q1, Q3]		24893	129.71 [118.87,143.79]	146.49 [132.15,161.41]	129.48 [118.70,143.23]	<0.001	-0.737
Mycoplasma hominis, Lg (GE), median [Q1,Q3]		37985	0.00 [0.00,0.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]	0.570	0.261
LV end-diastolic diameter, mm, median [Q1, Q3]		35132	49.59 [46.28,53.39]	52.45 [49.03,56.00]	49.41 [46.25,53.33]	<0.001	-0.549
Prolactin, mIU/mL, median [Q1, Q3]		36718	349.30 [246.00,494.50]	263.25 [179.52,409.82]	350.10 [246.70,496.60]	0.006	0.274
HDL-C, mmol/L, median [Q1, Q3]		29142	1.36 [1.11,1.68]	1.18 [0.98,1.42]	1.37 [1.12,1.69]	<0.001	0.534
Triglycerides, mmol/L, median [Q1, Q3]		28043	1.21 [0.85,1.76]	1.81 [1.32,2.60]	1.19 [0.84,1.72]	<0.001	-0.597
LV end-systolic volume, cm^3, median [Q1,Q3]		35133	40.74 [33.14,50.24]	51.59 [41.51,63.04]	40.40 [33.07,49.86]	<0.001	-0.633
LA diameter, mm, median [Q1,Q3]		35132	36.47 [32.94,40.00]	40.31 [36.85,44.12]	36.32 [32.78,40.00]	<0.001	-0.780
Liver left lobe thickness, mm, median [Q1, Q3]		24895	62.80 [54.91,72.84]	72.15 [63.14,82.69]	62.58 [54.79,72.59]	<0.001	-0.627
Monocytes, %, median [Q1, Q3]		4163	8.80 [7.50,10.30]	8.30 [7.10,9.85]	8.80 [7.60,10.30]	<0.001	0.206
VLDL-C, mmol/L, median [Q1,Q3]		29152	0.56 [0.39,0.81]	0.83 [0.61,1.20]	0.55 [0.39,0.79]	<0.001	-0.642
Right ovary volume, cm^3, median [Q1, Q3]		32005	6.08 [3.25,9.45]	2.15 [1.41,4.96]	6.11 [3.29,9.47]	<0.001	0.283
Left ovary volume, cm^3, median [Q1, Q3]		31993	5.68 [3.16,8.87]	2.67 [1.28,5.00]	5.70 [3.19,8.90]	<0.001	0.329
Head of the pancreas thickness, mm, median [Q1 ,Q3]		26198	25.24 [21.76,28.89]	28.99 [25.75,32.55]	25.17 [21.70,28.77]	<0.001	-0.641
Basophiles, 10^9 cells/L, median [Q1, Q3]		4163	0.02 [0.02,0.03]	0.03 [0.02,0.04]	0.02 [0.02,0.03]	<0.001	-0.275
EF, %, median [Q1, Q3]		35132	64.36 [61.47,67.52]	61.44 [56.70,64.62]	64.48 [61.59,67.62]	<0.001	0.692
Eosinophiles, 10^9/L, median [Q1, Q3]		4163	0.13 [0.08,0.21]	0.15 [0.10,0.23]	0.13 [0.08,0.21]	<0.001	-0.106
Neutrophils, 10^9/L, median [Q1, Q3]		4164	3.40 [2.66,4.35]	4.16 [3.22,5.23]	3.38 [2.65,4.33]	<0.001	-0.443
LV end-systolic diameter, mm, median [Q1, Q3]		35132	31.93 [29.37,34.82]	35.25 [32.18,38.28]	31.85 [29.31,34.71]	<0.001	-0.697
RDW-SD, fL, median [Q1, Q3]		4173	42.20 [40.40,44.40]	43.30 [41.20,45.60]	42.20 [40.30,44.40]	<0.001	-0.257
MCH, pg, median [Q1, Q3]		4155	29.60 [28.40,30.60]	29.60 [28.20,30.60]	29.60 [28.40,30.60]	0.819	-0.030

Urea, mmol/L, median [Q1, Q3]		18426	4.50 [3.65,5.60]	5.49 [4.38,6.85]	4.50 [3.62,5.54]	<0.001	-0.511
LV end-diastolic volume, cm^3, median [Q1, Q3]		35133	115.98 [98.73,137.65]	132.13 [113.27,155.25]	115.03 [98.62,137.32]	<0.001	-0.550
Eosinophiles, %, median [Q1, Q3]		4452	2.10 [1.40,3.20]	2.10 [1.40,3.10]	2.10 [1.40,3.20]	0.396	0.061
Uric acid, mmol/L, median [Q1, Q3]		28563	295.00 [240.00,362.00]	337.00 [284.00,409.00]	293.00 [239.00,360.00]	<0.001	-0.481
ALT, U/L, median [Q1, Q3]		10282	16.60 [11.60,25.60]	23.30 [15.30,36.50]	16.50 [11.50,25.30]	<0.001	-0.263
Alpha amylase, U/L, median [Q1, Q3]		22652	54.70 [42.00,69.80]	45.40 [33.50,60.00]	55.00 [42.30,70.00]	<0.001	0.328
RR interval, msec, median [Q1, Q3]		26083	842.00 [750.00,938.00]	797.50 [693.00,889.75]	845.00 [752.00,939.00]	<0.001	0.344
LV posterior wall thickness, mm, median [Q1, Q3]		35133	9.41 [8.37,10.40]	10.00 [9.26,11.00]	9.41 [8.36,10.30]	<0.001	-0.530
RDW-CV, %, median [Q1, Q3]		4172	13.10 [12.60,13.80]	13.50 [12.80,14.20]	13.10 [12.60,13.80]	<0.001	-0.211
WBC, 10^9 cells/L, median [Q1, Q3]		4155	6.42 [5.35,7.70]	7.47 [6.23,8.95]	6.39 [5.33,7.66]	<0.001	-0.408
Atherogenic index of plasma, median [Q1, Q3]		29144	2.68 [1.89,3.66]	3.36 [2.54,4.52]	2.65 [1.87,3.63]	<0.001	-0.485
Body of the pancreas thickness, mm, median [Q1, Q3]		26196	13.00 [10.57,15.80]	15.90 [13.13,18.63]	13.00 [10.53,15.65]	<0.001	-0.670
Left kidney width, mm, median [Q1, Q3]		29433	51.02 [47.00,55.54]	53.12 [49.71,58.18]	51.00 [46.91,55.48]	<0.001	-0.399
Left ovary length, mm, median [Q1, Q3]		31879	28.50 [23.48,33.36]	22.54 [17.34,28.02]	28.55 [23.58,33.40]	<0.001	0.596
Left ovary thickness, mm, median [Q1, Q3]		31879	17.23 [13.49,20.78]	12.66 [9.72,16.25]	17.27 [13.53,20.82]	<0.001	0.554
Right kidney length, mm, median [Q1, Q3]		29428	107.66 [102.10,114.00]	110.94 [103.99,118.60]	107.60 [102.07,113.92]	<0.001	-0.300
Basophiles, %, median [Q1, Q3]		4163	0.40 [0.20,0.50]	0.40 [0.20,0.50]	0.40 [0.20,0.50]	0.160	-0.023
Left kidney length, mm, median [Q1, Q3]		29434	108.93 [102.85,115.64]	112.16 [105.89,119.80]	108.88 [102.81,115.50]	<0.001	-0.320

Table 1. General characteristics of training dataset featuring T2D model predictors. P-values of continuous and categorical variables are measured by Kruskal-Wallis test and Chi-squared test, respectively. 0 male sex, 1 female sex, SMD standardized mean difference, SBP systolic blood pressure**,** DBP systolic blood pressure, ESR erythrocyte sedimentation rate, LV left ventricle of the heart, LA left atrium of the heart, HDL-C high-density lipoprotein cholesterol, VLDL-C very low-density lipoprotein cholesterol, EF ejection fraction of the heart, RDW-SD red cell distribution width-standard deviation, RDW-CV red cell distribution width-coefficient of variation, MCH mean cell hemoglobin, ALT alanine aminotransferase, WBC white blood cells.

	Split	AUC	EER	F1 score	Macro F1 score	Sensitivity	Specificity	Precision	NPV
MLP	#1	0.926	15.773	0.224	0.558	0.890	0.808	0.213	0.996
	#2	0.960	9.716	0.346	0.644	0.915	0.893	0.213	0.997
	#3	0.955	10.464	0.358	0.653	0.886	0.903	0.224	0.996

	Mean	0.947	11.984	0.309	0.618	0.897	0.868	0.188	0.996
Ridge classifier	#1	0.946	12.836	0.430	0.698	0.789	0.940	0.295	0.993
	#2	0.916	16.753	0.310	0.626	0.776	0.898	0.194	0.992
	#3	0.917	15.794	0.305	0.622	0.801	0.891	0.188	0.993
	Mean	0.926	15.128	0.348	0.649	0.789	0.910	0.226	0.993
	#1	0.958	10.993	0.431	0.697	0.850	0.934	0.289	0.995
Ridge classifier with polynomial features	#2	0.946	12.254	0.325	0.633	0.862	0.891	0.200	0.995
	#3	0.947	13.001	0.256	0.583	0.902	0.837	0.149	0.996
	Mean	0.950	12.083	0.337	0.638	0.871	0.887	0.213	0.995
XGBoost	#1	0.961	10.515	0.343	0.643	0.902	0.894	0.212	0.997
	#2	0.939	12.588	0.321	0.630	0.862	0.889	0.197	0.995
	#3	0.934	15.121	0.363	0.658	0.809	0.916	0.234	0.993
	Mean	0.945	12.741	0.342	0.644	0.858	0.900	0.214	0.995
Bootstrap aggregating	#1	0.956	11.482	0.386	0.670	0.862	0.917	0.249	0.995
	#2	0.933	13.484	0.272	0.596	0.874	0.856	0.161	0.995
	#3	0.934	14.628	0.339	0.643	0.825	0.904	0.213	0.994
	Mean	0.941	13.198	0.332	0.636	0.854	0.892	0.208	0.995
Stacked Generalization	#1	0.962	10.957	0.380	0.666	0.882	0.913	0.242	0.996
	#2	0.955	10.925	0.374	0.662	0.878	0.911	0.237	0.996
	#3	0.961	10.093	0.402	0.679	0.890	0.920	0.260	0.996
	Mean	0.959	10.658	0.385	0.669	0.883	0.915	0.246	0.996

Table 2. Performance metrics for the first train/test split.

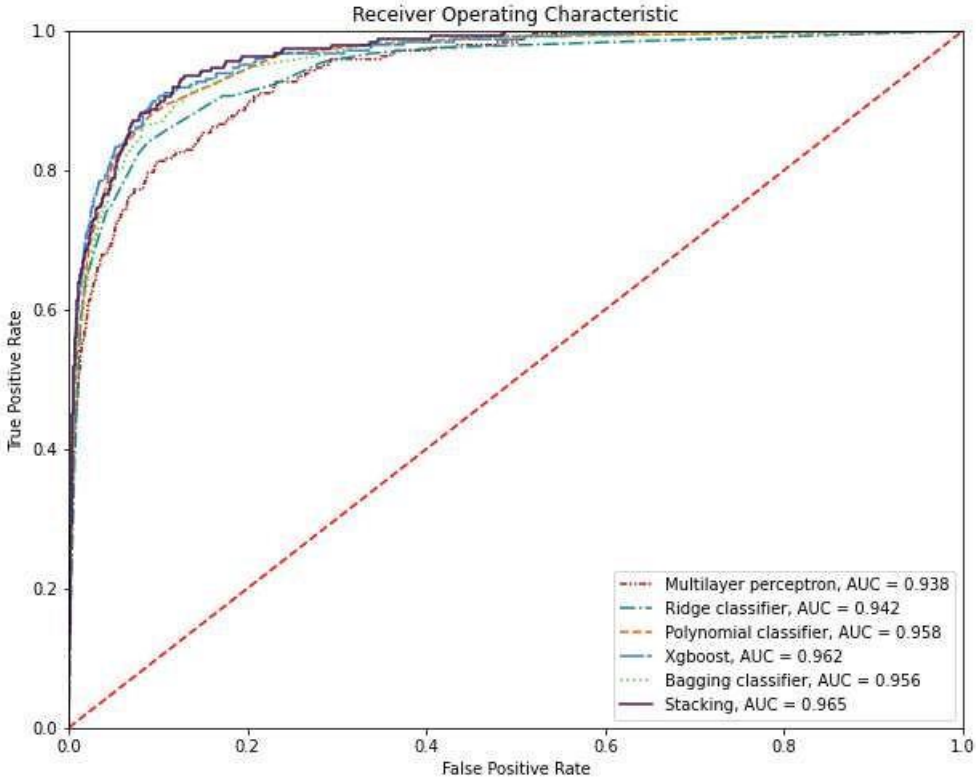

Figure 2. ROC curves of all the models.

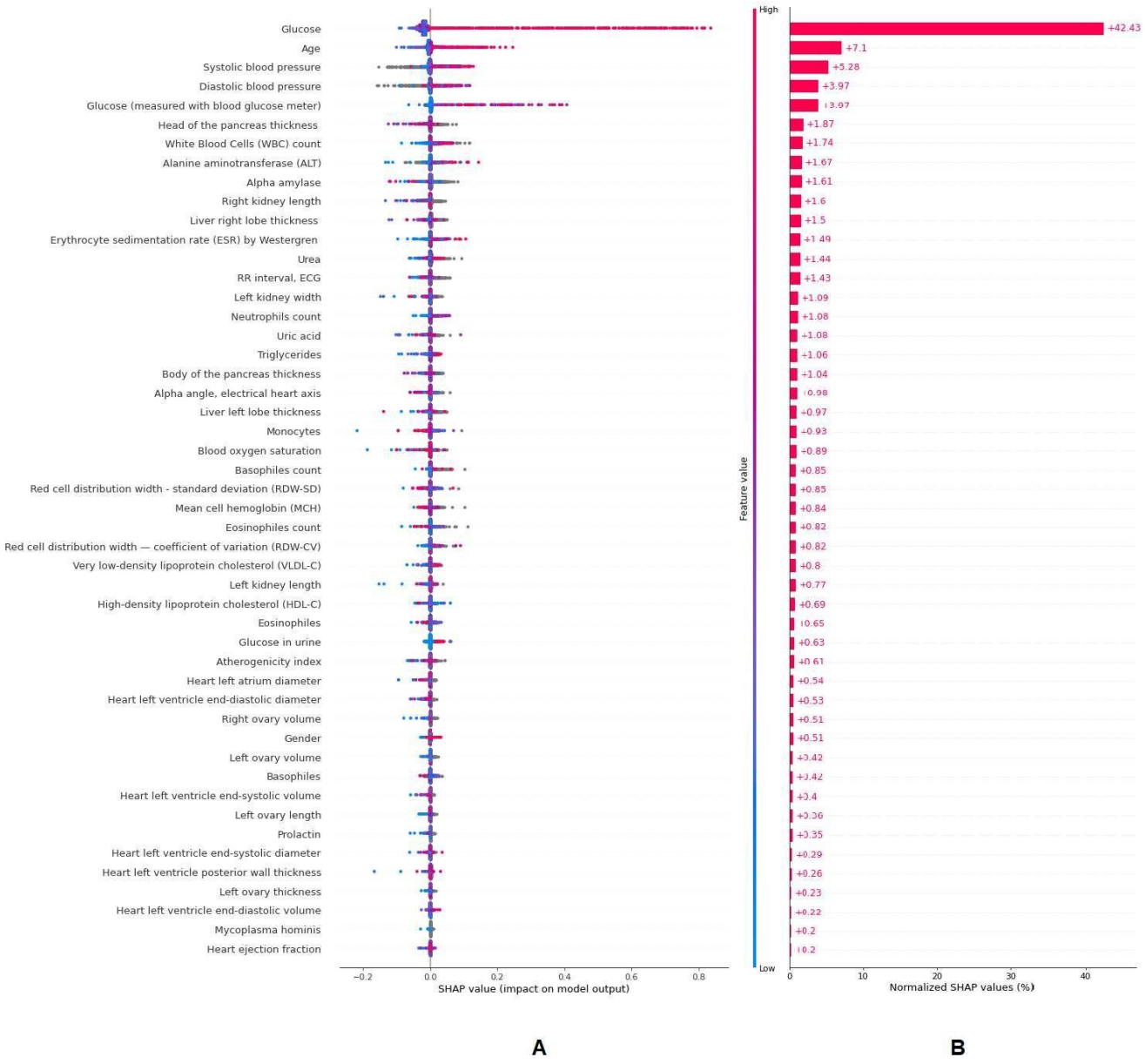

Figure 3. A: beeswarm plots. Each dot corresponds to some patient case. The dot's position on the horizontal axis shows the impact that feature had on the prediction. B: bar chart of the normalized average absolute SHAP values.

		Missing	Overall	diabetic	healthy	P-Value	SMD (diabetic,healthy)
n			4011	654	3357		
Eosinophiles, %, median [Q1,Q3]		4	2.40 [1.60,3.70]	2.50 [1.70,3.80]	2.40 [1.50,3.70]	0.020	-0.069
Systolic blood pressure, mm Hg, median [Q1,Q3]		393	120.67 [110.00,133.67]	129.00 [115.67,144.67]	119.33 [109.33,131.67]	<0.001	-0.457
Basophiles, %, median [Q1,Q3]		4	0.80 [0.60,1.00]	0.80 [0.60,1.00]	0.80 [0.60,1.00]	0.010	-0.103
Uric acid, mmol/L,		90	315.20	327.10	315.20	<0.001	-0.201

median [Q1,Q3]			[261.70,374.70]	[269.17,392.60]	[261.70,368.80]		
Alanine aminotransferase (ALT), U/L, median [Q1,Q3]		90	17.00 [13.00,25.00]	18.00 [14.00,27.00]	17.00 [12.00,25.00]	0.001	-0.060
Neutrophils count, 10^9/L, median [Q1,Q3]		4	3.70 [2.80,4.70]	4.20 [3.30,5.40]	3.60 [2.80,4.60]	<0.001	-0.381
Eosinophiles count, 10^9/L, median [Q1,Q3]		4	0.20 [0.10,0.20]	0.20 [0.10,0.30]	0.20 [0.10,0.20]	<0.001	-0.170
High-density lipoprotein cholesterol (HDLC), mmol/L, median [Q1,Q3]		72	1.32 [1.09,1.60]	1.22 [1.02,1.42]	1.34 [1.11,1.63]	<0.001	0.377
Age, median [Q1,Q3]		0	51.00 [34.00,64.00]	63.00 [54.00,72.00]	47.00 [32.00,62.00]	<0.001	-0.963
Triglycerides, mmol/L, median [Q1,Q3]		90	1.15 [0.82,1.66]	1.45 [1.06,2.00]	1.09 [0.79,1.57]	<0.001	-0.324
Mean cell hemoglobin (MCH), pg, median [Q1,Q3]		2	30.00 [28.60,31.10]	29.80 [28.40,31.00]	30.00 [28.60,31.10]	0.046	0.065
Diastolic blood pressure, mm Hg, median [Q1,Q3]		393	73.33 [66.33,81.33]	73.00 [65.00,81.00]	73.33 [66.33,81.33]	0.155	0.079
Monocytes, %, median [Q1,Q3]		4	8.00 [6.80,9.50]	7.90 [6.70,9.20]	8.00 [6.80,9.50]	0.032	0.059
White Blood Cells (WBC) count, 10^9/L, median [Q1,Q3]		2	6.50 [5.40,7.90]	7.10 [5.90,8.50]	6.30 [5.30,7.70]	<0.001	-0.346
Urea, mmol/L, median [Q1,Q3]		90	5.00 [3.93,6.07]	5.71 [4.28,7.50]	4.64 [3.93,6.07]	<0.001	-0.537
Red cell distribution width — coefficient of variation (RDWCV), %, median [Q1,Q3]		2	13.60 [13.10,14.30]	13.90 [13.30,14.80]	13.50 [13.00,14.20]	<0.001	-0.271
Sex, median [Q1,Q3]	0	0	1951 (48.64)	348 (53.21)	1603 (47.75)	0.012	0.109
	1		2060 (51.36)	306 (46.79)	1754 (52.25)		
Basophiles count, 10^9/L, median [Q1,Q3]		4	0.00 [0.00,0.10]	0.10 [0.00,0.10]	0.00 [0.00,0.10]	<0.001	-0.292
Glucose, mmol/L, median [Q1,Q3]		0	5.72 [5.30,6.33]	7.99 [6.50,10.20]	5.61 [5.27,6.00]	<0.001	-1.245

Table 3. General characteristics of external validation dataset. P-values of continuous and categorical variables are measured by Kruskal-Wallis test and Chi-squared test, respectively. 0 male sex, 1 female sex, SMD standardized mean difference.

	T2D status = true	T2D status = false
Fasting plasma glucose level ≥ 7 mmol/L	739 patients	773 patients
Fasting plasma glucose level < 7 mmol/L	491 patients	38 036 patients

Table 4. Contingency table of fasting plasma glucose level and diabetes status.

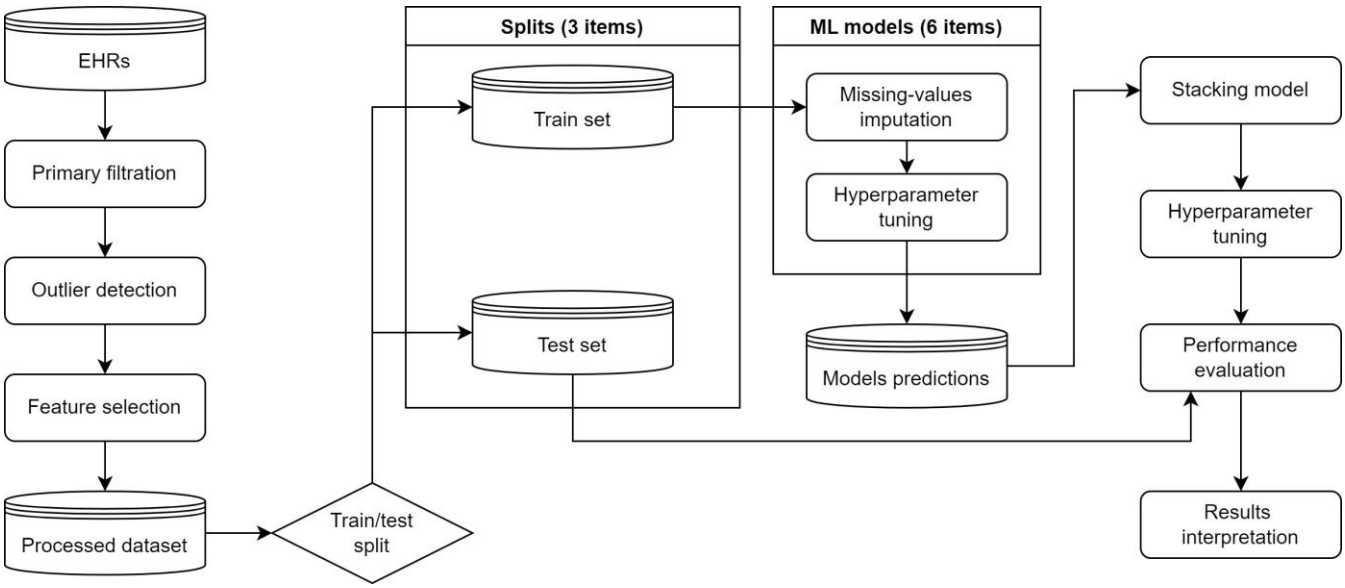

Figure 4. Workflow chart of diagnostic model development.