Computers in Biology and Medicine 187 (2025) 109771

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

REMED-T2D: A robust ensemble learning model for early detection of type 2 diabetes using healthcare dataset☆

Le Thi Phan [a], Rajan Rakkiyappan [b], Balachandran Manavalan [a,*]

[a] Computational Biology and Bioinformatics Laboratory, Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, 16149, Gyeonggi-do, Republic of Korea
[b] Department of Mathematics, Bharathiar University, Coimbatore, 641046, Tamil Nadu, India

ARTICLE INFO

Keywords:
Diabetes
Ensemble learning
Machine learning
Pima indian diabetes
Random sampling technique

ABSTRACT

Early diagnosis and timely treatment of diabetes are critical for effective disease management and the prevention of complications. Undiagnosed diabetes can lead to an increased risk of several health issues. Although numerous machine learning (ML) models have been designed to detect diabetes, many exhibit unsatisfactory performance, are not publicly available, and lack validation on external datasets. To address these limitations, we have developed REMED-T2D, an advanced ensemble ML approach that enhances predictive accuracy and robustness through the integration of diverse ML algorithms. Our approach involves a rigorous data preprocessing process and systematic evaluation of 20 different algorithms, encompassing both conventional ML and deep learning for diabetes prediction. Firstly, we applied an under-sampling approach to an imbalanced Pima Indian Diabetes dataset and generated five balanced datasets. Using these datasets, we investigated various computational strategies to select the optimal model for accurate diabetes classification. Our results demonstrate that REMED-T2D outperformed state-of-the-art methods on the training dataset, with notable improvements in ACC (1.40–4.60%) and MCC (3.50–9.80%). Extensive external validations revealed that the model trained on a five-feature subset achieved ACC of 92.61 % and 92.26 % on the RTML1 and Pabna datasets, respectively. Moreover, a model based on a seven-feature subset improved ACC by 2.80 % and MCC by 13.27 % on the RTML2 dataset. These results suggest the potential of REMED-T2D to predict diabetes in Asian females. Notably, this is the first study to conduct such a comprehensive analysis using the Pima dataset, incorporating a diverse set of ML algorithms. Furthermore, we have developed a publicly accessible web server (https://balalab-skku.org/REMED-T2D/) to facilitate self-monitoring and timely medical interventions. We believe REMED-T2D will assist healthcare professionals in detecting diabetes earlier and implementing preventive measures, ultimately improving health outcomes for those at risk.

1. Introduction

Diabetes is a chronic metabolic disorder characterized by insufficient insulin production or impaired insulin action and stands as one of the top ten leading causes of death globally [1–3]. The disruption in insulin signaling results in elevated blood glucose levels, leading to a cascade of health complications if it is not well-managed [1,4,5]. Diabetes is broadly classified into four types: type 1 diabetes, which is prevalent in children and requires insulin for survival [2–5]; type 2 diabetes (T2DM), the most common form in adults, marked by inadequate insulin production [1–4]; prediabetes, a state of impaired glucose tolerance with elevated blood glucose levels not yet meet the criteria for T2DM [2,3,6]; and gestational diabetes, which occurs during pregnancy and poses health risks for both mother and baby [2,3,5,6]. The exact cause of diabetes remains unclear, but research suggests that genetic predisposition, unhealthy lifestyle choices, and environmental factors like obesity and physical inactivity may contribute to its development [6,7].

The prevalence of diabetes has increased fourfold over the past 35 years, emerging as a major global health issue [1,7]. The International Diabetes Federation reported that in 2021, 536.6 million adults aged

☆ All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

* Corresponding author.
E-mail address: bala2022@skku.edu (B. Manavalan).

https://doi.org/10.1016/j.compbiomed.2025.109771
Received 30 October 2024; Received in revised form 31 December 2024; Accepted 29 January 2025
Available online 5 February 2025

20–79 years were living with diabetes, with the number expected to increase to 783.2 million by 2045 [2,3]. Notably, T2DM accounts for more than 90 % of these cases [8,9] and is a major cause of disability and mortality worldwide [1]. In 2021, T2DM and its complications, such as diabetic retinopathy, cardiovascular disease, neuropathy, nephropathy, foot ulcers, and stroke, led to a staggering 6.7 million deaths [2]. This underscores the critical importance of early screening and diagnosis to identify at-risk individuals and enable timely interventions. Regular check-ups, including glucose level monitoring and biomarker assessment, can facilitate early detection and help prevent or delay complications [10–12]. However, diagnosing early diabetes, especially in the initial stages, remains challenging [5,13]. The current diagnostic process requires analyzing a range of clinical data such as plasma glucose concentration, HbA1c test results, blood pressure, family history, smoking habits, body mass index, cholesterol levels, triglycerides, serum insulin levels, and age [14]. This process can be time-consuming, taking weeks or even months, and increasing the workload of physicians [15]. To address these challenges, computer-based systems that predict T2DM based on diabetes-associated risk factors offer a promising solution. These systems have the potential to enhance both the efficiency and accuracy of diabetes diagnosis [16].

Pima females, a Native American tribe residing in the southwestern United States, face a disproportionately high risk of developing diabetes due to a complex genetic interplay, lifestyle, and environmental factors [17]. Historically, the Pima people developed a thrifty metabolism that allowed them to store energy efficiently, adapting to their ancestral environment's scarcity of resources. However, in today's environments with abundant food and sedentary lifestyles, this thrifty metabolism has contributed to increased rates of obesity and insulin resistance among Pima women. Research has identified specific genetic variants within the Pima population are associated with reduced insulin sensitivity and impaired insulin secretion, thereby increasing their susceptibility to diabetes [18]. The high prevalence of diabetes among Pima females is linked to severe health complications, including cardiovascular diseases, renal impairment, retinopathy, neuropathy, and adverse pregnancy outcomes such as gestational diabetes and preeclampsia [19,20].

The Pima Indian Diabetes (PID) dataset is a comprehensive collection of medical and demographic data (268 diabetes cases and 500 non-diabetes cases) that serves as a crucial benchmark for evaluating machine learning (ML) models in healthcare research. Numerous ML algorithms have been applied to this dataset for diabetes prediction. Chang et al. [21] reported an accuracy (ACC) range of 74.78–79.57 % using Naïve Bayes (NB), random forest (RF), and J48 classifiers. Kalagotla et al. [22] achieved 78.20 % ACC using a stacking technique with support vector machine (SVM), multilayer perceptron (MLP), and logistic regression (LR) based on important features (glucose, BMI, age). Kumari et al. [23] proposed a soft voting approach combining RF, LR, and NB, attaining 79.04 % in ACC. Tasin et al. [24] addressed data imbalance using the adaptive synthetic sampling approach (ADASYN) and trained it with XGBOOST (XGB), achieving 81.00 % ACC. Reza et al. [25] utilized the synthetic minority over-sampling technique (SMOTE) with SVM, obtaining 85.50 % ACC, while Hairani et al. [26] applied SMOTE-Tomek Links with RF, reaching 86.40 % in ACC. Ramesh et al. [27] also employed SMOTE with SVM, achieving an ACC of 83.20 %. Wang et al. [28] employed ADASYN with RF, reached 86.20 % ACC. Despite significant advancements in diabetes prediction, current methodologies face three major limitations that reduce their effectiveness: (i) These methods tend to focus on improving accuracy with the PID dataset without assessing the robustness of their models on other diabetes-related datasets, thereby limiting their practical application; (ii) Lack accessible web-based platforms prevents widespread use by physicians for proactive screening and preventive interventions; (iii) There is still considerable room for enhancing the accuracy and the generalizability of the existing predictors.

In this study, we meticulously pre-processed the PID dataset by addressing dataset imbalance through under-sampling and exploring different imputation techniques. We generated five balanced datasets and rigorously evaluated the performance of 20 different classifiers. Our analysis included 14 conventional ML models (RF, extremely randomized trees (ERT), AB, XGB, gradient-boosting machine (GB), light GB (LGB), CatBoost (CB), SVM with radial basis function (rbf), linear, polynomial (poly), and sigmoid kernels, MLP, NB, and LR) and 6 deep learning (DL) algorithms (long short-term memory (LSTM), bidirectional LSTM (Bi-LSTM), gated recurrent unit (GRU), bidirectional GRU (Bi-GRU), convolutional neural network (CNN), and deep neural network (DNN)) for diabetes prediction. Subsequently, we systematically explored various strategies to identify the most optimal model for diabetes classification (Fig. 1). Notably, our ensemble-based aggregation approach demonstrated superior performance compared to existing models trained on the PID dataset. This robust performance was further validated on publicly available datasets from Bangladeshi women, highlighting its potential for accurate diabetes prediction among Asian females. To our knowledge, this is the first study to utilize such an extensive range of ML algorithms for diabetes prediction, offering the first large-scale analysis of the PID dataset. Additionally, we have developed a publicly accessible web server available at https://balalab-skku.org/REMED-T2D/ for real-time online diabetes testing, providing a valuable resource for researchers and enabling practical applications in diabetes prediction.

2. Materials and methods

2.1. Benchmark and case study datasets

The PID dataset, obtained from the National Institute of Diabetes and Digestive and Kidney Diseases, is publicly accessible at https://data.world/uci/pima-indians-diabetes [29]. This dataset comprises 768 samples from Pima Indian women, aged 21–81 years. Of these samples, 500 represent women who do not have diabetes (denoted as "0" in the Outcome), while 268 represent women diagnosed with diabetes (denoted as "1" in the Outcome). The dataset includes eight attributes: diastolic blood pressure (DBP), plasma glucose concentration following a 2-hour oral glucose tolerance test (glucose), 2-hour serum insulin (insulin), triceps skinfold thickness (skinfold), age, body mass index (BMI), number of pregnancies (preg), and diabetes pedigree function (pedi). For further details, please refer to Table 1.

In contrast to prior studies, we evaluated the transferability of our trained model using additional publicly available datasets, which are as follows.

(1) The RTML private dataset, obtained from a previous study [27], includes 203 female participants (aged 18–77 years old) in Dhaka, Bangladesh. It consists of features: glucose, preg, DBP, skinfold, insulin, BMI, and age. Of those samples, 34 had diabetes, and 169 instances with non-diabetes. The dataset is available at https://github.com/tansin-nabil/Diabetes-Prediction-Using-Machine-Learning. It is divided into two subsets: RTML1, which contains 203 samples (34 diabetes, 169 non-diabetes) and five-feature (glucose, BMI, age, DBP, and preg), and RTML2, which contains 109 samples (23 diabetes, 86 non-diabetes) and seven-feature (five-feature with the addition of insulin and skinfold).
(2) The Pabna dataset [30], includes data from 465 female patients (aged 21–86 years) at Pabna Diabetes Hospital, Bangladesh. The dataset consists of 372 diabetic and 93 non-diabetic patients. It includes variables such as preg, age, BMI, DBP, genetic factors, insulin, and glucose concentration (measured during an oral glucose tolerance test after 2-hour). The dataset can be downloaded from https://github.com/ruhul256/Pabna-Diabetes-Dataset-Bangladesh.

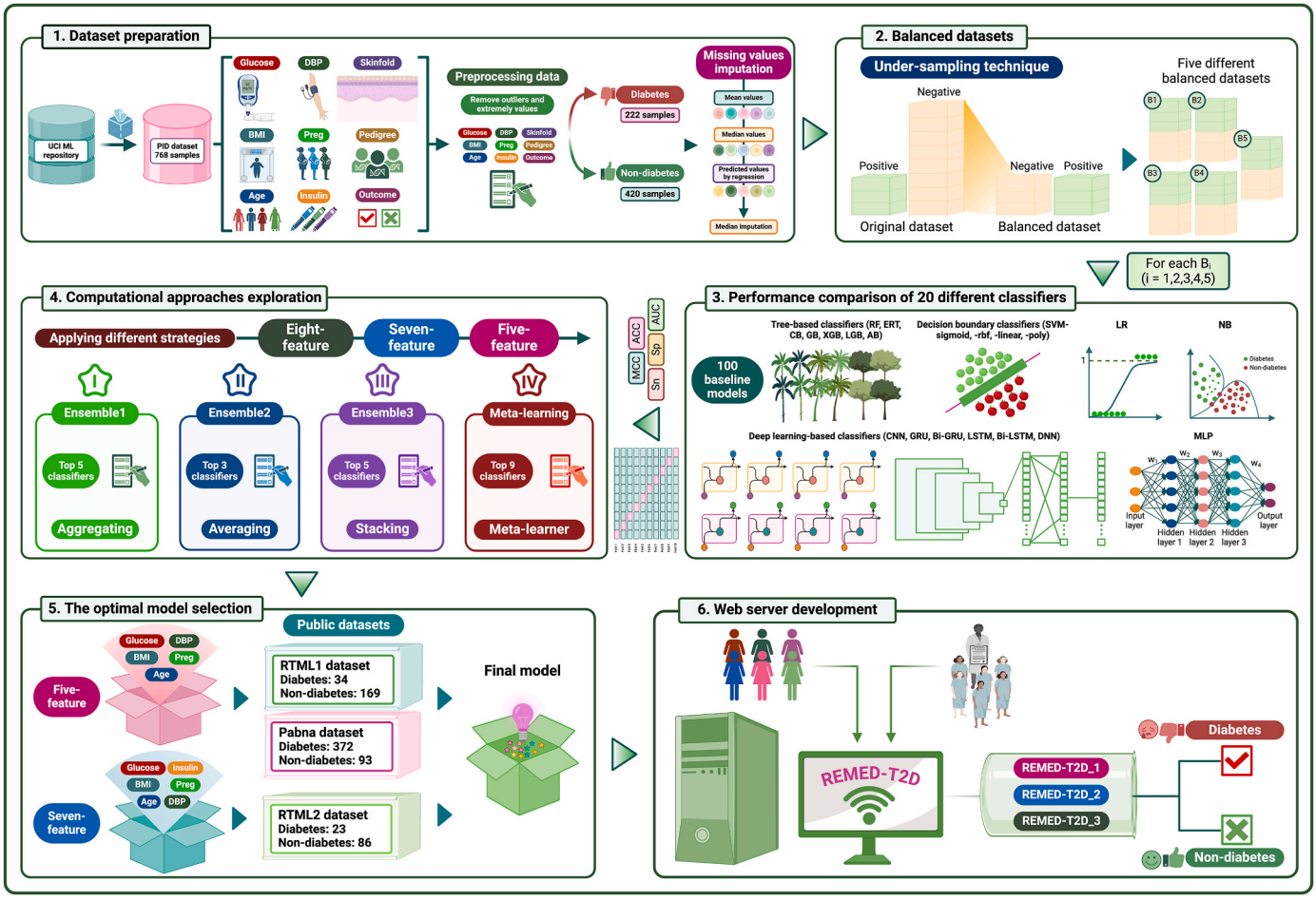

Fig. 1. The overall workflow of the REMED-T2D. It comprises the following steps: (i) Dataset preparation, including collecting and preprocessing of the PID data; (ii) Generation of balanced datasets using under-sampling technique; (iii) Performance evaluation of 20 classifiers on balanced datasets; (iv) Exploration of different computational strategies; (v) The optimal model selection; (vi) Web-based diabetes prediction tool development.

Table 1
Description of the PID dataset.

Attribute	Type	Description	Min	Max	Mean ± SD	Collection method
Preg	Integer	The number of times a woman has been pregnant	0	17	3.845 ± 3.370	Demographic
Glucose	Integer	Glucose concentration in the plasma (mg/dL)	0	199	120.895 ± 31.973	Laboratory test
DBP	Integer	The blood pressure in the arteries between heartbeats when the heart is resting (mmHg)	0	122	69.105 ± 19.356	Laboratory test
Skinfold	Integer	A value used to estimate body fat levels (mm)	0	99	20.536 ± 15.952	Laboratory test
Insulin	Integer	Insulin concentration in the bloodstream (mu U/ml)	0	846	76.799 ± 115.244	Laboratory test
BMI	Float	Body mass index (kg/m^2)	0	67.100	31.992 ± 7.884	Laboratory test
Pedi	Float	Quantifying the likelihood of diabetes based on family history	0.078	2.420	0.472 ± 0.331	Laboratory test
Age	Integer	Years	21	81	33.241 ± 11.760	Demographic
Outcome	Integer	Non-diabetes (500): 0 Diabetes (268): 1				

Preg, pregnant; DBP, diastolic blood pressure; BMI, body mass index; Pedi, pedigree; SD, standard deviation.

2.2. Data preprocessing

The original PID dataset contains unreasonable zero values in several attributes (glucose, DBP, skinfold, insulin, and BMI) as well as outliers. To ensure the robustness of our ML models, we preprocessed the data into two steps: (i) outliers and extreme values removal and (ii) missing value imputation.

Initially, we employed the interquartile range method [31] to identify and remove outliers and extreme values. This reduced the number of valid observations from 768 to 648 (Table 2). Subsequently, we addressed missing values in the cleaned PID dataset using three imputation methods.

(1) Regression-based imputation: we utilized regression models to impute missing values for insulin and skinfold variables, which had a significant number of missing entries (Table 2). We divided data into different subsets based on the presence or absence of these values: subdata1 and subdata2 for insulin, and subdata3 and subdata4 for skinfold measurements. Subsequently, we trained 15 distinct regression models on subdata1 and subdata3 with an extensive search range (Table S1). Of those, we have selected the top three models based on the root mean squared error (RMSE), mean absolute error (MAE), and the correlation coefficient (CC). This resulted in CB, AB, and SVM-rbf for insulin prediction, and CB, AB, and LGB for skinfold prediction

Table 2
The number of missing values for each attribute in the PID dataset before and after removing outlier and extreme values.

Attributes	Original dataset	After applying an interquartile range
Preg	0	0
Glucose	5	5
DBP	35	33
Skinfold	227	210
Insulin	374	345
BMI	11	9
Pedi	0	0
Age	0	0
Class	Non-diabetes: 500 Diabetes: 268	Non-diabetes: 462 Diabetes: 222

(Table S2). We applied these models to predict missing values in the corresponding subsets, averaged the predicted values from the top three models, and used these values to replace the missing entries in the original dataset. Finally, we imputed the remaining missing values for glucose, DBP, and BMI using median values.

(2) Mean-based imputation: We replaced missing values with the mean of their respective attributes.

(3) Median-based imputation: We replaced missing values with the median of their respective attributes.

After applying these imputation processes, we constructed three new datasets. Subsequently, we assessed the performance of each imputation method using 20 different classifiers (see section 2.5 for details). These analyses enabled us to identify the most suitable imputation approach for diabetes classification.

2.3. Construction of balanced training datasets

The constructed dataset exhibits a class imbalance, with more negative (non-diabetes) instances than positive (diabetes) instances. This imbalance can bias predictions and significantly impact model performance [32]. To address the skewed class distribution in the training dataset, we employed an under-sampling technique to create five balanced datasets. Specifically, we randomly selected 222 (matching the minority class samples) from the original pool of 462 non-diabetes samples, repeating this process five times to create five distinct subsets. Each of these five subsets was then combined with all 222 diabetes cases, resulting in five balanced datasets (BD1–BD5), each containing 444 samples.

2.4. Development and investigation of base-classifiers

For each BD, we trained 100 baseline models using 20 distinct classifiers (including 14 ML-based (SVM with four different kernels, NB, LR, MLP, RF, ERT, CB, GB, XGB, LGB, and AB) and 6 DL-based (GRU, LSTM, Bi-GRU, Bi-LSTM, CNN, and DNN). During model training, we employed 10-randomized 10-fold cross-validation (CV) along with an extensive grid search to optimize relevant hyperparameters (Table S3).

2.4.1. Machine learning-based classifiers

2.4.1.1. Support vector machine (SVM). SVM is a powerful algorithm broadly used in different problems [33–35]. It maps input features into a high-dimensional space to identify the optimal hyperplane that maximizes the separation between diabetes and non-diabetes. Four SVM kernels were employed, including linear, rbf, poly, and sigmoid. Two SVM parameters (gamma (G) and penalty (C)) were optimized using a grid search range (Table S3). Strikingly, preliminary analysis showed that SVM-rbf outperformed other kernel functions (SVM-sigmoid, -poly, and -linear).

2.4.1.2. Naïve Bayes (NB). NB is a classification algorithm that simplifies learning by assuming the independence of predictors [36]. Based on Bayes' theorem, NB considers that the presence of a particular feature in a class is independent of the presence of any other features. Consequently, each feature contributes equally to the output [37]. NB can be mathematically represented as:

$$\widehat{y} = \underset{y}{\operatorname{argmax}} P(y)\prod_{i=1}^{n} P(x_i|y) \quad (1)$$

where, $x_1, x_2, \ldots, x_n$ are the features, y is the target variable, $P(y)$ is the prior probability of y, and $P(x_i|y)$ are the posterior probabilities.

2.4.1.3. Logistic regression (LR). LR is a commonly used statistical ML method. It predicts an output value (y) by linearly combining input values (x) with coefficient values. The predicted output ($yhat$) is a real value between 0 and 1, calculated with the sigmoid function as:

$$yhat = \frac{1}{1 + e^{-(b0+b1*x1)}} \quad (2)$$

where $b0$ is the intercept, $b1$ is the coefficient, and e is Euler's number [38].

2.4.1.4. Multilayer perceptron. MLP is a neural network that comprises input, output, and hidden layers [39]. The input layer receives data, and the output layer produces results. Hidden layers lie between the input and output layers, allowing MLPs to learn and capture nonlinear relationships in the data. Generally, neural networks mimic the behavior of human neurons, but processing time can be lengthy. Three MLP parameters (hidden_layer_sizes, max_iter, and learning_rate) were optimized using the grid search.

2.4.1.5. Decision tree-based classifiers. Decision tree-based classifiers are suitable for handling unnormalized features [40,41]. Hence, we utilized seven decision tree-based classifiers (RF, ERT, AB, GB, XGB, LGB, and CB) to construct baseline models due to their successful application in bioinformatics [42–44]. The implementation of classifiers following established procedures from previous studies [35,40,41,45]. Appropriate hyperparameters were selected for each tree-based model to find their optimal model (Table S3).

2.4.2. Deep learning-based classifiers

DL-based models can enhance diabetes prediction by capturing temporal dependencies and relevant features within the data [46]. In this study, we utilized various DL-based algorithms to construct baseline models. Specifically, we used GRU to effectively capture sequential dependencies through its gating mechanisms [47,48], while Bi-GRU further improved context understanding by processing data in both backward and forward directions [48]. Unlike GRU, LSTM was used to address the vanishing gradient problem and capture long-term dependencies in the data, leveraging its advanced memory units [46,48,49]. Like Bi-GRU, Bi-LSTM also processed data bidirectionally, but with a larger number of memory units [49]. Apart from utilizing recurrent neural network methods, we further employed CNN to extract local features and DNN to learn complex patterns and relationships within the data [46,50,51]. To optimize the DL-based models, we performed an extensive grid search for hyperparameter tuning (Table S3).

2.5. Identifying key risk factors for diabetes using the feature importance score approach

To identify the most crucial risk factors for diabetes, we computed the feature importance score (FIS). While tree-based classifiers are commonly used to rank features based on their FIS [52,53], these scores alone may not accurately reflect the true importance of each feature. Therefore, we processed the original features through seven classifiers

(RF, ERT, AB, GB, XGB, LGB, and CB), and normalized each method's FIS values to ensure consistency. We then calculated the average FIS across all classifiers, which we termed an ensemble FIS.

Based on the ensemble FIS values, we systematically divided the original feature set into five subsets, ranging from 3- to 7-dimensional (D) features, by iteratively adding the top-ranked features. We then used these subsets to train by top five classifiers and averaged their performances across five balanced datasets. The performance of these models was compared and identified the most predictive risk factors for diabetes. This methodology can be applied to other disease-related problems, such as predicting lung, liver, and kidney-related diseases. Generally, by identifying the most influential indicators, researchers can develop more accurate and targeted diagnostic and treatment strategies.

2.6. Exploration of various computational frameworks

Ensemble models have been shown to consistently outperform their baseline models [41,54–57]. Ensemble learning techniques leverage the strengths of multiple predictive models, resulting in a more robust meta-model. To the best of our knowledge, this study is the first to employ a wide range of ML classifiers and explore different ensemble strategies for predicting diabetes. The adopted strategies are described as follows.

(1) Aggregation strategy (Ensemble1): We select the top five classifiers across all balanced datasets (BD1–BD5). The probability scores (PCs) from these classifiers are concatenated and an average score is computed for each sample.
(2) Ensemble2: We select the top three classifiers for each BD. The average PCs of these classifiers and calculate performance metrics per BD. Finally, we obtain the overall performance metrics by averaging the results across BD1–BD5.
(3) Ensemble 3: Similar to Ensemble1, the top five classifiers are selected for each BD. PCs from these classifiers across BD1–BD5 are concatenated to create a new 25D feature set. This feature set is used to train the model with seven classifiers: RF, CB, ERT, GBM, XGB, LGB, and AB. Finally, we select the model with the best performance.
(4) Meta-learner strategy (Meta-learner): The top nine classifiers are selected for each BD and their PCs are concatenated to create a 9-D feature vector for each. This feature set is then fed into 10 different classifiers, including RF, CB, ERT, GBM, XGB, LGB, AB, MLP, SVM, and LR. The average performance of each classifier is computed across BD1-BD5, and the final model with the best performance is selected.

2.7. Performance evaluation

The performance evaluation of REMED-T2D and its baseline models was conducted using four widely recognized performance metrics: sensitivity (Sn), specificity (Sp), ACC, and MCC [58–61]. The calculation for each metric is shown as follows:

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} & (3) \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} & (4) \\ \mathrm{ACC} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} & (5) \\ \mathrm{MCC} = \dfrac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} & (6) \end{cases}$$

In these equations, TP and TN denote the number of correctly classified diabetes and non-diabetes instances, respectively. Conversely, FP and FN represent the number of non-diabetes and diabetes instances that are incorrectly classified, respectively.

3. Results

3.1. Feature distribution analysis and their contribution to class discrimination in the PID dataset

We visualized the feature distribution for diabetes and non-diabetes in the original PID dataset using box and whisker plots (Fig. S1). The result shows that features like preg, glucose, skinfold, insulin, BMI, pedi, and age (excluding DBP) exhibit significant variation between these two groups, confirmed by their *p*-values. Among these, glucose, BMI, and age are the most critical factors in predicting diabetes. The plot also highlights the presence of outliers among these risk factors, which could potentially affect the model performance. Therefore, it is essential to address these outliers prior to model training.

We employed the interquartile range method to eliminate outliers [31], resulting in a cleaned PID dataset of 648 female participants. Of those, 222 with diabetes, and 462 with non-diabetes. We examined pairwise correlations using a scatterplot to further understand the feature relationships before the modeling (Fig. S2). Specifically, the correlation coefficients (CC) were categorized as follows: strong positive correlations ($CC > 0.5$), moderate positive correlations ($0.3 \leq CC \leq 0.5$), and weak positive correlations ($CC < 0.3$). Strong positive correlations were identified between age and pregnancy, and between BMI and skinfold. Moderate positive correlations were found between insulin and glucose, and between age and DBP. The remaining pairwise comparisons displayed weak correlations. Subsequently, we applied dimension reduction techniques (t-SNE, UMAP, and PCA) to the original features to explore their discriminative capacity between two classes in the cleaned PID dataset (Fig. 2). The substantial overlap between the two classes in the reduced-dimension space indicated that using all eight features together offered limited discrimination. This complex interplay of these features makes accurate diabetes prediction challenging.

3.2. Evaluating different imputation methods on an imbalance dataset

The cleaned PID dataset exhibited invalid zero or missing values in key attributes, including glucose, DBP, skinfold, insulin, and BMI. To address these issues, we implemented three distinct imputation methods: (1) regression model-based imputation, (2) median-based imputation, and (3) mean-based imputation (described in Section 2.2). Each method produced a unique, imbalanced dataset. We then evaluated these datasets using 20 different classifiers, with models trained using a 10-randomized 10-fold CV strategy.

Among the imputation techniques, the regression-based imputation exhibited the poorest performance across all classifiers, with MCC values ranging from 0.174 to 0.496 and ACC values from 0.699 to 0.788 (Fig. 3A). In contrast, the mean-based imputation showed slight improvements, achieving MCC values between 0.157 and 0.526 and ACC values from 0.694 to 0.794 (Fig. 3B). Notably, five classifiers (AB, CB, ERT, RF, and MLP) achieved MCC values of $\geq$0.500 using this approach. The median-based imputation method yielded the best overall results, with MCC values ranging from 0.185 to 0.534 and ACC values from 0.683 to 0.798 (Fig. 3C). Additionally, seven classifiers (MLP, CB, ERT, RF, LGB, SVM-poly, and XGB) demonstrated MCC values of $\geq$0.500 using this method.

Evaluation of three imputation methods revealed that the MLP model consistently performed the best, achieving MCC values of 0.496, 0.526, and 0.534 across all imputation types (Fig. 3). Specifically, the median-based imputation achieved the highest MCC values (0.534), outperforming the best classifiers from the other imputation methods by 0.78–3.78 %. Consequently, we selected the median-based imputation for subsequent analyses. Interestingly, conventional ML models demonstrated superior performance compared to DL models, likely due to the limited dataset size, which may have hindered the effectiveness of DL algorithms. Additionally, all baseline models showed high specificity but low sensitivity, indicating a bias toward the majority class due to the

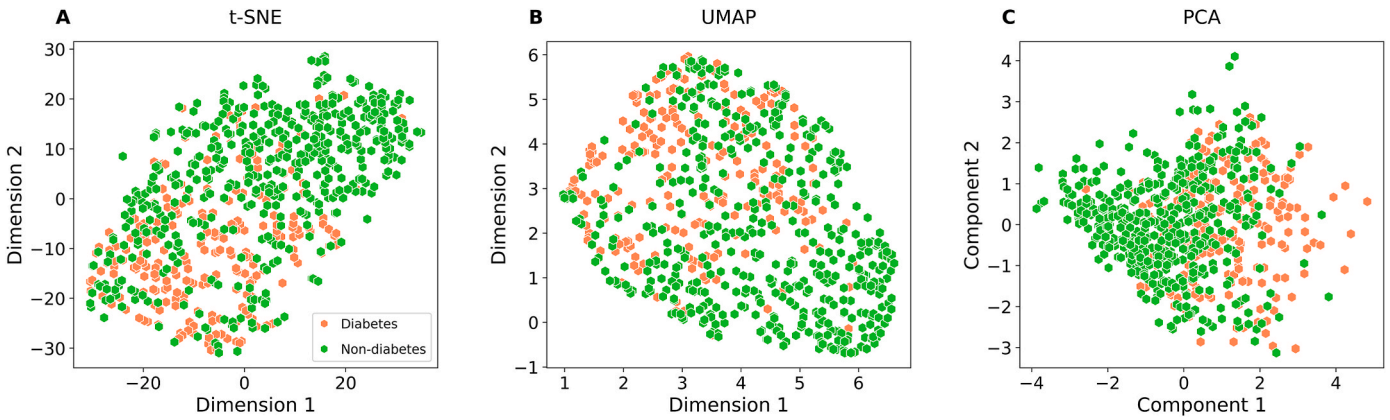

Fig. 2. Visualizing diabetes distribution in 2D space using dimensionality reduction techniques in the clean PID dataset. (A) t-distributed stochastic neighbor embedding (t-SNE). (B) Uniform manifold approximation and projection (UMAP). (C) Principal component analysis (PCA).

imbalanced nature of the dataset. This class imbalance poses a significant challenge for training ML models for accurate diabetes prediction and underscores the need for effective techniques to address this issue.

3.3. Performance of various ML models on balanced datasets

To address the challenges posed by the imbalanced dataset, we employed an under-sampling technique to generate five balanced datasets (BD1–BD5). We then evaluated the performance of 20 different classifiers on each dataset, resulting in 100 baseline models (20 classifiers × 5 datasets) (Fig. 4A–E). To evaluate the overall performance, we computed average results across BD1–BD5 (Fig. 4F). Our results indicate that training on these balanced datasets resulted in improvements in MCC, ACC, and AUC values for traditional ML models, with enhancements ranging from 2.55% to 15.05%, 1.16% to 4.13%, and 0.40% to 7.45%, respectively. Furthermore, Sn increased by 10.71–30.63 % in traditional ML models and 21.14–33.39 % in DL models. Specifically, among DL models, GRU and LSTM showed MCC improvements of 4.49 % and 5.66 %, respectively. However, we observed a decrease in ACC across all DL models, likely due to the reduced size of the training samples.

We further categorized the classifiers into three groups (high, moderate, and low) based on their discriminative abilities: eight classifiers (CB, RF, ERT, GB, XGB, LGB, AB, and SVM-rbf) achieved an excellent performance (high) with MCC >0.600; five classifiers (MLP, SVM-poly, SVM-sigmoid, LR, and SVM-linear) achieved a moderate performance with MCC in the range of 0.500–0.600; and seven classifiers (NB, DNN, BiLSTM, CNN, LSTM, Bi GRU, and GRU) achieved low performance with MCC <0.500). Notably, CB demonstrated strong performance in both scenarios. It ranked second when trained on the imbalanced dataset. However, when trained on BDs, CB achieved the top rank, significantly improving its MCC by 11.92 % and ACC by 2.81 %, compared to the best classifier trained on the imbalanced dataset. Furthermore, within the high group, CB outperformed other classifiers, showing increases in MCC ranging from 0.49 % to 4.78 % and an improvement in ACC ranging from 0.18 % to 2.39 % (Fig. 4).

To comprehend the key factors driving these performance differences, we analyzed the influence of individual features on model performance. As shown in Fig. 2, the eight features in the PID dataset presented challenges in accurately distinguishing between diabetes and non-diabetes cases. To identify the most influential features, we adopted the ensemble FIS approach (detailed in Section 2.5). This analysis revealed that glucose, BMI, age, pedi, and DBP were the most critical features in enhancing model performance. Conversely, preg, insulin levels, and skinfold had minimal impact (Fig. S3).

Based on ensemble FIS, we created five subsets of features: Top3 (glucose, BMI, and age), Top4 (Top3 + pedi), Top5 (Top4 + DBP), Top6 (Top5 + preg), and Top7 (Top6 + skinfold). We then assessed the performance of the top five classifiers (CB, ERT, RF, GB, and XGB) on these subsets. Subsequently, we computed the average performance for each subset across the five balanced datasets (BD1–BD5) and compared it to the control, a model based on all eight features (Fig. S4). The results showed that training CB on the Top3, Top4, Top5, Top6, and Top7 subsets resulted in reductions in MCC by 4.84 %, 5.13 %, 2.29 %, 1.75 %, and 1.27 %, respectively, and in ACC by 2.49 %, 2.53 %, 1.19 %, 0.91 %, and 0.64 %, respectively (Fig. S4). This trend of decreased performance with reduced features was also observed with the other top four classifiers (ERT, RF, GB, and XGB) when trained on these feature subsets (Table S4).

Considering the significance of Top3, Top4, and Top5 feature subsets on model performance, we conducted an ablation analysis to assess the impact of sequentially removing the top contributing features with CB. Specifically, we removed the Top3 features (Rm_Top3), the Top4 features (Rm_Top4), and the Top5 features (Rm_Top5) from each BD. The results show that the removal of these features consistently led to significant declines in model performance. Specifically, we observed MCC decreases of 17.99 %, 23.24 %, and 28.70 %, and ACC decreases of 9.04 %, 11.67 %, and 14.66 % for the Rm_Top3, Rm_Top4, and Rm_Top5 subsets, respectively (Fig. S5). These findings underscore the importance of key risk factors: age, BMI, glucose, pedi, and DBP, which are strongly associated with the likelihood of diabetes and can be considered influential predictive indicators.

3.4. Construction of REMED-T2D and its evaluation on case studies

(i) Evaluation of different computational frameworks

While CB emerges as the optimal baseline model for predicting diabetes and non-diabetes, our objective extended toward developing more advanced prediction strategies. Inspired by the THRONE method [62], we explored several alternative strategies: Ensemble1, Ensemble2, Ensemble3, and Meta-learner strategies (detailed descriptions in Section 2.6). Our initial analysis revealed that the Ensemble3 strategy, trained on a 25D feature set using CB (25D + CB), significantly outperforms other approaches when applied to the eight-feature subset, which includes all diabetes-associated risk factors in the PID dataset. Ensemble3 achieved a 16.66–29.15 % higher MCC and an 8.56–14.51 % higher ACC compared to Ensemble1, Ensemble2, and Meta-learner strategies (Fig. 5C and Table 3).

These promising results required further validation to confirm the model's robustness and generalizability on independent datasets. However, the available public datasets lack all eight features, limiting the ability to evaluate transferability directly. To address this limitation, we developed two additional models using distinct feature subsets: a

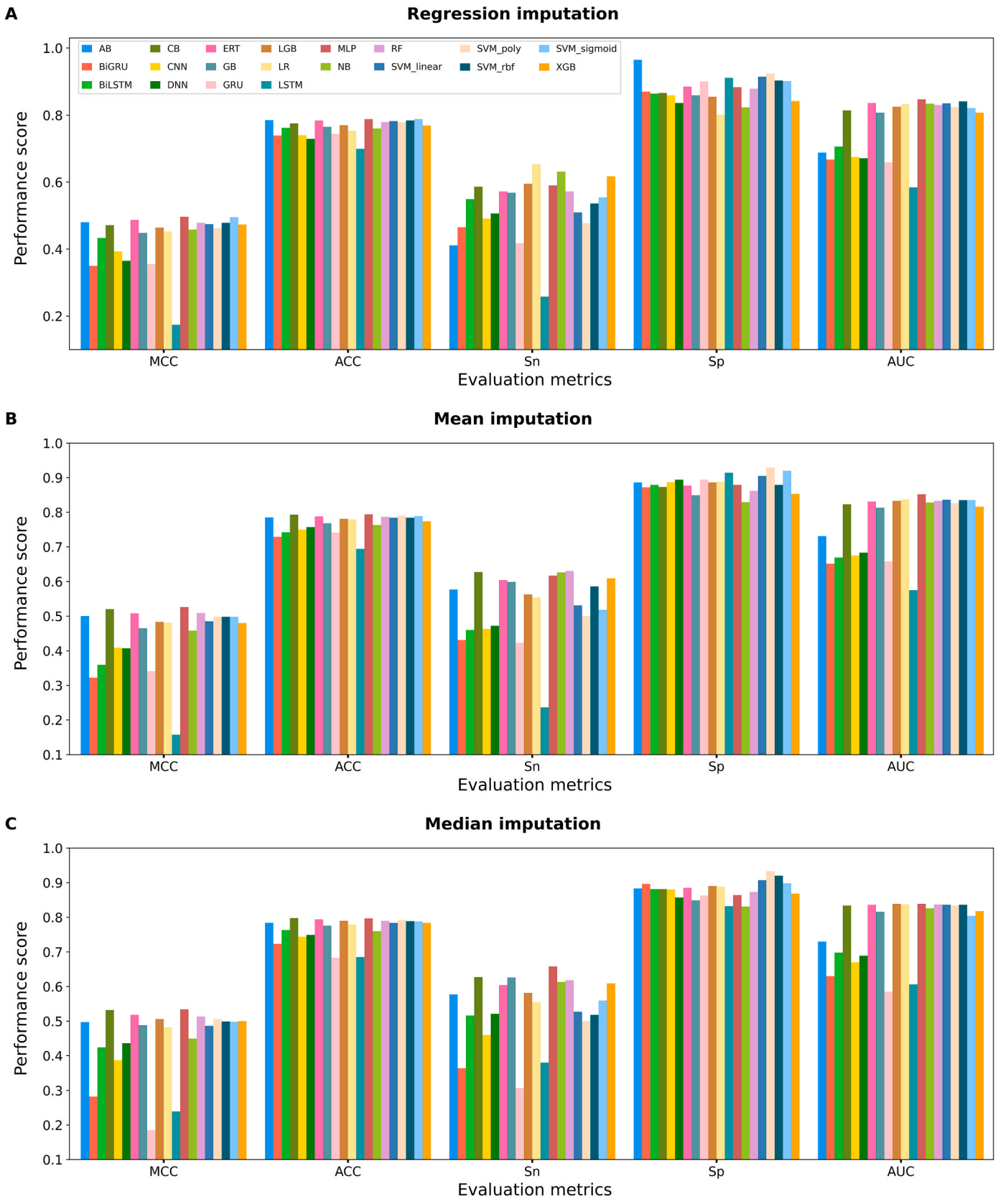

Fig. 3. Performance comparison of 20 classifiers on an imbalanced dataset employing various imputation methods. (A) Median imputation, (B) Mean imputation, and (C) Regression-based imputation. The performance of 20 classifiers was evaluated and represented with various colors. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

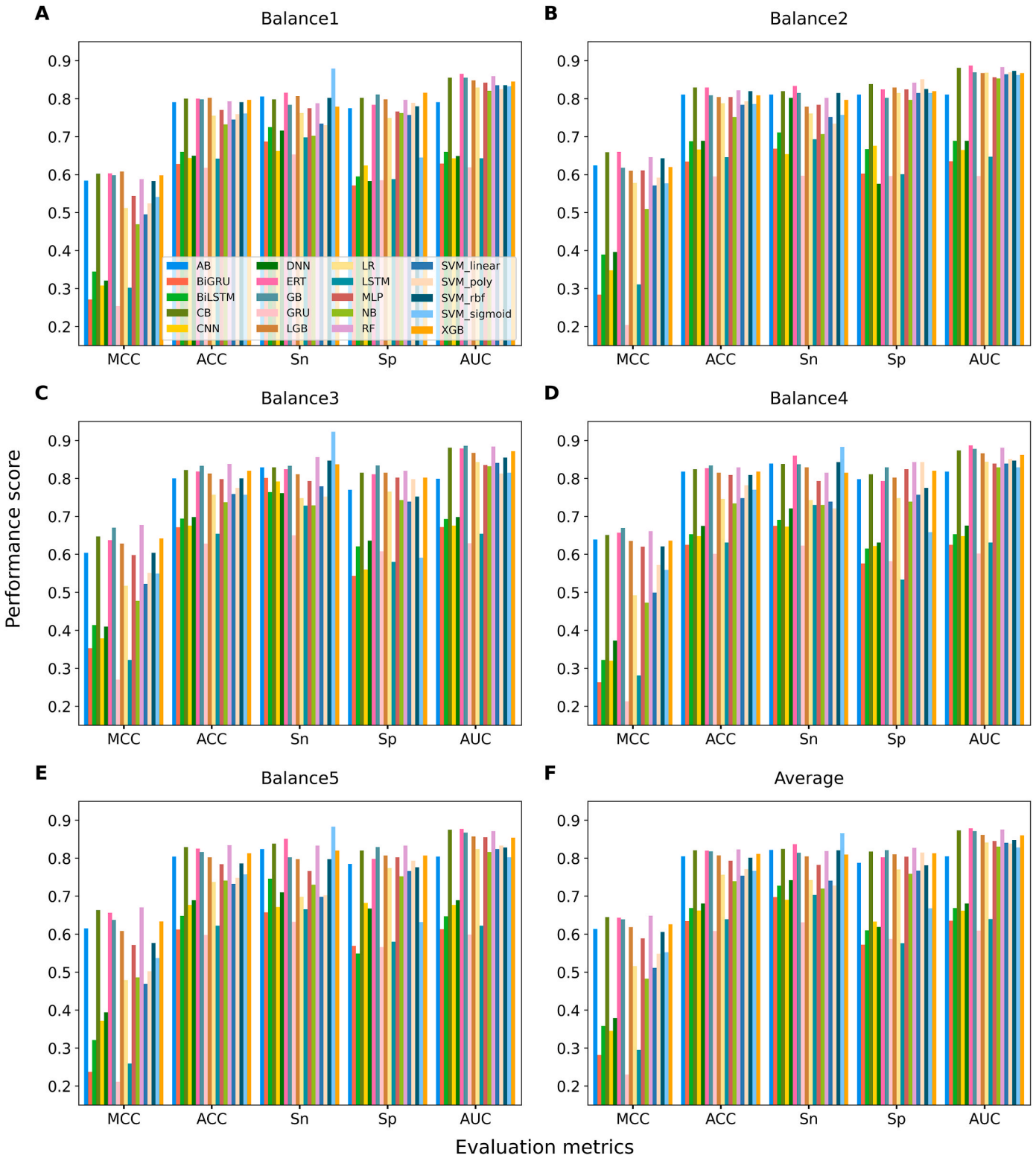

Fig. 4. Performance comparison of 20 different classifiers on five balanced datasets. (A–E) Performance of ML and DL classifiers on Balance1, Balance2, Balance3, Balance4, and Balance5 datasets, respectively. (F) An average performance of each classifier across five balanced datasets.

five-feature subset (incorporating glucose, BMI, age, DBP, and preg) and a seven-feature subset (incorporating five-feature, insulin, and skinfold). We then assessed the performance of 12 ML classifiers (excluding DL classifiers due to their suboptimal performance on the clean PID dataset) on these two subsets. As shown in Figs. S6 and S7, CB consistently outperformed the other classifiers across both datasets, aligning with the findings in Fig. 4F, where CB was identified as the top classifier.

(ii) Selection of REMED-T2D

We applied ensemble strategies to these feature subsets and validated each on independent public datasets (RTML1, RTML2, and

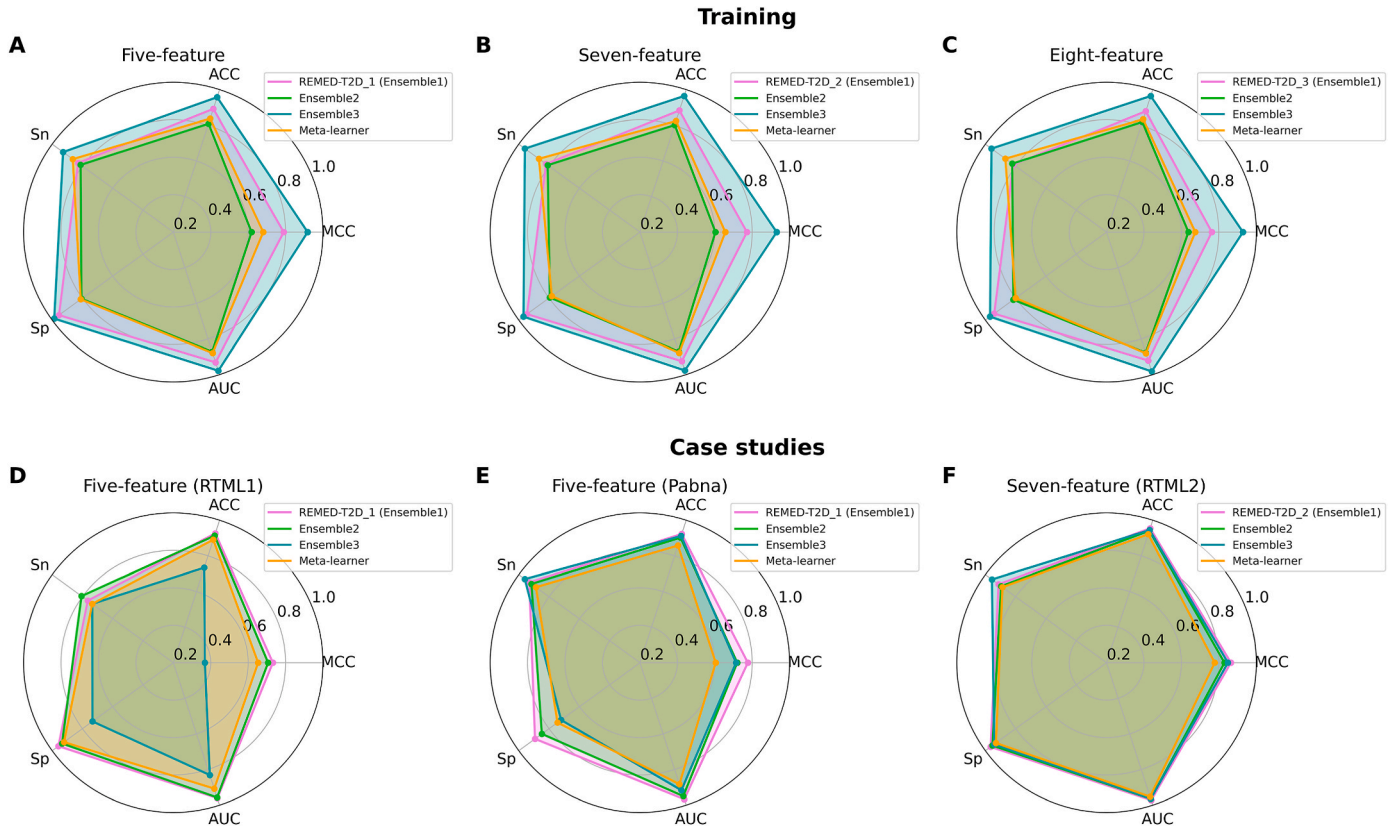

Fig. 5. Exploration of various strategies using different feature subsets on balanced datasets while comparing their performance through case studies. Panels (A–C) show the performance of different methods (Ensemble1, Ensemble2, Ensemble3, and Meta-learner) on the five-feature, seven-feature, and eight-feature subsets, respectively. Panels (D–F) illustrate case study results: (D) the performance of REMED-T2D_1 on the RTML1 dataset, (E) the performance of REMED-T2D_1 on the Pabna dataset, and (F) the performance of REMED-T2D_2 on the RTML2 dataset. The five-feature subset includes glucose, BMI, age, DBP, and preg; the seven-feature subset extends the five-feature subset by adding insulin and skinfold; the eight-feature subset further includes pedi.

Pabna). As shown in Table 3, Ensemble3 consistently achieved the best training performance across the five-feature and seven-feature subsets. Specifically, for the seven-feature subset, Ensemble3 (25D + CB) increased MCC by 15.93–32.75 % and ACC by 8.11–16.22 % on the training set compared to other strategies. Similarly, for the five-feature subset, Ensemble3 (25D + RF) improved MCC by 12.82–30.00 % and ACC by 6.54–14.83 %. However, when validated on public datasets, Ensemble3's performance deteriorated, likely due to overfitting and the inherent complexity of this approach. In contrast, Ensemble1 demonstrated greater stability and robustness across both training and validation sets (Fig. 5). Based on its consistent performance, Ensemble1 was selected as the final model for diabetes prediction, which we named REMED-T2D. Furthermore, REMED-T2D incorporates three distinct models, each trained on a different subset of features: the five-feature model (REMED-T2D_1), the seven-feature model (REMED-T2D_2), and the eight-feature model (REMED-T2D_3). Importantly, this study is the first to validate a model trained on the PID dataset using publicly available datasets related to diabetes.

(iii) Case studies

To assess the generalizability of our model, we validated it on three datasets representing Bangladeshi females (RTML1, RTML2, and Pabna), each with different class distributions. For instance, RTML1 had five times more non-diabetes cases, Pabna had four times more diabetes cases, and RTML2 had over 3.5 times more non-diabetes cases (Fig. S8). Fig. S9 visualizes different distributions of glucose, BMI, age, DBP, and pregnancy status across the PID, RTML1, RTML2, and Pabna datasets, while insulin and skinfold factors are available only for the PID and RTML2 datasets.

Despite imbalances and demographic differences, the PID dataset-trained model demonstrated robust cross-population performance, highlighting its potential for broad applicability across diverse demographic contexts. Specifically, we validated REMED-T2D_1 (the five-feature model) on two independent datasets: RTML1 and the Pabna dataset. REMED-T2D_1 achieved an MCC of 0.790 and an ACC of 0.892 on the training dataset (Fig. 5A), outperforming other methods by 10.94–17.18 % in MCC and 5.43–8.29 % in ACC. On the RTML1 dataset, it attained an MCC of 0.732 and an ACC of 0.926, exceeding other methods by 2.53–7.85 % in MCC and 1.28–3.55 % in ACC (Fig. 5D). Similarly, on the Pabna dataset, REMED-T2D_1 attained an MCC of 0.777 and an ACC of 0.923, outperforming other models by 5.66–17.15 % in MCC and 2.02–6.45 % in ACC (Fig. 5E). We also evaluated REMED-T2D_2 (the seven-feature model). On the training dataset, it achieved an MCC of 0.772 and an ACC of 0.883 (Fig. 5B), outperforming other methods by 11.55–16.82 % in MCC and 5.84–8.11 % in ACC. Validation on the RTML2 dataset further confirmed its efficacy, yielding an MCC of 0.865 and an ACC of 0.954, with improvements of 3.42–8.60 % in MCC and 1.28–3.30 % in ACC compared to other approaches (Fig. 5F). These findings demonstrate that REMED-T2D exhibits strong performance across different datasets with varying class distributions, highlighting its potential for accurate diabetes prediction in various demographic contexts.

(iv) Evaluating REMED-T2D against leading baseline models across training and case studies

REMED-T2D models outperformed the top five baseline models

Table 3
Performance of different ensemble models on datasets with different groups of attributes.

Datasets	Approach	Performance		AUC	MCC	ACC	Sn	Sp
Five-feature	Ensemble1	Aggregation		0.931	0.790	0.892	0.829	0.955
	Ensemble2	Dataset	Balance1	0.858	0.608	0.804	0.797	0.811
			Balance2	0.878	0.613	0.806	0.811	0.802
			Balance3	0.880	0.636	0.818	0.847	0.788
			Balance4	0.880	0.640	0.820	0.815	0.824
			Balance5	0.862	0.595	0.797	0.788	0.806
		Average		0.872	0.618	0.809	0.812	0.806
	Ensemble3	Classifiers	RF	0.978	0.918	0.957	0.928	0.986
			CB	0.979	0.917	0.957	0.946	0.969
			ERT	0.985	0.918	0.957	0.928	0.987
			GB	0.980	0.915	0.955	0.919	0.991
			XGB	0.977	0.905	0.951	0.928	0.973
			LGB	0.982	0.914	0.955	0.955	0.955
			AB	0.948	0.901	0.948	0.919	0.978
	Meta-learner	Classifiers	RF	0.878	0.663	0.829	0.849	0.809
			CB	0.880	0.681	0.838	0.864	0.811
			ERT	0.877	0.658	0.827	0.843	0.809
			GB	0.874	0.655	0.825	0.841	0.808
			XGB	0.868	0.663	0.830	0.849	0.809
			LGB	0.868	0.662	0.829	0.850	0.806
			AB	0.828	0.662	0.829	0.848	0.809
			MLP	0.874	0.655	0.825	0.849	0.801
			SVM	0.858	0.654	0.825	0.849	0.800
			LR	0.874	0.635	0.815	0.826	0.804
Seven-feature	Ensemble1	Aggregation		0.925	0.772	0.883	0.820	0.946
	Ensemble2	Dataset	Balance1	0.863	0.604	0.802	0.811	0.793
			Balance2	0.861	0.599	0.800	0.793	0.806
			Balance3	0.876	0.613	0.806	0.820	0.793
			Balance4	0.886	0.613	0.806	0.811	0.802
			Balance5	0.867	0.590	0.795	0.811	0.779
		Average		0.871	0.604	0.802	0.809	0.795
	Ensemble3	Classifiers	RF	0.982	0.926	0.962	0.942	0.982
			CB	0.977	0.931	0.964	0.959	0.969
			ERT	0.983	0.930	0.964	0.946	0.982
			GB	0.983	0.927	0.962	0.937	0.987
			XGB	0.977	0.909	0.953	0.941	0.964
			LGB	0.987	0.921	0.960	0.955	0.964
			AB	0.960	0.921	0.960	0.942	0.978
	Meta-learner	Classifiers	RF	0.865	0.637	0.816	0.839	0.792
			CB	0.878	0.656	0.825	0.867	0.782
			ERT	0.879	0.643	0.819	0.851	0.786
			GB	0.865	0.621	0.807	0.834	0.780
			XGB	0.870	0.631	0.812	0.850	0.774
			LGB	0.870	0.635	0.814	0.855	0.773
			AB	0.815	0.635	0.815	0.845	0.784
			MLP	0.862	0.622	0.808	0.838	0.779
			SVM	0.868	0.636	0.814	0.864	0.764
			LR	0.871	0.608	0.801	0.809	0.794
Eight-feature	Ensemble1	Aggregation		0.921	0.763	0.878	0.815	0.941
	Ensemble2	Dataset	Balance1	0.860	0.590	0.795	0.806	0.784
			Balance2	0.885	0.658	0.829	0.820	0.838
			Balance3	0.883	0.649	0.824	0.833	0.815
			Balance4	0.883	0.653	0.827	0.824	0.829
			Balance5	0.876	0.640	0.820	0.829	0.811
		Average		0.877	0.638	0.819	0.823	0.815
	Ensemble3	Classifiers	RF	0.990	0.917	0.957	0.942	0.973
			CB	0.982	0.929	0.964	0.960	0.969
			ERT	0.983	0.899	0.948	0.932	0.964
			GB	0.990	0.921	0.959	0.941	0.978
			XGB	0.981	0.916	0.957	0.946	0.969
			LGB	0.990	0.920	0.959	0.955	0.964
			AB	0.955	0.911	0.955	0.946	0.964
	Meta-learner	Classifiers	RF	0.881	0.655	0.825	0.844	0.806
			CB	0.881	0.671	0.832	0.858	0.806
			ERT	0.883	0.662	0.828	0.842	0.815
			GB	0.871	0.640	0.817	0.813	0.821
			XGB	0.885	0.674	0.834	0.866	0.801
			LGB	0.880	0.675	0.834	0.868	0.800
			AB	0.823	0.651	0.823	0.840	0.807
			MLP	0.873	0.653	0.824	0.840	0.808
			SVM	0.858	0.657	0.825	0.849	0.801
			LR	0.875	0.638	0.816	0.821	0.811

across both training and case studies (Fig. S10). During the training phase, REMED-T2D_1 surpassed CB, ERT, RF, MLP, and AB by 15.83–17.60 % in MCC and 7.78–8.63 % in ACC (Fig. S10A). REMED-T2D_2 achieved even higher performance, exceeding CB, ERT, RF, GB, and AB with increases in MCC by 13.51–15.62 % and ACC by 6.66–7.69 % (Fig. S10B). Similarly, REMED-T2D_3 demonstrated improvements, with MCC rising by 10.95–13.72 % and ACC by 5.35–6.71 % over CB, ERT, RF, GB, and XGB (Fig. S10C).

During the case studies validation phase, REMED-T2D_1 achieved an MCC of 0.732, ACC of 0.926, Sn of 0.765, Sp of 0.959, and AUC of 0.961 on the RTML1 dataset (Fig. S10D), improving ACC and MCC over the leading baseline models by 3.30–10.10 % and 1.11–4.51 %, respectively. On the Pabna dataset, REMED-T2D_1 achieved an MCC of 0.777, ACC of 0.923, Sn of 0.930, Sp of 0.893, and AUC of 0.967 (Fig. S10E), surpassing the top baseline models by 5.17–17.47 % in MCC and by 1.06–7.66 % in ACC. Meanwhile, REMED-T2D_2 showed remarkable performance on the RTML2 dataset (Fig. S10F), with an ACC of 0.954, MCC of 0.865, Sn of 0.913, Sp of 0.965, and AUC of 0.972. This model improved MCC and ACC by 0.42–6.58 % and 0.18–2.39 % over these five baseline models. However, due to limitations in feature availability, a comparison between the eight-feature-based model and the top baseline models was not feasible. Overall, REMED-T2D demonstrated robust performance, strong convergence, and impressive generalization capabilities across both training and case studies.

3.5. Comparison of REMED-T2D with existing predictors on the training dataset

We compared our eight-feature model with the state-of-the-art predictors on the PID dataset [25–28]. To ensure a fair comparison, all methods, including the current approach, used the PID dataset for training, employed a 10-fold CV, and addressed dataset imbalances. As shown in Table 4, REMED-T2D_3 outperformed existing methods in terms of ACC and MCC. In particular, our model achieved significantly higher ACC (1.40–4.60 %) and MCC (3.50–9.80 %) values compared to the existing approaches. However, we could not evaluate our model against the existing methods in case studies, due to the lack of publicly available source code or web servers for the existing methods.

3.6. Web server development

We developed REMED-T2D, an online platform designed for quick and convenient diabetes prediction to support healthcare professionals in making prompt, informed decisions. The web server is accessible at https://balalab-skku.org/REMED-T2D/ and allows users to submit jobs by either inputting information directly or uploading a CSV file. The web server was constructed using Python, Django, HTML, CSS, and JavaScript, incorporating a PostgreSQL database for efficient storage and retrieval of job outcomes. Users can choose from three prediction options to suit their needs.

(1) REMED-T2D_1: Patient_ID, Pregnancy_times, Glucose, Diastolic_blood_pressure, Body_mass_index, Age.
(2) REMED-T2D_2: Patient_ID, Pregnancy_times, Glucose, Diastolic_blood_pressure, Skin_thickness, Insulin_level, Body_mass_index, Age.
(3) REMED-T2D_3: Patient_ID, Pregnancy_times, Glucose, Diastolic_blood_pressure, Skin_thickness, Insulin_level, Body_mass_index, Diabetes_pedigree_function, Age.

A detailed help page (https://balalab-skku.org/REMED-T2D/help/) provides comprehensive usage guidelines and dataset download links. After submission, users can view the prediction results directly on the website. If an email is provided, the results will also be sent via email.

4. Discussion

Diabetes has emerged as a global epidemic and represents a major public health challenge [3,63]. In response, AI-based models have become valuable tools for early diabetes diagnosis and management, enabling timely intervention to prevent complications [64–67]. Despite this potential, there are significant challenges in enhancing the accuracy of diabetes prediction models, particularly when addressing several issues inherent to the PID dataset: (i) missing data; (ii) class imbalance; and (iii) the overall performance of the classifier. To address these challenges, we implemented a comprehensive preprocessing strategy. First, we pre-processed the data to handle missing values and ensure data quality, exploring various techniques to determine the optimal approach. Next, we employed an under-sampling technique to mitigate class imbalance and potential model bias, creating five different balanced datasets (BD1–BD5) with equal representation of positive and negative diabetes cases. Finally, we conducted a comprehensive evaluation of multiple ML models, encompassing both conventional ML algorithms and DL approaches. Using a rigorous 10-randomized 10-fold CV strategy with extensive hyperparameter optimization, we compared model performance across all balanced datasets. Our analysis revealed that the CB algorithm consistently outperformed other models. Furthermore, we identified five key risk factors as the most significant predictors of diabetes: glucose, BMI, age, pedigree function, and DBP.

Through comprehensive exploration of diverse modeling strategies, we ultimately adopted an ensemble-based aggregation approach that demonstrated robust and consistent performance across both our training dataset and independent case studies. The improved performance is mainly attributed to the ensembled approach combined with the under-sampling approach for handling class imbalance [25–28]. Notably, our study represents the validation of a PID-based diabetes prediction model in an Asian population, achieving high predictive accuracy when tested on external datasets of Bangladeshi females, demonstrating its potential for a broader demographic application context. To facilitate real-world application and ensure widespread accessibility, we developed the REMED-T2D web server (https://balalab-skku.org/REMED-T2D/), a user-friendly web tool enabling remote users to utilize our predictive model for early diabetes detection. Further, the reliable and robust approach of REMED-T2D can be easily adaptable to other structured datasets, with potential applications in areas like heart disease [68], breast cancer [69], lung cancer [66], and

Table 4
Comparative performance analysis of our proposed model and the existing predictors on the PID dataset. We compared the performance of our proposed models to recently published methods.

Study	Year, classifiers	Preprocessing approach	Imbalanced treatment	Performance metrics					Evaluation strategy
				ACC	MCC	Sn	Sp	AUC	
Wang et al. [28]	RF, 2019	NB method	ADASYN	0.862	0.709	0.857	0.865	0.926	10-fold CV
Ramesh et al. [27]	SVM, 2021	Multivariate imputation	SMOTE	0.832	0.665	0.872	0.790	NA	10-fold CV
Hairani et al. [26]	RF, 2022	NA	SMOTE-Tomek Link	0.864	0.728	0.882	0.849	NA	10-fold CV
Reza et al. [25]	SVM, 2023	Median imputation	SMOTE	0.855	0.711	0.870	0.841	0.855	10-fold CV
REMED-T2D_3	Aggregation, 2024	Median imputation	Under-sampling technique	0.878	0.763	0.815	0.941	0.921	Randomized 10-fold CV

SMOTE, synthetic minority oversampling technique; NA, not available.

hydrogen production from microbial electrolysis cells [70]. It is worth noting that REMED-T2D was trained on females so the model should not be used to predict diabetes in males. However, the methodology used in this study can apply to a dataset that includes both male and female samples.

Despite REMED-T2D demonstrating impressive performance in predicting diabetes, there is room for improvement. The model was trained on a relatively smaller dataset (444 samples for each BD), and missing features posed a challenge to model robustness. Future studies should prioritize the acquisition of larger, more comprehensive patient datasets to further enhance model performance. Exploring the application of both conventional ML and DL algorithms on these expanded datasets could yield valuable improvements in predictive accuracy and precision.

CRediT authorship contribution statement

Le Thi Phan: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation. **Rajan Rakkiyappan:** Supervision, Validation, Writing – review & editing. **Balachandran Manavalan:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Data availability statement

The web server can be accessed via https://balalab-skku.org/REMED-T2D/ and all the processed data used in this study can be downloaded from the web server.

Ethics in publishing statement

This research presents an accurate account of the work performed, all data presented are accurate and methodologies are detailed enough to permit others to replicate the work.

This manuscript represents entirely original works and if work and/or words of others have been used, this has been appropriately cited or quoted and permission has been obtained where necessary.

This material has not been published in whole or in part elsewhere.

The manuscript is not currently being considered for publication in another journal.

That generative AI and AI-assisted technologies have not been used to create or alter images unless specifically used as part of the research design where such use must be described in a reproducible manner in the methods section.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors utilized ChatGPT 4.0 to enhance sentence fluency and improve readability. After using this tool, the authors carefully reviewed and edited the content as necessary. The authors take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2021R1A2C1014338 and RS-2024-00344752). This research was also supported by the Department of Integrative Biotechnology, Sungkyunkwan University (SKKU), and the BK21 FOUR Project. The authors thank the Computational Biology and Bioinformatics Laboratory's members for their valuable discussion.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.109771.

References

[1] Y. Zheng, S.H. Ley, F.B. Hu, Global aetiology and epidemiology of type 2 diabetes mellitus and its complications, Nat. Rev. Endocrinol. 14 (2018) 88–98.
[2] P. Aschner, New IDF clinical practice recommendations for managing type 2 diabetes in primary care, Diabetes Res. Clin. Pract. 132 (2017) 169–170.
[3] I. Federation, International diabetes federation. IDF Diabetes Atlas, tenth ed., 2021. *Brussels, Belgium*, https://www.diabetesatlas.org.
[4] L. Perreault, J.S. Skyler, J. Rosenstock, Novel therapies with precision mechanisms for type 2 diabetes mellitus, Nat. Rev. Endocrinol. 17 (2021) 364–377.
[5] A.D. Assoc, Diagnosis and classification of diabetes mellitus, Diabetes Care 36 (2013) S67–S74.
[6] R.A. DeFronzo, E. Ferrannini, L. Groop, R.R. Henry, et al., Type 2 diabetes mellitus, Nat. Rev. Dis. Prim. 1 (2015).
[7] B. Zhou, Y. Lu, K. Hajifathalian, J. Bentham, et al., Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants, Lancet 387 (2016) 1513–1530.
[8] N. Holman, B. Young, R. Gadsby, Current prevalence of Type 1 and Type 2 diabetes in adults and children in the UK, Diabet. Med. 32 (2015) 1119–1120.
[9] G. Bruno, C. Runzo, P. Cavallo-Perin, F. Merletti, et al., Incidence of type 1 and type 2 diabetes in adults aged 30-49 years, Diabetes Care 28 (2005) 2613–2619.
[10] S. Park, D. Choi, M. Kim, W. Cha, et al., Identifying prescription patterns with a topic model of diseases and medications, J. Biomed. Inf. 75 (2017) 35–47.
[11] W.H. Herman, W. Ye, S.J. Griffin, R.K. Simmons, et al., Early detection and treatment of type 2 diabetes reduce cardiovascular morbidity and mortality: a simulation of the results of the anglo-Danish-Dutch study of intensive treatment in people with screen-detected diabetes in primary care (ADDITION-Europe), Diabetes Care 38 (2015) 1449–1455.
[12] K. Ogurtsova, L. Guariguata, N.C. Barengo, P.L. Ruiz, et al., IDF diabetes Atlas: global estimates of undiagnosed diabetes in adults for 2021, Diabetes Res. Clin. Pract. 183 (2022) 109118.
[13] U.M. Butt, S. Letchmunan, M. Ali, F.H. Hassan, et al., Machine learning based diabetes classification and prediction for healthcare applications, J Healthc Eng 2021 (2021) 9930985.
[14] O. Rolandsson, M. Norberg, L. Nystrom, S. Soderberg, et al., How to diagnose and classify diabetes in primary health care: lessons learned from the Diabetes Register in Northern Sweden (DiabNorth), Scand. J. Prim. Health Care 30 (2012) 81–87.
[15] R.B. Balaban, A physician's guide to talking about end-of-life care, J. Gen. Intern. Med. 15 (2000) 195–200.
[16] K. Roy, M. Ahmad, K. Waqar, K. Priyaah, et al., An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values, Complexity 2021 (2021).
[17] R.G. Nelson, W.C. Knowler, D.J. Pettitt, M.F. Saad, P.H. Bennett, Diabetic kidney disease in Pima Indians, Diabetes Care 16 (1993) 335–341.
[18] S. Lillioja, Impaired glucose tolerance in Pima Indians, Diabet. Med. 13 (1996) 127–132.
[19] V. Kulshrestha, N. Agarwal, Maternal complications in pregnancy with diabetes, JPMA. The Journal of the Pakistan Medical Association 66 (2016) S74–S77.
[20] U. Schaefer-Graf, A. Napoli, C.J. Nolan, D.P.S. Group, Diabetes in pregnancy: a new decade of challenges ahead, Diabetologia 61 (2018) 1012–1021.
[21] V. Chang, J. Bailey, Q.A. Xu, Z.L. Sun, Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms, Neural Computing & Applications, 2022.
[22] S.K. Kalagotla, S.V. Gangashetty, K. Giridhar, A novel stacking technique for prediction of diabetes, Comput. Biol. Med. 135 (2021).
[23] D.K. Saloni Kumari, Mamta Mittal an ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, International Journal of Cognitive Computing in Engineering 2 (2021) 40–46.
[24] I. Tasin, T.U. Nabil, S. Islam, R. Khan, Diabetes prediction using machine learning and explainable AI techniques, Healthc Technol Lett 10 (2023) 1–10.
[25] M.S. Reza, U. Hafsha, R. Amin, R. Yasmin, S. Ruhi, Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: insights from the PIMA dataset, Computer Methods and Programs in Biomedicine Update 4 (2023) 100118.
[26] H. Hairani, A. Anggrawan, D. Priyanto, Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link, JOIV: international journal on informatics visualization 7 (2023) 258–264.
[27] J. Ramesh, R. Aburukba, A. Sagahyroon, A remote healthcare monitoring framework for diabetes prediction using machine learning, Healthc Technol Lett 8 (2021) 45–57.
[28] Q. Wang, W.J. Cao, J.W. Guo, J.D. Ren, et al., DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values, IEEE Access 7 (2019) 102232–102238.
[29] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, 1998.

[30] M.S. Reza, R. Amin, R. Yasmin, W. Kulsum, S. Ruhi, Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data, Heliyon 10 (2024) e24536.
[31] C.S.K. Dash, A.K. Behera, S. Dehuri, A. Ghosh, An outliers detection and elimination framework in classification task of data mining, Decision Analytics Journal 6 (2023) 100164.
[32] X. Zhang, L. Wei, X. Ye, K. Zhang, et al., SiameseCPP: a sequence-based Siamese network to predict cell-penetrating peptides by contrastive learning, Briefings Bioinf. 24 (2023).
[33] S. Basith, B. Manavalan, T.H. Shin, G. Lee, SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome, Mol. Ther. Nucleic Acids 18 (2019) 131–141.
[34] X. Wang, C. Li, F. Li, V.S. Sharma, et al., SIMLIN: a bioinformatics tool for prediction of S-sulphenylation in the human proteome based on multi-stage ensemble-learning models, BMC Bioinf. 20 (2019) 602.
[35] M.M. Hasan, S. Basith, M.S. Khatun, G. Lee, et al., Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework, Briefings Bioinf. 22 (2021).
[36] I. Rish, IJCAI 2001 workshop on empirical methods in artificial intelligence, Cités (2001) 41–46.
[37] Z. Abbas, H. Tayara, K.T. Chong, Alzheimer's disease prediction based on continuous feature representation using multi-omics data integration, Chemometr Intell Lab 223 (2022).
[38] J. Brownlee, Machine Learning Algorithms from Scratch, 2021.
[39] R. Saxena, S.K. Sharma, M. Gupta, G.C. Sampada, A novel approach for feature selection and classification of diabetes mellitus: machine learning methods, Comput. Intell. Neurosci. 2022 (2022) 3820360.
[40] Y.J. Jeon, M.M. Hasan, H.W. Park, K.W. Lee, B. Manavalan, TACOS: a novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization, Briefings Bioinf. 23 (2022).
[41] S. Basith, G. Lee, B. Manavalan, STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction, Briefings Bioinf. 23 (2022).
[42] P. Charoenkwan, C. Nantasenamat, M.M. Hasan, M.A. Moni, et al., StackDPPIV: a novel computational approach for accurate prediction of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides, Methods 204 (2022) 189–198.
[43] M.M. Hasan, M.A. Alam, W. Shoombuatong, H.W. Deng, et al., NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning, Briefings Bioinf. 22 (2021).
[44] R. Xie, J. Li, J. Wang, W. Dai, et al., DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy, Briefings Bioinf. 22 (2021).
[45] S. Basith, M.M. Hasan, G. Lee, L. Wei, B. Manavalan, Integrative machine learning framework for the identification of cell-specific enhancers from the human genome, Briefings Bioinf. 22 (2021).
[46] G. Swapna, R. Vinayakumar, K.P. Soman, Diabetes detection using deep learning algorithms, Ict Express 4 (2018) 243–246.
[47] P.N. Srinivasu, J. Shafi, T.B. Krishna, C.N. Sujatha, et al., Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data, Diagnostics 12 (2022).
[48] G. Geetha, K.M. Prasad, Stacking ensemble learning-based convolutional gated recurrent neural network for diabetes miletus, Intelligent Automation and Soft Computing 36 (2023) 703–718.
[49] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, et al., An optimization-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment, Appl Sci-Basel 12 (2022).
[50] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[51] H. Naz, S. Ahuja, Deep learning approach for diabetes prediction using PIMA Indian dataset, J. Diabetes Metab. Disord. 19 (2020) 391–403.
[52] B. Manavalan, T.H. Shin, G. Lee, PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine, Front. Microbiol. 9 (2018) 476.
[53] B. Manavalan, T.H. Shin, G. Lee, DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest, Oncotarget 9 (2018) 1944–1956.
[54] P. Charoenkwan, W. Chiangjong, C. Nantasenamat, M.M. Hasan, et al., StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides, Briefings Bioinf. 22 (2021).
[55] H. Yan, L.Y. Fu, Y. Qi, D.J. Yu, Q.L. Ye, Robust ensemble method for short-term traffic flow prediction, Future Generat. Comput. Syst. 133 (2022) 395–410.
[56] M. Ullah, F. Hadi, J. Song, D.J. Yu, PScL-DDCFPred: an ensemble deep learning-based approach for characterizing multiclass subcellular localization of human proteins from bioimage data, Bioinformatics 38 (2022) 4019–4026.
[57] Q. Wu, Z. Deng, X. Pan, H.B. Shen, et al., MDGF-MCEC: a multi-view dual attention embedding model with cooperative ensemble learning for CircRNA-disease association prediction, Briefings Bioinf. 23 (2022).
[58] R. Su, J. Hu, Q. Zou, B. Manavalan, L.Y. Wei, Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools, Briefings Bioinf. 21 (2020) 408–420.
[59] B. Manavalan, M.M. Hasan, S. Basith, V. Gosu, et al., Empirical comparison and analysis of web-based DNA N-4-Methylcytosine site prediction tools, Mol. Ther. Nucleic Acids 22 (2020) 406–420.
[60] S. Basith, B. Manavalan, T. Hwan Shin, G. Lee, Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening, Med. Res. Rev. 40 (2020) 1276–1314.
[61] B. Manavalan, S. Basith, T.H. Shin, G. Lee, Computational prediction of species-specific yeast DNA replication origin via iterative feature representation, Briefings Bioinf. 22 (2021).
[62] W. Shoombuatong, S. Basith, T. Pitti, G. Lee, B. Manavalan, THRONE: a new approach for accurate prediction of human rna N7-methylguanosine sites, J. Mol. Biol. 434 (2022) 167549.
[63] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, et al., IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045, Diabetes Res. Clin. Pract. 183 (2022) 109119.
[64] F. Mohsen, H.R.H. Al-Absi, N.A. Yousri, N. El Hajj, Z.B. Shah, A scoping review of artificial intelligence-based methods for diabetes risk prediction, Npj Digit Med 6 (2023).
[65] S.C.Y. Wang, G. Nickel, K.P. Venkatesh, M.M. Raza, J.C. Kvedar, AI-based diabetes care: risk prediction models and implementation concerns, Npj Digit Med 7 (2024).
[66] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, M.Z. Ghobadi, Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review, Diabetol. Metab. Syndrome 14 (2022).
[67] M. Ravaut, H. Sadeghi, K.K. Leung, M. Volkovs, et al., Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data, Npj Digit Med 4 (2021).
[68] A.U. Rahman, Y. Alsenani, A. Zafar, K. Ullah, et al., Enhancing heart disease prediction using a self-attention-based transformer model, Sci. Rep. 14 (2024) 514.
[69] H.A. Essa, E. Ismaiel, M.F.A. Hinnawi, Feature-based detection of breast cancer using convolutional neural network and feature engineering, Sci. Rep. 14 (2024) 22215.
[70] J. Yoon, D.-Y. Cheong, G. Baek, Predicting current and hydrogen productions from microbial electrolysis cells using random forest model, Appl. Energy 371 (2024) 123641.