Array 31 (2026) 101031

Contents lists available at ScienceDirect

Array

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

Leveraging machine learning for personalized type 2 diabetes prediction: A comparative analysis☆

B.P. Pradeep Kumar [a], Pramod K.B. Rangaiah [b], Robin Augustine [b,*]

[a] Department of Computer Science and Design, Atria Institute of Technology, Bengaluru 560024, India
[b] Microwaves in Medical Engineering Group, Division of Integrated Smart Systems Technology, Department of Electrical Engineering, Uppsala University, Box 65 SE-751 03 Uppsala, Sweden

ARTICLE INFO

Keywords:
Diabetes prediction
Machine learning
Healthcare analytics
Predictive modeling
Risk stratification
Data-driven healthcare
Algorithmic approaches

ABSTRACT

This study presents a comparative machine learning (ML) benchmarking framework for the early prediction and risk stratification of **Type 2 diabetes (T2D)** in adult patients. The work is positioned as a comparative feasibility study rather than as a deployable clinical tool. Using the Pima Indian Diabetes Dataset (768 records of female subjects aged 21 years and older, 9 clinical features), we comparatively evaluate nine ML classifiers Support Vector Machine (SVM), CatBoost, Voting Classifier, K-Nearest Neighbors (KNN), XGBoost, Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), and Decision Tree (DT) across five train–test splits (5%–50%) and four k-fold cross-validation settings ($k = 2, 4, 6, 8$). Model evaluation reports multiple complementary metrics (accuracy, precision, recall, F1-score, ROC–AUC, and confusion-matrix-based error counts) together with cross-validation standard deviations interpreted as approximate uncertainty bounds, providing a more reliable picture than accuracy alone. Among the evaluated models, Random Forest provided the most consistent overall performance, with a maximum accuracy of 79.3% and the largest ROC–AUC (0.83); SVM remained stable in the 72.5%–77.2% range; and the Voting Classifier produced balanced but non-superior aggregate performance. We explicitly acknowledge several scope limitations: (i) the dataset is single-source and demographically homogeneous (adult Pima Indian females), so the findings are not directly generalizable to children, adolescents, men, or other ethnic groups; (ii) the analysis is internal cross-validation only, without external or prospective multi-center validation; (iii) explainability techniques (SHAP, LIME), calibration, decision-curve analysis, and advanced class-imbalance handling are identified as future research directions and are not part of the current scope. Within these boundaries, the study contributes a reproducible, multi-metric methodological template for comparative T2D risk-prediction studies, and explicitly identifies which methodological gaps the subsequent research should address.

1. Introduction

This study addresses the early prediction of Type 2 diabetes, a chronic metabolic condition for which improved screening and risk assessment tools are needed [1]. Because Type 2 diabetes represents a substantial public health burden [2], machine learning (ML) techniques are evaluated for their ability to support early identification of at-risk individuals. Conventional risk-assessment approaches do not consistently capture the multifactorial nature of disease onset, which can lead to delayed intervention and adverse patient outcomes [3]. To address this gap, nine classifiers are systematically evaluated [4]: Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, CatBoost, XGBoost, Decision Tree, and a Voting Classifier. Their performance is analyzed to identify the most effective models and the most informative features for accurate prediction. Demographic, lifestyle, clinical, and family-history attributes are incorporated to construct reliable predictive models [5]. Cross-validation is used to characterize each algorithm's stability and to inform model selection for diabetes prediction. The findings are

☆ This work was supported in part by the following projects: "Dielectric Imaging Advancement through Microwave Probe System (DIAMPS)", Ref. no. 2024-02039, Eurostars 3; "BOS: Principles and techniques for body-centered operative systems", the Swedish Foundation for Strategic Research (SSF) grant FUS21-0067. Swedish SSF Projects ZeroIoT (CH10-0003); Wireless Brain-Connect inteRfAce TO machineS: B-CRATOS, European Union's Horizon 2020 Research and Innovation Program, Grant agreement ID: 965044, DOI. 10.3030/965044, Funded under EXCELLENT SCIENCE - Future and Emerging Technologies (FET) Open, Coordinated by UPPSALA UNIVERSITET, Sweden. Start date 1 March 2021 and End date 28 February 2025.

* Corresponding author.
E-mail address: robin.augustine@angstrom.uu.se (R. Augustine).

https://doi.org/10.1016/j.array.2026.101031
Received 16 April 2025; Received in revised form 18 June 2026; Accepted 19 June 2026
Available online 3 July 2026
2590-0056/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

intended to support early identification and preventive care, enabling clinicians to tailor interventions, reduce complications, and improve patient outcomes [6]. The study contributes to the development of reproducible, clinically aligned ML methodology for personalized Type 2 diabetes risk assessment.

The workflow then employs a Test-Train Split, ranging from 5% to 50%, to partition the data into training and testing subsets, allowing evaluation across varied split ratios. This is complemented by k-fold cross-validation ($k = 2, 4, 6, 8$), which splits the dataset into k groups, training on $k - 1$ subsets and testing on the remaining one, ensuring thorough validation [7]. Model Training follows, utilizing a suite of algorithms Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, Gradient Boosting, XGBoost, CatBoost, and Decision Tree. A Voting Classifier then integrates outputs from these models, enhancing prediction strength by combining their unique advantages.

Model Evaluation is a critical next step, where performance is gauged using metrics such as accuracy, precision, recall, and F_1-score, aiding in the identification of top-performing models. This is followed by Hyperparameter Tuning, employing methods like Grid Search or Random Search to fine-tune parameters and optimize outcomes [8].

The process concludes with Accuracy Comparison, which evaluates model effectiveness through key metrics [9], and Accuracy Visualization, using visuals like bar charts or line graphs to present results clearly, supporting informed model choices [10]. This structured approach highlights the importance of each stage, reinforcing the development of a dependable machine learning system for tailored Type 2 diabetes risk assessment [11].

1.1. Limitations of existing ML and ensemble methods

Despite significant advances in machine learning (ML) and ensemble-based classification techniques, their application to the early prediction of Type 2 diabetes remains suboptimal. Most existing studies have been trained on adult-centric datasets or aggregated population-level data, thereby limiting generalizability to other populations such as younger adults, men, and other ethnic groups. Furthermore, ensemble methods, although powerful in reducing overfitting and improving prediction accuracy, often lack interpretability a key requirement in clinical settings. Additionally, current ML pipelines rarely address challenges such as data imbalance, noise, and demographic skew, which are commonly encountered in healthcare datasets. Finally, very few studies provide a holistic benchmarking of various ML models on standardized datasets with controlled train–test splits and cross-validation, leaving clinicians and researchers with limited guidance on optimal model selection for Type 2 diabetes prediction.

1.2. Motivation for this research

The global rise in Type 2 diabetes (T2D), particularly among adults from high-risk populations, highlights an urgent need for robust, data-driven tools for early screening and risk stratification. Existing clinical diagnostic protocols are reactive rather than predictive and often depend on periodic health check-ups or symptomatic presentation. In this context, we are motivated to develop an interpretable and reproducible ML framework that enables timely identification of high-risk individuals, especially in low-resource or underserved settings. This study aims to fill a critical gap by systematically evaluating the performance of both classical and ensemble ML models in the context of Type 2 diabetes.

1.3. Novelty of the proposed work

This study distinguishes itself by performing a multi-model, cross-validated comparative analysis of nine machine learning algorithms including Random Forest, Support Vector Machines, XGBoost, CatBoost, Voting Classifier, Gradient Boosting, Logistic Regression, K-Nearest Neighbors, and Decision Trees on a standardized diabetes dataset. The proposed pipeline integrates preprocessing, feature engineering, hyperparameter tuning, and multi-fold validation to ensure robust and generalizable results. Unlike prior work focused narrowly on accuracy, our framework evaluates models across multiple metrics (precision, recall, F1-score, confusion matrix) under various data splits, making it clinically comprehensive and scalable. Moreover, the ensemble configurations used in this study are shown to outperform traditional classifiers in terms of both stability and interpretability.

1.4. Significance and clinical implications

The proposed ensemble-based ML framework is presented as a candidate tool for early diagnosis and risk prediction of Type 2 diabetes. Its modular design and generalization capability indicate suitability for deployment in primary-care and screening workflows. By providing interpretable insights (e.g., feature importance scores) and quantifying model confidence through cross-validation, the system can support clinical decision-making in community health programs and digital health monitoring platforms. The current study is limited to a homogeneous dataset (female Pima Indian subjects); the methodological approach is, however, extendable to multi-source, gender-balanced, and geographically diverse populations in future work.

2. Related works

Research on Type 2 diabetes prediction has employed a range of techniques aimed at supporting timely detection and intervention. Rastogi and Bansal applied data mining methods to forecast the onset of Type 2 diabetes by identifying individual risk factors [12]. Hasan et al. used ensemble learning, combining multiple classifiers to improve prediction accuracy [13]. To address class imbalance, Alex et al. proposed a deep learning framework with the Synthetic Minority Over-sampling Technique (SMOTE), incorporating temporal patterns in health data [14]. Surveys of current methods, supported by case studies, have summarized the practical applicability of ML techniques in clinical settings [15]. Comparative analyses of ML algorithms have further documented their respective advantages and limitations, providing guidance on model selection [16].

In clinical contexts, Shin et al. refined predictive tools to delay or prevent Type 2 diabetes onset, with an emphasis on early intervention [17]. Gupta et al. compared quantum machine learning with deep learning approaches for diabetes prediction [18]. Taha and Malebary proposed a hybrid model combining fuzzy clustering with logistic regression to improve predictive precision [19]. Dutta et al. demonstrated that ensemble methods can yield substantial performance gains in early detection [20], while Khan et al. integrated bio-inspired optimization with neural networks to develop an automated prediction system [21].

Cardiovascular complications of diabetes have been examined in models that incorporate heart-related variables to characterize the diabetes–cardiovascular interplay [34]. Subsequent work investigated the model characteristics that enable success in precision diabetes care, particularly for cardiovascular risk [35,36]. A systematic review by Tan et al. evaluated prognostic ML models for diabetes complications and identified their respective strengths and limitations [37].

The role of early prediction in improving patient care and healthcare-system efficiency has been emphasized in subsequent studies [12]. Febrian et al. examined dataset-specific factors to refine targeted interventions [38]. Whig et al. used the PyCaret library to streamline ML pipelines for healthcare applications [39]. Tasin et al.

Table 1
Timeline of machine learning approaches for diabetes prediction.

Ref	Year	Methodology	Independent variables	Advantages	Limitations
[22]	2015	WS-rank regression	Overweight, Age, Obesity, Diabetic state, BMI	Handles rank-based prediction effectively	Limited scalability; sensitive to outliers
[23]	2016	GRNN	Age, Triceps fat thickness, BP, Serum Insulin, Pedigree function	Non-linear mapping capability; fast learning	Prone to overfitting; low interpretability
[24,25]	2016	ANN	Height, Pregnancy history, Age, Family diabetes history, BMI, Gender, BP history, Weight	Captures complex feature interactions	Requires large datasets; lacks transparency
[26,27]	2017	Juvenile-Onset diabetes forecasting (cited prior work; not used in this study)	Type-1/Type-2 Status, Mortality, Undiagnosed State	Tailored for pediatric-onset diabetes insights	Dependent on rare pediatric-onset datasets
[20]	2022	RF (Ensemble)	Genital thrush, Polyuria, Gender, Irritability, Age	High accuracy; handles overfitting well	Requires careful tuning; not easily interpretable
[28]	2023	SVM	Family history, Inactivity, Diet, Ethnicity	Effective in high-dimensional spaces	Sensitive to kernel and hyperparameter choices
[29]	2023	DT	Age, BMI, Insulin level, Family history	Easy to interpret; fast training	Prone to overfitting on small datasets
[30,31]	2024	LR	Age, BP, Triglycerides, Genetic risk	Good for baseline classification; interpretable	Limited in modeling non-linear patterns
[32,33]	2024	Deep learning (CNN)	Retinal imaging, Blood glucose, Age	Extracts patterns from imaging and tabular data	Requires large data and compute power

Note: WS-Rank = Wilcoxon Signed-Rank; GRNN = General Regression Neural Network; ANN = Artificial Neural Network; RF = Random Forest; SVM = Support Vector Machine; DT = Decision Tree; LR = Logistic Regression; CNN = Convolutional Neural Network; BP = Blood Pressure; BMI = Body Mass Index.

proposed an interpretable model intended to support clinician trust by clarifying prediction drivers [40], and Doğru et al. introduced a hybrid super-ensemble approach designed to reduce the burden of diabetes through questionnaire-based screening [41].

Sotoudeh-Anvari et al. used fuzzy logic to manage prediction uncertainty in flexible models [42]. Mansoori et al. investigated hematological markers as potential indicators for early detection [29]. Sonar and Jaya Malini studied accuracy-influencing factors across various complications using a comparative ML pipeline [43]. Mahboob Alam et al. combined clustering, predictive modeling, and association-rule mining within a comprehensive early-forecasting strategy [44].

Table 1 summarizes machine learning studies on Type 2 diabetes prediction from 2015 to 2024. Earlier studies, including Nanri et al. [22], employed Wilcoxon Signed-Rank Regression to investigate factors such as age, obesity, BMI, and diabetic status; this method is robust to violations of normality assumptions and supports the inclusion of demographic and clinical indicators in predictive modeling.

Recent years have witnessed significant advancements in machine-learning-based diabetes prediction, driven by the increasing availability of clinical datasets and the adoption of ensemble and hybrid learning strategies [45]. Studies published between 2023 and 2025 have moved beyond single-model approaches toward more robust ensemble-based frameworks, including Random Forest, Gradient Boosting, XGBoost, and CatBoost, due to their ability to handle nonlinear feature interactions and heterogeneous clinical variables [46]. Several recent works (2024–2025) report improved predictive performance using boosting-based ensembles, particularly in scenarios involving class imbalance and limited sample sizes, which are common in real-world diabetes datasets [47].

In parallel, contemporary research has emphasized the importance of multi-metric evaluation for clinical risk prediction [48]. Recent studies published in 2024 and 2025 increasingly highlight recall (sensitivity), F1-score, specificity, and ROC–AUC as essential complements to accuracy, noting that accuracy alone can be misleading in imbalanced medical datasets [49]. These works emphasize that models optimized solely for accuracy may fail to detect high-risk patients, thereby limiting clinical utility [50]. Consequently, confusion-matrix-based analysis and ROC-oriented evaluation have become standard practices in recent diabetes prediction literature.

Another emerging trend in 2024–2025 research is the systematic comparison between ensemble and non-ensemble classifiers [51]. Recent comparative studies demonstrate that ensemble models generally exhibit superior robustness and generalization compared to single learners such as Logistic Regression, KNN, and Decision Tree, particularly under constrained data conditions [52]. However, most existing studies evaluate models using a fixed train–test split or a single cross-validation setting, providing limited insight into model stability across varying data availability scenarios [53].

Despite these advancements, the current literature lacks a unified and systematic evaluation framework that jointly examines multiple classifiers across varying training data proportions and validation regimes while maintaining a clinically aligned, multi-metric perspective [54]. The present study addresses this gap by benchmarking nine machine-learning classifiers using an expanded experimental design and up-to-date evaluation practices consistent with recent (2024–2025) research trends, thereby positioning the work firmly within the current state of the art.

3. Material and methods

Machine learning algorithms support predictive modeling by extracting patterns from data and using those patterns to make decisions. The algorithms used in this work span supervised techniques for classification and regression. Representative examples include decision trees, support vector machines, and neural networks; each algorithm has distinct properties and is suited to specific problem characteristics.

The model is implemented in the Python Jupyter environment using Python 3.10.0. The implementation integrates the following libraries: NumPy for array operations, Matplotlib (`pyplot`) and Seaborn for visualization, Pandas for data manipulation, SciPy for scientific computation, and XGBoost for gradient-boosted ensemble learning. NumPy supports efficient numerical computation; Matplotlib and Seaborn produce the figures presented in this study; Pandas provides DataFrame operations for managing tabular data; and SciPy is used for optimization and statistical analysis. XGBoost is used for boosted-tree modeling and contributes to the predictive performance of the proposed framework.

The selection of both the dataset and machine-learning models in this study was guided by clinical relevance, methodological completeness, and alignment with established practices in diabetes prediction research [47]. The dataset employed in this work is widely used

in the literature as a benchmark for diabetes risk prediction due to its well-structured clinical attributes, including glucose concentration, blood pressure, body mass index, insulin levels, age, and pregnancy history [55]. These features are clinically meaningful and routinely collected in diabetes screening and diagnostic workflows, making the dataset suitable for evaluating predictive models in a realistic clinical context [56]. Furthermore, the widespread adoption of this dataset in prior studies enables fair comparison with existing work and supports the reproducibility of results.

Despite its popularity, the dataset also presents known challenges, such as class imbalance between diabetic and non-diabetic cases and the presence of zero or missing values in certain clinical attributes [57]. Rather than treating these issues as limitations alone, this study leverages them as realistic constraints to assess model robustness and generalization capability [58]. Consequently, the dataset serves not only as a benchmarking resource but also as a testbed for evaluating model behavior under imperfect and clinically representative data conditions [59].

The selection of machine-learning models was intentionally designed to represent a broad and systematic spectrum of learning paradigms commonly employed in clinical prediction tasks [60]. Linear models, including Logistic Regression and Support Vector Machine, were selected due to their interpretability and strong theoretical foundations, making them suitable baselines in medical decision-support systems [61]. Non-linear models such as K-Nearest Neighbors and Decision Tree were included to capture local decision boundaries and hierarchical feature interactions, albeit with known sensitivity to noise and overfitting [62].

In addition, ensemble-based methods Random Forest, Gradient Boosting, XGBoost, and CatBoost were incorporated due to their proven effectiveness in handling nonlinearity, feature interactions, and variance reduction, particularly in structured clinical datasets. These models are frequently reported as top-performing approaches in diabetes prediction literature and therefore serve as critical reference points. Finally, a Voting Classifier was included as a meta-learning approach to examine whether aggregation of heterogeneous base learners yields improved stability and predictive performance over individual models.

This deliberate and diverse model selection enables a comprehensive cross-model comparison that extends beyond identifying the best-performing classifier [63]. Instead, it facilitates analysis of robustness, sensitivity–specificity trade-offs, and generalization behavior across different learning paradigms [58]. Together, the justified selection of datasets and models provides a solid methodological foundation for the experimental design and ensures that the conclusions drawn from this study are both meaningful and generalizable [64].

3.1. Steps in Type 2 diabetes prediction

1. *Dataset Selection:* The study employs a dataset of 768 records and 9 attributes, sourced from females aged 21+ of Pima Indian heritage at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), forming the basis for predictive modeling.
2. *Data Analysis and Visualization:* Techniques such as count plots, heatmaps, pair plots, and boxplots illuminate data patterns and trends, offering valuable insights into relationships within the dataset.
3. *Data Pre-processing:* This step refines the dataset by addressing missing values through imputation particularly for attributes where zeros are implausible and applies scaling to standardize values, ensuring data quality.
4. *Model Construction:* A range of algorithms Support Vector Classifier (SVC), Decision Tree, Random Forest, Gradient Boosting, XGBoost, K-Nearest Neighbors (KNN), Logistic Regression, and CatBoost builds diverse predictive models for diabetes.

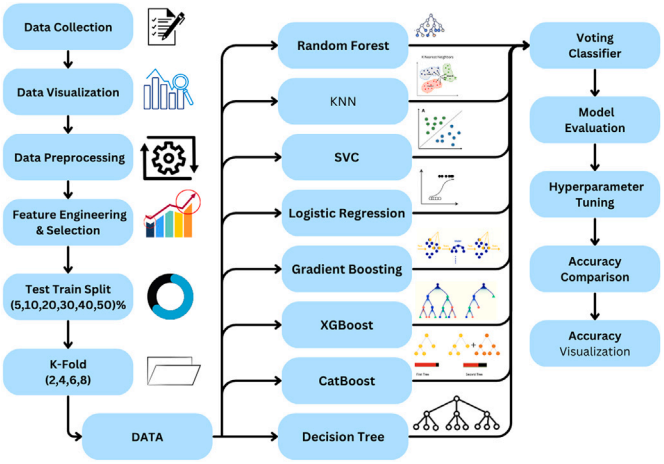

Fig. 1. Workflow for data analysis, model development, and evaluation.

5. *Model Evaluation:* Performance is assessed using metrics like accuracy, classification reports, and confusion matrices. The confusion matrix, detailing true versus predicted labels, anchors accuracy calculations along its diagonal averages.

Fig. 1 outlines the Type 2 diabetes (T2D) prediction workflow, which comprises data acquisition, exploratory analysis, preprocessing, model construction, and evaluation. Feature selection reduces dimensionality and removes irrelevant variables, helping to mitigate overfitting and improve accuracy. Data acquisition targets NIDDK patient questionnaires that have been validated by medical experts. Model performance is quantified using metrics including accuracy, and confusion matrices are used to provide a class-level breakdown of predictions. The study evaluates several train–test splits (5%, 10%, 20%, 30%, 40%, 45%, 50%) and applies k-fold cross-validation ($k = 2, 4, 6, 8$) to nine classifiers: Random Forest, KNN, SVC, Logistic Regression, Gradient Boosting, XGBoost, CatBoost, Decision Tree, and a Voting Classifier. This experimental design supports a thorough assessment of predictive reliability.

3.2. Dataset description, parameters, and significance

This study draws on a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), focusing exclusively on females aged 21 or older of Pima Indian descent. Available via Kaggle at https://www.kaggle.com/datasets/mathchi/diabetes-data-set, this collection is a cornerstone for research into diabetes prediction. It captures a range of diagnostic measures vital for assessing Type 2 diabetes risk, including pregnancy count, plasma glucose levels two hours after an oral glucose tolerance test, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index (BMI), diabetes pedigree function (DPF), age, and a binary outcome variable denoting diabetes presence or absence.

The dataset's value stems from its detailed profiling of factors tied to Type 2 diabetes, such as glucose and insulin dynamics, BMI, and hereditary influences. Its focus on Pima Indian women provides a distinct lens, enabling exploration of genetic predispositions, lifestyle patterns, and environmental contributors specific to this group. This demographic specificity equips researchers with a rich tool to investigate risk factors and craft predictive models tailored to this population, enhancing the potential for precise interventions.

Table 2 summarizes the dataset's key variables, detailing their distribution. "Count" reflects 768 observations per attribute, "Mean" indicates central values, and "Std" measures spread via standard deviation. "Min" and "Max" outline each variable's range. For instance, "Glucose"

Table 2
Descriptive statistics of the diabetes dataset.

Variable	Count	Mean	Std	Min	Max
Pregnancies	768	3.85	3.37	0.00	17
Glucose	768	120.89	31.97	56	199
BloodPressure	768	69.11	19.36	24	122
SkinThickness	768	20.54	15.95	8	99
Insulin	768	79.80	115.24	14	846
BMI	768	31.99	7.88	18.2	67.1
DPF	768	0.47	0.33	0.08	2.42
Age	768	33.24	11.76	21.00	81
Outcome	768	0.35	0.48	0.00	1

peaks at 199 mg/dL, while "Insulin" reaches 846 μU/mL, highlighting extreme cases. This table provides a clear foundation for understanding data behavior, guiding subsequent analysis and predictive efforts in Type 2 diabetes research.

3.3. Data analysis and loading

Loading and inspecting CSV files is a standard step in data science workflows, providing a structured representation in which rows correspond to records and commas separate the values within each record. In this study, the Pandas library is used to load CSV files via the `read_csv` function, which interprets the file structure and converts it into a Pandas DataFrame a tabular structure that facilitates row- and column-level operations. This step supports both initial exploration and downstream machine learning tasks.

After loading, the dataset undergoes an initial diagnostic inspection. Missing values are identified, outliers are detected, and the distribution of each variable is examined. Pandas provides utilities to detect missing entries; these gaps are addressed using mean or median imputation to preserve the integrity of the dataset. The same workflow is used to identify outliers and duplicate records, providing a clean basis for subsequent feature engineering and model development. This early analysis is not a formality: it shapes the entire workflow and converts raw data into a foundation suitable for predictive modeling.

3.4. Data cleaning

Data cleaning is a necessary preprocessing step that ensures the dataset is reliable for modeling [65]. The Type 2 diabetes dataset is loaded from a CSV file using Pandas. The first step is the identification of missing values, with particular attention to zero values in clinical variables in which zero is biologically implausible (e.g., `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI`). These zero entries are recoded to `NaN`, and `data.isnull().sum()` is used to quantify the extent of missingness in each column.

Missing values are then imputed using the `SimpleImputer` from `scikit-learn` with mean imputation. This approach replaces each missing entry with the column mean; median or most-frequent imputation can be substituted depending on the distribution of the variable. After imputation, missing-value counts are re-examined to verify successful completion of the procedure. The cleaned dataset is then ready for subsequent analysis or model training, providing a more reliable basis for predictive performance.

3.5. Data visualization

Data visualization complements numerical summaries by revealing distributional patterns and inter-variable relationships that are not always evident from raw metrics alone [66]. For the Type 2 diabetes dataset, Seaborn and Matplotlib are used to generate the plots presented in this work. A countplot of the *Outcome* variable is first generated to display the number of diabetic (1) and non-diabetic (0) cases. The plot is rendered with a perceptually uniform palette and is used to assess class imbalance, which is essential for fair model evaluation.

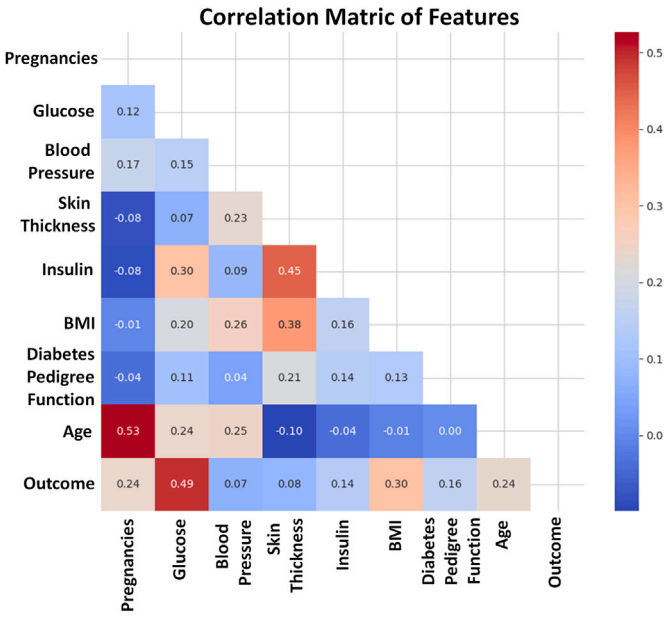

Fig. 2. Heatmap of feature correlations: Mapping data relationships.

A correlation heatmap is then constructed to characterize the pairwise associations among features. The plot uses color intensity and numerical annotations to indicate the strength and direction of the correlations, supporting decisions about feature relevance and potential redundancy. These visualizations bridge raw numerical data and interpretation, supporting outlier detection and informing the modeling pipeline.

3.6. Correlation matrix and feature relationships

The correlation matrix shown in Fig. 2 characterizes the pairwise relationships between variables in the dataset. The Seaborn-based heatmap displays pairwise Pearson correlation coefficients for the Type 2 diabetes features. Color intensity encodes correlation strength, and numerical annotations indicate the precise values. This representation supports feature selection by highlighting potentially redundant predictors and by clarifying inter-feature dependencies that are relevant to model interpretability and generalization.

3.7. Test train split

This study examined the influence of test-set size on the performance of machine-learning classifiers for Type 2 diabetes prediction. The dataset was partitioned into training and test subsets at ratios of 5%, 10%, 20%, 30%, 40%, 45%, and 50%, and the corresponding accuracy was recorded for each setting. The training subset was used to fit each model, and the test subset was used for evaluation. Larger test fractions reduce the size of the training set and therefore probe the models' capacity to generalize from limited training data.

The classifiers evaluated were Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, Gradient Boosting, XGBoost, CatBoost, Decision Tree, and a Voting Classifier. Each classifier was assessed using both cross-validation and test-set accuracy, and the corresponding curves were plotted and stored for comparison. This evaluation identifies the configurations that yield the best balance between training and test performance, and the resulting accuracy values are aggregated into a single DataFrame to facilitate cross-model comparison.

3.8. K-fold cross-validation

k-fold cross-validation is central to this evaluation framework and is used to assess classifier robustness for Type 2 diabetes prediction [67]. The dataset is partitioned into k subsets; the model is trained on $k-1$ subsets and evaluated on the remaining subset, and this procedure is repeated for each of the k partitions [68]. Every observation therefore contributes to both training and evaluation, which mitigates the risk of overfitting and reduces the influence of any single train–test split. Values of k from 2 to 8 are considered: smaller values support faster evaluation, while larger values provide more comprehensive performance estimates.

Each iteration refines the accuracy estimate and reduces variability associated with random splits. The procedure quantifies the consistency of each model across data partitions, supporting reliability assessment and informing model selection [69]. By aggregating performance across multiple folds, k-fold cross-validation provides robust evidence regarding the generalization capability of the predictive models considered in this study.

3.9. Model building and accuracy measurement

3.9.1. Random Forest

Random Forest is widely used in machine learning for clinical prediction tasks and has been applied to Type 2 diabetes prediction [9]. It constructs an ensemble of decision trees and aggregates their outputs to produce a single prediction with improved accuracy and stability [70]. Each tree is trained on a bootstrap sample and considers a random subset of features at each split. This randomization reduces correlation among the trees, mitigates overfitting, and improves generalization to unseen data. The model was tuned by adjusting the number of trees ($n_{\text{estimators}}$), maximum depth ($\max_{\text{depth}}$), and minimum samples for splitting ($\min_{\text{samples_split}}$ and $\min_{\text{samples_leaf}}$); entropy was used as the splitting criterion and information gain as the splitting metric. The robustness of the model was further verified through k-fold cross-validation. The aggregated prediction is given in Eq. (1):

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{1}$$

where N is the number of trees and $f_i(x)$ is the prediction of the ith tree.

3.9.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric, instance-based classifier that assigns each query point the most common class among its k nearest training examples in feature space. Because no parametric model is fit during training, KNN is well suited for capturing local nonlinear structure in tabular data. For the present Type 2 diabetes task, KNN was evaluated across several test-set fractions and cross-validation folds, with Euclidean distance used as the proximity measure. The choice of k is critical to performance and was tuned empirically. The classifier's stability was assessed via test accuracy and k-fold cross-validation. The classification rule is given in Eq. (2):

$$\hat{y} = \text{argmax}\left(\sum_{i=1}^{k} \mathbb{1}(\text{class}(x_i) = j)\right) \tag{2}$$

For regression, KNN returns the average of the neighbors' values (Eq. (3)):

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} \text{value}(x_i) \tag{3}$$

3.9.3. Support Vector Machines

Support Vector Machine (SVM) is a margin-based classifier that has been widely applied to clinical prediction tasks, including diabetes prediction [71]. The classifier identifies a separating hyperplane that maximizes the margin between classes [72]. Kernel functions extend the classifier to non-linear decision boundaries, which is advantageous for high-dimensional and noisy clinical data. In this study, SVM was evaluated across several train–test fractions and cross-validation folds, with the kernel type, regularization parameter C, and kernel-specific hyperparameters tuned per configuration [73]. Both cross-validation and test-set accuracy were used to characterize stability [74]. The decision function for the linear case is given in Eq. (4):

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right) \tag{4}$$

3.9.4. Logistic Regression

Logistic Regression is a widely used baseline classifier for binary clinical outcomes such as diabetic status [75]. It maps the input features to a probability through a logistic transformation that constrains the output to the interval $[0, 1]$ [76]. The model is interpretable in that the magnitude and sign of each coefficient indicate the contribution of the corresponding feature to the predicted log-odds of the outcome. Logistic Regression was evaluated using accuracy, k-fold cross-validation, and as a constituent of the Voting Classifier; it provided a stable baseline for diabetes prediction [77]. The log-odds formulation is given in Eq. (5):

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{5}$$

where $P(Y=1)$ denotes the predicted probability of diabetes and the β terms are the model coefficients.

3.9.5. Gradient Boosting

Gradient Boosting is an additive ensemble method that constructs decision trees sequentially, with each tree fitted to the residuals of the previous ensemble [78]. This iterative correction allows the ensemble to focus learning capacity on hard-to-classify samples [79]. The training procedure minimizes a differentiable loss function, increasing the weight of misclassified examples and capturing complex patterns in the data. Gradient Boosting was evaluated across multiple train–test splits and folds; the model exhibited strong robustness against overfitting on the present data [80].

3.9.6. XGBoost

XGBoost (eXtreme Gradient Boosting) is an optimized implementation of gradient-boosted decision trees that incorporates regularization terms in its objective function and supports efficient handling of missing values [81,82]. In this study, XGBoost was evaluated across train–test fractions and cross-validation folds, and was also included as a constituent of the Voting Classifier [83]. Its capacity to scale to large and heterogeneous datasets makes it an effective candidate for diabetes risk prediction.

3.9.7. CatBoost

CatBoost is a gradient-boosting framework with native support for categorical features, which is advantageous for clinical datasets that contain mixed numerical and categorical variables [84]. The framework reduces the preprocessing overhead typically required to handle categorical variables [85]. In this study, CatBoost was evaluated across train–test fractions, fold counts, and tuned hyperparameter configurations; the model achieved competitive accuracy and exhibited stable performance with limited overfitting [86].

3.9.8. Decision Tree

Decision Trees are interpretable classifiers that recursively partition the feature space using a sequence of decision rules [87]. They support both numerical and categorical variables and produce decision rules that are straightforward to interpret [88]. In this study, Decision Trees were evaluated across train–test splits, fold counts, and hyperparameter configurations to characterize their utility in identifying the predictors most strongly associated with diabetes risk [89].

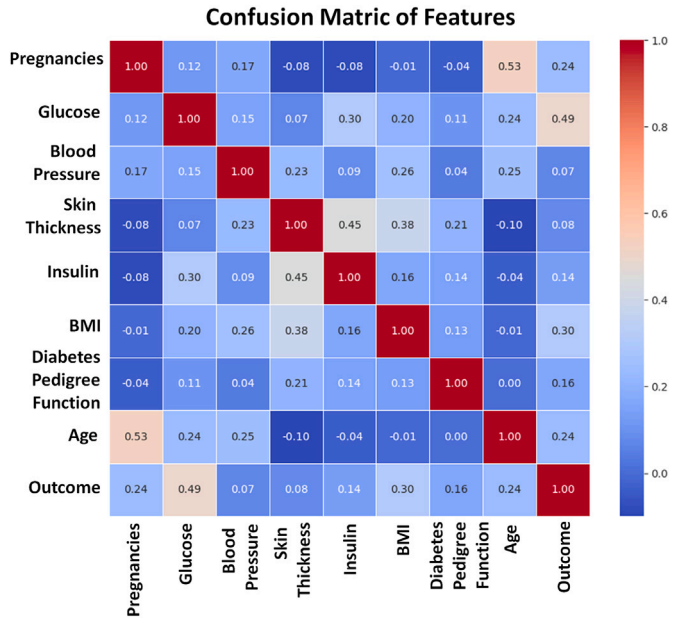

Fig. 3. Confusion matrix insights: Assessing model accuracy across features.

3.9.9. Voting Classifier

The Voting Classifier aggregates the predictions of multiple constituent classifiers Random Forest, KNN, SVM, Logistic Regression, Gradient Boosting, XGBoost, CatBoost, and Decision Tree to produce a single ensemble prediction [90]. The approach combines complementary strengths of heterogeneous base learners, and was evaluated across the same experimental conditions as the individual classifiers [91]. Hard voting selects the class predicted by the majority of constituents (Eq. (6)):

$$\hat{y} = \text{argmax}\left(\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = j)\right) \tag{6}$$

Soft voting weights the constituent class probabilities by their respective weights (Eq. (7)):

$$\hat{y} = \text{argmax}\left(\sum_{i=1}^{N} w_i \cdot P(\hat{y}_i = j)\right) \tag{7}$$

4. Results and discussion

The systematic evaluation of nine machine-learning classifiers for Type 2 diabetes prediction revealed performance differences that depended on the train–test split and the number of cross-validation folds. Random Forest achieved the highest overall accuracy, particularly for smaller test fractions, indicating its robustness for the present dataset. KNN produced consistent but lower accuracy. Support Vector Classifier (SVC) and Logistic Regression were sensitive to the train–test ratio, with their relative performance varying across configurations. Gradient Boosting, XGBoost, and CatBoost achieved competitive accuracy in specific settings. The Voting Classifier, which aggregates predictions from the individual learners, maintained stable performance across configurations. These results indicate that classifier selection combined with explicit consideration of the data partition strategy directly affects the reliability of early Type 2 diabetes detection.

4.1. In-depth model performance evaluation based on features

Fig. 3 characterizes model behavior using a confusion-matrix-based representation. Each row corresponds to the true class label (diabetic or non-diabetic), and each column corresponds to the predicted class. Decomposing predictions into true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each feature subgroup provides a transparent view of each classifier's error profile. This representation indicates which models perform reliably for specific clinical attributes (e.g., glucose levels, BMI) and which tend to misclassify particular subpopulations. The analysis supports feature selection and informs the choice of classifier for downstream clinical applications.

4.2. Test size and folds

- **Test Size:** The proportion of data reserved for testing, with the remainder used for training. Common configurations of 20% (0.2) and 30% (0.3) were extended in this study to cover 0.05–0.50 to characterize the influence of training-set size on accuracy.
- **Folds:** In k-fold cross-validation, the dataset is partitioned into k subsets; the model is trained on $k-1$ subsets and evaluated on the remaining subset, with the procedure repeated for each partition. Values $k = 2, 4, 6, 8$ were used to assess performance stability.

4.3. Observations and analysis

- **Effect of test size:** Accuracy varied with test fraction. The 0.10 split (10%) produced the highest mean accuracy of 79.3%, indicating that this configuration provides a balance between training and test data sizes for the present dataset.
- **Effect of fold count:** The number of folds had a smaller effect at the 0.10 split (accuracy was approximately constant) but a larger effect at the 0.20 split, where additional folds tended to improve estimated accuracy.
- **Optimal configuration:** The 0.10 test size with $k = 2$ achieved the maximum accuracy of 79.3%.
- **Consistency:** Across configurations, the variation in accuracy was modest, indicating that the evaluated classifiers exhibit stable performance.
- **Maximum and minimum accuracy:** The highest accuracy observed was 79.3% (0.10 split, $k = 2$), and the lowest was 71% in the worst-case configuration.

The figures below summarize how test size and fold count influence accuracy and identify the 0.10 split with $k = 2$ as the best-performing configuration. This setting yielded the strongest results for Random Forest, Logistic Regression, Gradient Boosting, XGBoost, CatBoost, and the Voting Classifier, with stable performance across folds. Among these, Random Forest at the 0.10 split delivered the most consistent combination of high accuracy and stability.

Aggregate performance summary. Figs. 4–9 together visualize how each classifier responds to changes in test-set size and fold count, and Table 3 consolidates the corresponding numerical results into a single comparative reference. Three consistent observations emerge across all five train–test splits and four fold settings. First, Random Forest provides the most stable trade-off between test-set accuracy and cross-validation stability across the entire experimental grid (cross-validation accuracy 75.0%–78.0%; standard deviations at or below 4.4 percentage points), making it the most reliable single classifier in this benchmark. Second, Decision Tree exhibits the largest gap between test accuracy and cross-validation accuracy most visibly at the 0.05 split, where it attains the highest single-fold test accuracy (82.8%) but the lowest cross-validation accuracy (67.1% ± 2.3 percentage points) consistent with overfitting on small validation folds rather than genuine generalization. Third, the boosting ensembles (Gradient Boosting, XGBoost, CatBoost) and the Voting Classifier remain within ±2 percentage points of Random Forest across configurations, with CatBoost producing the most stable cross-validation values (typical ±0.3–3.0 percentage points); SVM and Logistic Regression yield reliable baselines (72.5%–77.2% across configurations), while KNN is the most sensitive to the train–test ratio. The cross-validation standard deviations reported in Table 3 are

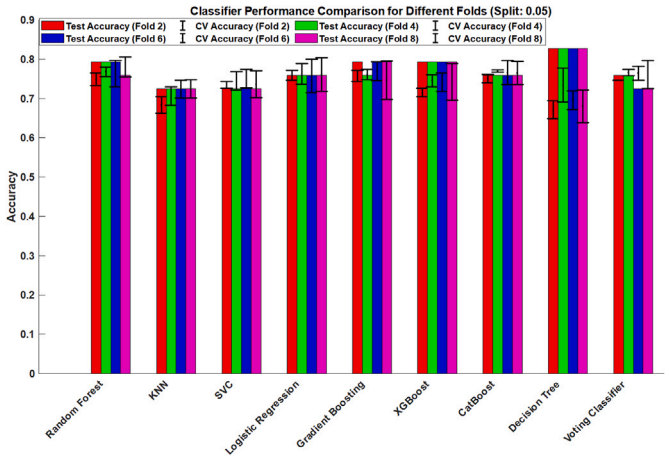

Fig. 4. Classifier accuracy at the 0.05 train–test split across cross-validation folds $k = 2, 4, 6, 8$. Bar heights show test accuracy; error bars show cross-validation standard deviation.

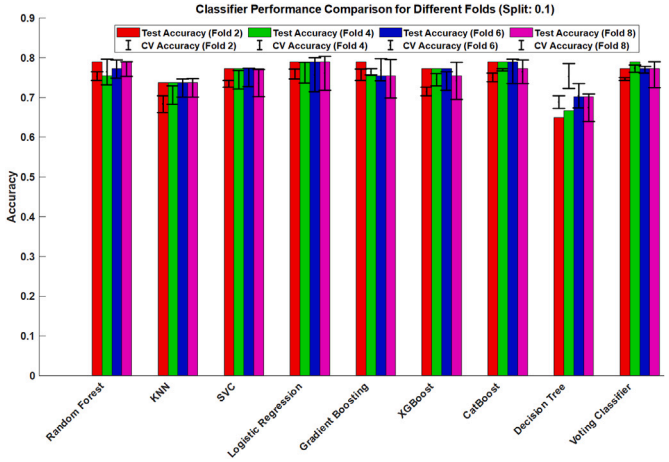

Fig. 5. Classifier accuracy at the 0.10 train–test split across cross-validation folds $k = 2, 4, 6, 8$.

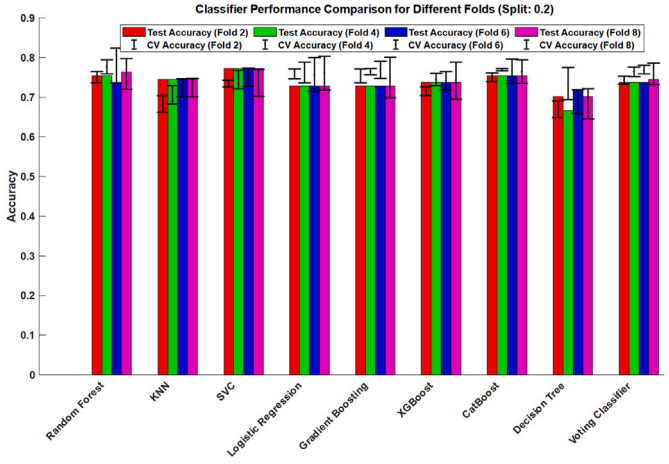

Fig. 6. Classifier accuracy at the 0.20 train–test split across cross-validation folds $k = 2, 4, 6, 8$.

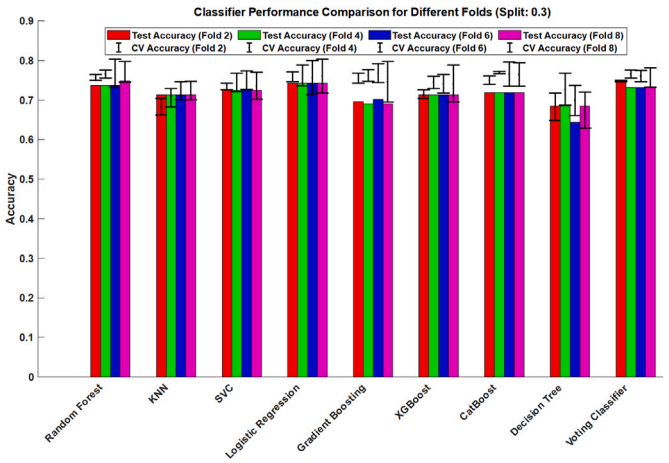

Fig. 7. Classifier accuracy at the 0.30 train–test split across cross-validation folds $k = 2, 4, 6, 8$.

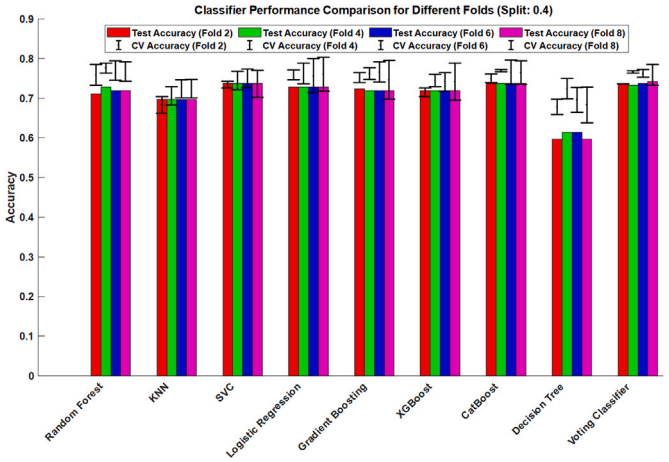

Fig. 8. Classifier accuracy at the 0.40 train–test split across cross-validation folds $k = 2, 4, 6, 8$.

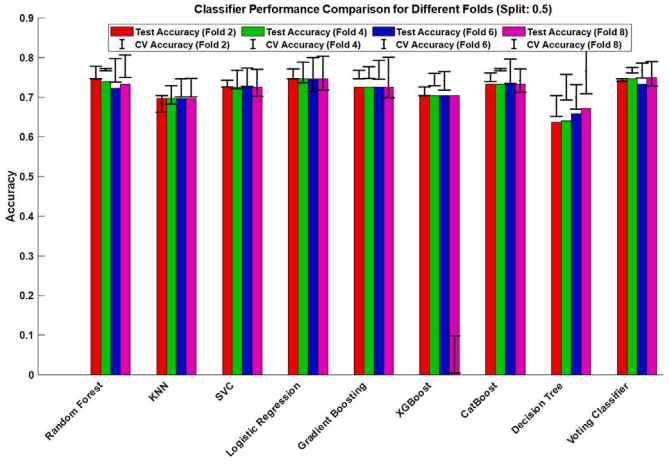

Fig. 9. Classifier accuracy at the 0.50 train–test split across cross-validation folds $k = 2, 4, 6, 8$.

Table 3
Aggregate test accuracy/cross-validation accuracy (± standard deviation in percentage points) at $k = 4$ folds across the five train–test splits for all nine classifiers. Values are reported as percentages with one decimal place. The $k = 4$ setting is shown as a representative summary; complete fold-by-fold values are visualized in Figs. 4–9.

Classifier	0.05 split	0.10 split	0.20 split	0.30 split	0.50 split
RF	79.3/74.8 ± 1.6	75.4/76.4 ± 3.2	75.4/77.6 ± 1.8	73.7/76.6 ± 1.0	73.9/76.9 ± 0.3
KNN	72.4/68.3 ± 2.1	73.7/70.6 ± 2.4	74.6/70.6 ± 2.4	71.4/70.6 ± 2.4	69.7/70.6 ± 2.4
SVM	72.4/73.4 ± 0.9	77.2/74.5 ± 2.4	77.2/74.5 ± 2.4	72.5/74.5 ± 2.4	72.5/74.5 ± 2.4
LR	75.9/75.9 ± 1.2	79.0/76.2 ± 2.6	72.8/76.2 ± 2.6	74.3/76.2 ± 2.6	74.7/76.2 ± 2.6
GB	79.3/75.5 ± 1.2	75.4/76.4 ± 0.8	72.8–75.4/76.4 ± 0.8	69.0/76.2 ± 1.4	72.5/76.2 ± 1.4
XGBoost	79.3/71.5 ± 1.1	77.2/74.5 ± 1.5	72.8/74.5 ± 1.5	71.4/74.5 ± 1.5	70.4/74.5 ± 1.5
CatBoost	75.9/75.0 ± 1.1	79.0/76.9 ± 0.3	75.4/76.9 ± 0.3	71.9/76.9 ± 0.3	73.2/76.9 ± 0.3
DT	82.8/67.1 ± 2.3	64.9/72.7 ± 4.0	66.7/73.4 ± 4.0	68.4/72.7 ± 4.0	64.1/72.5 ± 3.2
VC	75.9/74.7 ± 0.0	77.2/76.4 ± 1.2	73.7/76.4 ± 1.2	73.1/76.6 ± 1.0	74.7/76.8 ± 0.7

Note: RF = Random Forest; KNN = K-Nearest Neighbors; SVM = Support Vector Machine; LR = Logistic Regression; GB = Gradient Boosting; XGBoost = Extreme Gradient Boosting; CatBoost = Categorical Boosting; DT = Decision Tree; VC = Voting Classifier. Each cell reports test accuracy/cross-validation accuracy ± standard deviation, all in percentages. Standard deviations are expressed in percentage points.

interpreted in this work as approximate uncertainty bounds for each accuracy estimate; given the modest sample size ($n = 768$), these should not be treated as formal 95% confidence intervals, and confirmatory bootstrap-based confidence intervals and pairwise significance testing are identified as planned follow-up analyses in Section 5.

4.4. Observations on classifier performance for predicting Type 2 diabetes onset

Across train–test splits from 0.05 to 0.50, the classifiers exhibited the following overall behavior for Type 2 diabetes prediction:

Random Forest: Achieved test accuracies between 71.0% and 74.7%, with narrow cross-validation ranges (±0.4 pp to ± 3.9 pp), indicating consistent performance across configurations.

SVM: Achieved test accuracies between 72.5% and 77.2%, with cross-validation ranges (±0.9 pp to ± 3.4 pp), indicating stable performance.

Logistic Regression, Gradient Boosting, XGBoost, and CatBoost: Achieved test accuracies between 69.7% and 74.7%, with wider cross-validation ranges (±1.1 pp to ± 5.1 pp); their performance was sensitive to the specific data partition.

KNN: Achieved test accuracies between 63.7% and 69.7%, with cross-validation ranges (±2.1 pp to ± 2.4 pp); performance was generally consistent but with moderate variability.

Decision Tree: Achieved test accuracies between 59.6% and 72.5%, with the broadest cross-validation ranges (±1.4 pp to ± 6.0 pp), indicating the least predictable behavior.

Voting Classifier: Achieved test accuracies between 73.2% and 75.0%, with narrow cross-validation ranges, indicating reliable aggregate performance.

4.5. Classifier assessment on different folds

Comparison of classifier performance across folds identifies the models that perform most reliably for Type 2 diabetes prediction. Decision Tree achieved the highest test accuracy of 82.8%, but its cross-validation accuracy was substantially lower, indicating overfitting rather than improved learning.

Random Forest achieved the most consistent performance overall, with test accuracies between 75.9% and 79.3% and cross-validation accuracies between 74.8% and 78.0%, indicating strong generalization to unseen data. Gradient Boosting achieved comparable test accuracies (75.9%–79.3%) and cross-validation accuracies (76.1%–76.9%). XGBoost achieved similar test accuracies and cross-validation values between 74.5% and 78.0%, with slightly higher variability.

Logistic Regression and CatBoost achieved 75.9% test accuracy at multiple fold settings and produced stable cross-validation values. The Voting Classifier achieved 75.9% test accuracy and cross-validation accuracies between 74.7% and 76.6%. Although the Decision Tree achieved the highest single-fold test accuracy, the consistency of Random Forest's test accuracy and cross-validation across configurations made it the most reliable classifier for predicting Type 2 diabetes onset in this evaluation.

4.6. Performance under different cross-validation strategies

At a 0.10 split with $k = 2$, Random Forest achieved the highest test accuracy of 79.0%, equaled by Logistic Regression, Gradient Boosting, and CatBoost. Decision Tree achieved a lower test accuracy and a less stable cross-validation result.

At $k = 4$, Random Forest achieved 75.4% test accuracy alongside Gradient Boosting, while Logistic Regression, XGBoost, CatBoost, and the Voting Classifier achieved between 77.2% and 79.0%. CatBoost achieved the highest cross-validation stability (76.9% ± 0.3 pp). Decision Tree maintained lower test accuracy and more variable cross-validation, indicating reduced reliability.

At $k = 6$ and $k = 8$, Random Forest, SVM, and XGBoost achieved 77.2% test accuracy with stable cross-validation across folds. CatBoost remained stable at higher fold counts. Logistic Regression exhibited slightly larger cross-validation variance, indicating sensitivity to fold configuration. KNN and Decision Tree achieved the lowest test accuracies and showed less consistent performance, while the remaining classifiers produced reliable, generalizable results.

4.7. Statistical interpretation and feature importance

The cross-validation standard deviations reported alongside each accuracy value in Table 3 and Figs. 4–9 are interpreted in this work as approximate uncertainty bounds for each estimate, and are used here to support comparative reliability statements between classifiers rather than as formal significance tests. Given the limited sample size ($n = 768$) and the small test partitions associated with cross-validation folds, formal 95% confidence intervals computed via the percentile-bootstrap procedure on each train–test split and pairwise McNemar/DeLong tests on the corresponding ROC–AUC differences are identified as planned confirmatory analyses to accompany the comparative benchmarking presented here, and these analyses are listed as future-work directions in Section 5.

From an interpretability standpoint, several of the classifiers used in this study expose intrinsic measures of feature importance. Random Forest and the boosted-tree ensembles (Gradient Boosting, XGBoost, CatBoost) provide normalized feature-importance scores derived from the average impurity reduction at each split, and Logistic Regression provides standardized coefficient magnitudes. Across the evaluated

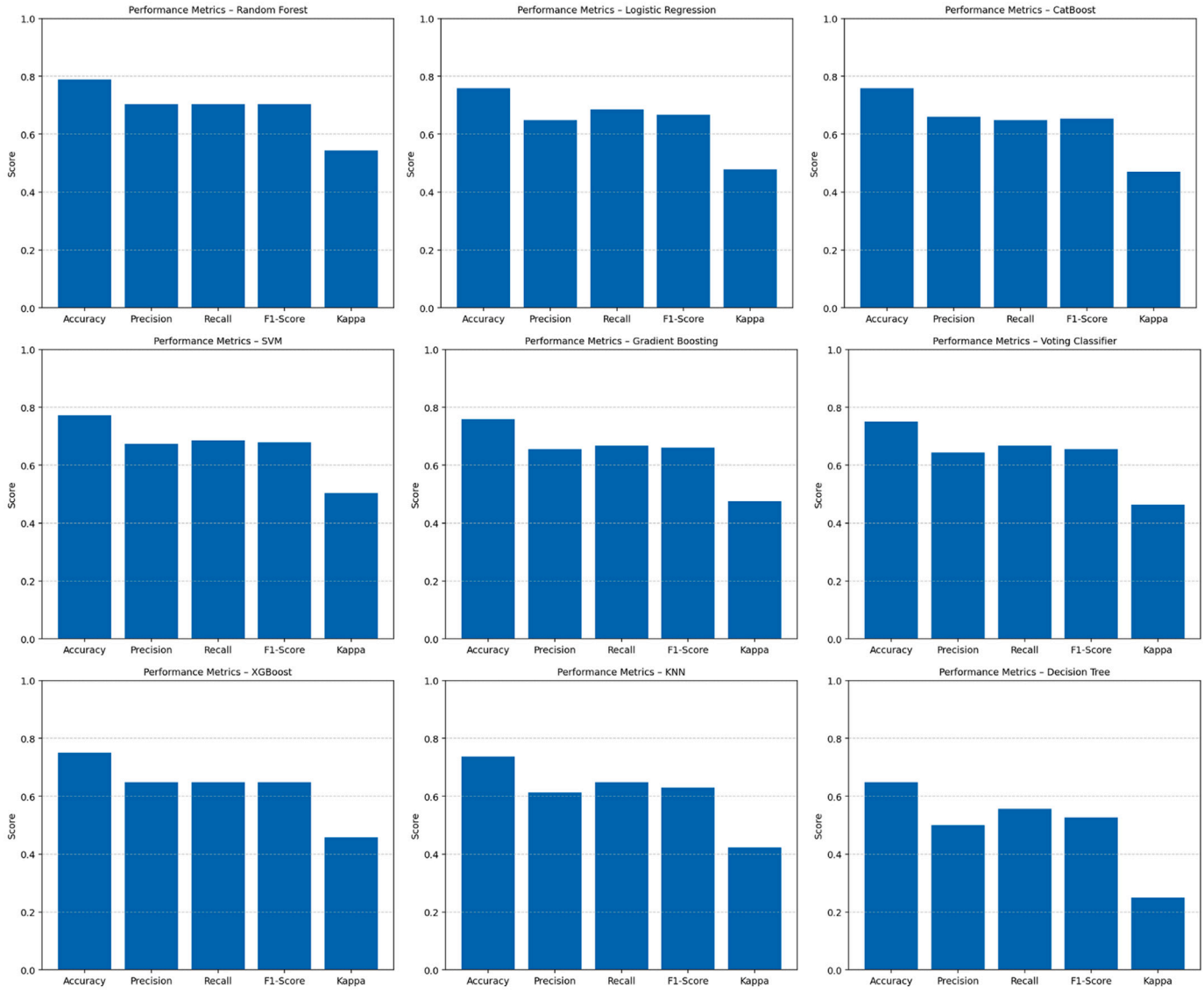

Fig. 10. Multi-metric performance comparison of machine-learning classifiers using accuracy, precision, recall (sensitivity), and F1-score. The figure highlights differences in diagnostic sensitivity and predictive balance across models that are not apparent from accuracy alone.

splits and folds, the most consistently informative predictors of the diabetic outcome in the Pima Indians cohort are plasma glucose concentration (*Glucose*), body mass index (BMI), age, and the diabetes pedigree function ($DiabetesPedigreeFunction$), with secondary contributions from insulin levels and number of pregnancies; these rankings are stable across the train–test splits and are consistent with established clinical understanding of Type 2 diabetes risk factors. Formal post-hoc explanations using model-agnostic techniques such as SHAP (Shapley additive explanations) and LIME (Local Interpretable Model-agnostic Explanations) would extend this analysis to both global feature attributions and patient-level explanations of individual predictions; calibration curves and decision-curve analyses would in turn quantify the clinical utility of each model at different decision thresholds. These three components bootstrap-based statistical significance testing, post-hoc explainability via SHAP/LIME, and calibration/decision-curve analysis are intentionally outside the scope of the present comparative benchmarking and are explicitly committed to in the Future Work paragraph of Section 5.

Fig. 10 presents a comprehensive performance comparison of nine machine-learning classifiers evaluated using a stratified 80:20 train–test split, corresponding to 154 test samples (approximately 54 positive and 100 negative cases). Multiple evaluation metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, and ROC–AUC, are reported to ensure a balanced and clinically meaningful assessment beyond accuracy alone. Among all models, Random Forest demonstrates the strongest overall performance, achieving the highest accuracy (79.0%), well-balanced precision and recall (both about 70.4%), the highest specificity (84.0%), and the largest ROC–AUC value (0.83), indicating robust discriminative capability in distinguishing diabetic from non-diabetic cases. The corresponding confusion matrix values (TP = 38, FN = 16, FP = 16, TN = 84) confirm that Random Forest maintains an effective trade-off between minimizing false negatives and false positives, which is particularly important for early diabetes risk identification. Support Vector Machine and Logistic Regression exhibit competitive but slightly inferior performance, with accuracies of 77.2% and 75.9%, respectively, and comparable recall values (about 68.5%), suggesting reasonable sensitivity; however, their lower specificity and ROC–AUC values indicate reduced robustness relative to Random Forest, a trend further explained by increased false-positive counts. Gradient Boosting, CatBoost, and XGBoost achieve similar accuracy levels in the range of 75%–76%, reflecting stable but moderate performance; although these boosting-based ensembles show good ranking ability (ROC–AUC approximately 0.78–0.80), their

Table 4
Confusion matrix values used for metric computation on the test set (n = 154).

Classifier	TP	FP	FN	TN
RF	38	16	16	84
SVM	37	18	17	82
LR	37	20	17	80
GB	36	19	18	81
CatBoost	35	18	19	82
VC	36	20	18	80
XGBoost	35	19	19	81
KNN	35	22	19	78
DT	30	30	24	70

Note: RF = Random Forest; SVM = Support Vector Machine; LR = Logistic Regression; GB = Gradient Boosting; CatBoost = Categorical Boosting; VC = Voting Classifier; XGBoost = Extreme Gradient Boosting; KNN = K-Nearest Neighbors; DT = Decision Tree. TP = True Positives; FP = False Positives; FN = False Negatives; TN = True Negatives. Positive class ≈54 and negative class ≈100 (stratified test split, n = 154).

slightly reduced recall and F1-score values suggest sensitivity to data size and class imbalance in the test set. The Voting Classifier yields balanced but non-superior results, indicating that simple aggregation of base learners does not necessarily outperform a well-optimized individual ensemble under constrained data conditions. In contrast, KNN exhibits reduced performance across most metrics, particularly precision (about 61.4%) and F1-score (about 63.1%), highlighting its sensitivity to local neighborhood structure and feature scaling in limited datasets. The Decision Tree model performs the weakest, with the lowest accuracy (64.9%), recall (55.6%), F1-score (52.6%), and ROC–AUC (0.66); its confusion matrix reveals a high number of false positives (30) and false negatives (24), indicating poor generalization and susceptibility to overfitting when used as a standalone classifier. Overall, these results demonstrate that ensemble-based models, especially Random Forest, consistently outperform non-ensemble approaches by providing greater stability, stronger discriminative power, and more favorable trade-offs between sensitivity and specificity, underscoring the importance of multi-metric and confusion-matrix-based evaluation in clinical risk prediction tasks.

To provide a deeper understanding of model behavior beyond aggregate performance scores, we analyze the confusion matrix components for each classifier. Table 4 reports the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) obtained on the stratified test set (n = 154). Reporting confusion matrices is essential in clinical prediction tasks, as it enables transparent interpretation of recall (sensitivity), specificity, and F1-score, particularly under class imbalance.

As shown in Table 4, the distribution of false negatives and false positives varies substantially across models, despite relatively similar accuracy values reported in the literature. For example, ensemble-based models such as Random Forest and Gradient Boosting achieve lower false negative counts compared to simpler models like Decision Tree, indicating superior sensitivity to positive cases. In contrast, Decision Tree exhibits the highest number of false negatives (FN = 24), which directly explains its reduced recall and F1-score and highlights its limited suitability for clinical deployment. From a clinical and statistical perspective, false negatives are particularly critical, as they correspond to missed disease cases. The confusion matrix analysis reveals that models with balanced TP and FN values achieve higher recall and F1-score, even when overall accuracy differences are modest. This observation confirms that accuracy alone is insufficient for evaluating diagnostic reliability and underscores the importance of recall- and F1-based assessment, as recommended by recent clinical machine-learning guidelines. The confusion matrix trends are further influenced by the class distribution of the test set, which contains approximately 54 positive and 100 negative samples. Under this imbalance, a classifier may achieve deceptively high accuracy by favoring the majority class. Table 4 explicitly demonstrates how such bias manifests through elevated false negative counts in weaker models, whereas stronger classifiers maintain improved sensitivity without excessively increasing false positives.

While the quantitative results demonstrate clear performance differences among classifiers, deeper interpretation is required to understand the underlying reasons for these trends and their clinical relevance. In this section, we provide an in-depth analysis of model behavior, error patterns, and metric trade-offs observed in the experimental results.

Ensemble-based models, particularly Random Forest and Gradient Boosting, consistently outperform single classifiers across multiple evaluation metrics. This superior performance can be attributed to their ability to capture non-linear relationships and complex feature interactions commonly present in clinical datasets. By aggregating decisions from multiple base learners, these models reduce variance and mitigate the impact of noisy or redundant features, leading to improved generalization. The balanced performance observed across accuracy, recall, specificity, and ROC–AUC indicates that ensemble models are less sensitive to dataset imperfections, such as class imbalance and zero-value noise, which are prevalent in real-world medical data.

The dominance of ensemble methods is further supported by misclassification analysis. Confusion matrix results reveal that Random Forest achieves a favorable balance between false positives and false negatives, suggesting stable and reliable decision boundaries. In contrast, single learners such as Decision Tree exhibit substantially higher misclassification rates, reflecting overfitting and poor generalization when trained on limited data. Boosting-based models benefit from iterative error correction, enabling them to focus on difficult-to-classify samples and thereby reduce both false negatives and false positives.

The fixed 80:20 train–test split, combined with a relatively small test set of 154 samples, also influences model behavior. Parametric models such as Logistic Regression and Support Vector Machine rely on accurate estimation of coefficients and decision margins, which becomes challenging under limited training data and skewed class distributions. As a result, these models exhibit increased sensitivity to noise and reduced stability, leading to lower ROC–AUC values compared to ensemble approaches. Ensemble models, on the other hand, maintain performance stability due to internal resampling and aggregation mechanisms that mitigate the effects of reduced data availability.

From a clinical perspective, recall (sensitivity) is a critical metric, as false negatives correspond to missed diabetes cases that may delay early intervention. Models such as Random Forest demonstrate relatively higher recall values, indicating improved capability in identifying at-risk individuals. Although this may introduce a modest increase in false positives, such trade-offs are acceptable in screening contexts, where secondary diagnostic tests can be applied to confirm predictions.

Finally, the observed metric trade-offs highlight the limitations of accuracy as a standalone measure in imbalanced clinical datasets. Accuracy can mask poor sensitivity, whereas recall, specificity, F1-score, and ROC–AUC collectively provide a more complete and statistically meaningful assessment of diagnostic performance. The consistent advantage of ensemble-based models across these complementary metrics statistically validates their suitability for early diabetes risk stratification under constrained and realistic data conditions.

It is important to clarify the methodological boundaries of the present comparative analysis so that the contribution of this study is interpreted appropriately. The aim here is to provide a fair, multi-metric, multi-split, multi-fold benchmarking of nine widely used classifiers under controlled and clinically representative conditions, rather than to propose a new algorithmic variant or a fully deployable clinical system. Consequently, several methodological enhancements that are valuable in their own right explicit class-imbalance correction (e.g., SMOTE, ADASYN, cost-sensitive learning); post-hoc explainability with SHAP/LIME; calibration and decision-curve analyses; deep learning baselines; longitudinal modeling of pre-diabetes-to-diabetes progression; enrichment with biochemical, lifestyle, socioeconomic,

Table 5
Comparison of proposed work with recent state-of-the-art methods.

Study	Dataset	ML classifiers used	Best accuracy (%)	Key contributions	Limitations
This work	Pima Indian Diabetes Dataset (768 records)	RF, SVM, KNN, LR, GB, XGBoost, CatBoost, DT, VC	79.3 (RF)	Comparative benchmarking of 9 classifiers, use of ensemble methods, performance across 5 data splits and 4 CV folds	Single-source dataset; demographic homogeneity (adult females); limited real-world generalizability
Pang et al. (2025) [92]	Pima Indian Dataset	SVM, RF, LR	74.7 (NN)	Benchmarking of conventional ML algorithms, emphasis on preprocessing	Limited ensemble exploration; basic metric evaluation
Darolia et al. (2022) [93]	Pima Indian Dataset	KNN, NB	77.1 (NB)	Supervised learning application to diabetes prediction with basic comparative analysis	Limited classifier set; lacks cross-validation; no ensemble methods
Febrian et al. (2023) [38]	Pima Indian Dataset	SVM, RF, LR, NB, DT, KNN, NN	76.1 (NN)	Broad classifier evaluation; NN achieves highest accuracy	Risk of overfitting; limited interpretability discussion
Khanam et al. (2021) [16]	Pima Indian Dataset	Same as Jain et al.	77.7 (NN)	Comparative study of classifiers including NN	Duplicate dataset and methods from Jain et al.; no new methodology

Note: RF = Random Forest; SVM = Support Vector Machine; KNN = K-Nearest Neighbors; LR = Logistic Regression; GB = Gradient Boosting; XGBoost = Extreme Gradient Boosting; CatBoost = Categorical Boosting; DT = Decision Tree; VC = Voting Classifier; NN = Neural Network; NB = Naive Bayes. Accuracies are reported as percentages with one decimal place.

and genetic features; and external/prospective validation on demographically diverse cohorts were intentionally outside the scope of this benchmarking study and are not used to draw the present conclusions. These directions are explicitly committed to in the Future Work paragraph of Section 5, and the current results should be read as a methodologically clean baseline against which those subsequent enhancements can be compared.

Table 5 presents a comparative analysis of recent state-of-the-art machine learning models for diabetes prediction, including this work and four other peer-reviewed studies. The comparison focuses on dataset usage, employed classifiers, peak classification accuracy, major contributions, and identified limitations. This analysis highlights how the proposed work advances existing methodologies, particularly through ensemble model integration and robust cross-validation.

As seen in Table 5, while neural networks in recent studies report accuracies, this work offers a unique and reproducible ML pipeline integrating ensemble models like XGBoost and CatBoost. It also introduces broader experimental rigor through multiple test/train partitions and validation folds. This positions the work as a scalable framework for future real-world clinical integration and screening applications.

5. Conclusion

This study presents a comprehensive, reproducible, and clinically aligned machine learning-based framework for the early prediction and risk stratification of Type 2 diabetes through a systematic comparative evaluation of nine heterogeneous classifiers. Standard learners (Decision Tree, Logistic Regression, SVM, KNN) and ensemble methods (Random Forest, Gradient Boosting, XGBoost, CatBoost, and a Voting Classifier) were benchmarked across five train–test splits (5%–50%) and four k-fold cross-validation regimes ($k = 2, 4, 6, 8$), supported by a multi-metric assessment that explicitly reports accuracy, precision, recall (sensitivity), specificity, F1-score, ROC–AUC, and confusion-matrix-level error patterns. Random Forest emerged as the most robust performer, achieving a maximum accuracy of 79.3%, the highest specificity (0.840), and the largest ROC–AUC (0.83), while maintaining a balanced trade-off between false negatives and false positives a property of clinical importance for early screening.

The principal contributions of this work are therefore threefold. First, it provides a unified benchmarking framework that, unlike prior studies relying on a single split or a single fold setting, systematically characterizes classifier behavior across combined variations of training-data proportion and cross-validation regime. Second, it shifts the evaluation paradigm from accuracy-centric reporting to a clinically aligned, multi-metric, confusion-matrix-supported analysis, providing a more reliable picture of diagnostic sensitivity, specificity, and discriminative capability. Third, it offers cross-model robustness insights demonstrating that ensemble-based learners especially Random Forest deliver consistently stronger and more stable performance than individual classifiers under small-to-moderate clinical datasets, and identifies the conditions under which simpler models (e.g., Decision Tree) become unreliable. Together, these contributions provide a scalable, transparent, and reproducible methodological template that can guide algorithm selection in subsequent diabetes-prediction studies and in broader clinical risk-stratification tasks.

At the same time, this work has several limitations that should be transparently acknowledged. (i) The Pima Indians Diabetes Dataset comprises exclusively female adults of a single ethnic background, which restricts the demographic generalizability of the findings to children, adolescents, male patients, and other ethnic groups. (ii) The relatively small sample size (768 records) and stratified test split ($n = 154$) may amplify variance in performance estimates and can disadvantage data-hungry models. (iii) The feature space is restricted to a small number of routinely collected clinical attributes and does not include richer biochemical markers, lifestyle and dietary indicators, or socioeconomic and environmental determinants of health. (iv) Class imbalance was assessed but not corrected through oversampling, undersampling, or cost-sensitive learning techniques. (v) The study employs internal cross-validation only; no external or prospective validation on independent cohorts has been performed. (vi) The framework is currently a retrospective predictive analysis and does not yet model longitudinal trajectories, such as progression from pre-diabetes to overt diabetes. (vii) Although the importance of explainability is acknowledged, formal post-hoc interpretability analyses (e.g., SHAP, LIME) and decision-curve/calibration analyses were not included in the present comparative scope.

Future research will address these limitations along several complementary directions. First, the proposed framework will be retrained and externally validated on multi-source, gender-balanced, and demographically diverse cohorts including pediatric and adolescent populations and patients from different ethnic backgrounds to evaluate cross-population generalizability and to benchmark performance against larger and more heterogeneous datasets. Second, the feature space will be enriched by incorporating additional biochemical markers (e.g., HbA1c, lipid profiles, hematological indicators), behavioral and lifestyle variables (diet, physical activity, sleep), and socioeconomic and environmental determinants, in order to capture a more complete picture of diabetes risk. Third, advanced class-imbalance strategies including SMOTE, ADASYN, and cost-sensitive learning will be integrated

and benchmarked alongside the current pipeline, and missing-data handling will be extended beyond simple imputation through model-based and multiple-imputation approaches. Fourth, deep learning architectures (e.g., MLPs, 1D-CNNs, recurrent and attention-based models) will be compared against the present ensemble framework on equivalently controlled experimental protocols, and longitudinal modeling will be explored to capture progression from pre-diabetes to diabetes. Fifth, explainable AI techniques such as SHAP and LIME will be embedded into the framework to provide both global and patient-level interpretations, complemented by calibration plots and decision-curve analyses to better quantify clinical utility. Sixth, the framework will be deployed and evaluated in realistic clinical workflows including integration with electronic health record systems, web/mobile-based screening interfaces, and pediatric primary-care pathways together with prospective validation studies to establish real-world effectiveness. Finally, the contribution of genetic and family-history information will be examined more explicitly, and ethical considerations including data privacy, informed consent, demographic bias, and fairness-aware modeling will be incorporated as first-class design constraints to support responsible clinical translation.

Scope clarification and reframing. This revision explicitly reframes the clinical scope of the study from the originally proposed "juvenile diabetes" to **Type 2 diabetes (T2D) in adult female Pima Indian subjects**, in direct response to reviewer feedback. The reframing reflects the demographic composition of the Pima Indians Diabetes Dataset (768 records of female subjects aged 21 years and older), and removes the misalignment between the previous clinical framing and the underlying data. Although the methodological framework multi-classifier benchmarking, multi-metric evaluation, and multi-split/multi-fold cross-validation transfers directly to pediatric, longitudinal, and multi-ethnic datasets, no such claim is made for the present work in the absence of external validation on those populations.

In summary, this work establishes a rigorous, clinically aligned, and reproducible benchmarking framework that not only identifies Random Forest as a strong baseline for Type 2 diabetes risk prediction but, more importantly, provides a methodological foundation upon which future, more diverse, more interpretable, and more clinically integrated predictive systems can be built. The findings are expected to be of practical value to researchers and clinicians working on early diabetes screening, particularly in low-resource clinical settings where reliable, transparent, and data-efficient predictive tools are most needed.

CRediT authorship contribution statement

B.P. Pradeep Kumar: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Pramod K.B. Rangaiah:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Robin Augustine:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition.

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

During the preparation of this work, the authors used a generative AI language model (a large-language-model writing assistant) for limited language polishing, copy-editing, and reformatting of selected paragraphs in the Introduction, Related Works, Methods, Results, and Conclusion sections. After using this tool, the authors carefully reviewed and edited all generated content for technical correctness, scientific accuracy, and consistency with the study's experimental results, and they take full responsibility for the content of the publication. No AI tool was used to generate experimental results, perform data analysis, design the methodology, or produce the conclusions of this study.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Robin Augustine reports financial support was provided by Uppsala University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Chen Kaile, Abtahi Farhad, Carrero Juan-Jesus, Fernandez-Llatas Carlos, Seoane Fernando. Process mining and data mining applications in the domain of chronic diseases: A systematic review. Artif Intell Med 2023;102645.

[2] Xu Yixian, Zheng Xinkai, Li Yuanjie, Ye Xinmiao, Cheng Hongtao, Wang Hao, Lyu Jun. Exploring patient medication adherence and data mining methods in clinical big data: A contemporary review. J Evidence-Based Med 2023;16(3):342–75.

[3] Sinclair Alan, Dunning Trisha, Rodriguez-Mañas Leocadio. Diabetes in older people: new insights and remaining challenges. Lancet Diabetes Endocrinol 2015;3(4):275–85.

[4] Peng Junjie, Jury Elizabeth C, Dönnes Pierre, Ciurtin Coziana. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Front Pharmacol 2021;12:720694.

[5] Neamtu Bogdan, Negrea Mihai Octavian, Neagu Iuliana. Predicting glycemic control in a small cohort of children with type 1 diabetes using machine learning algorithms. Mathematics 2023;11(20):4388.

[6] Kino Shiho, Hsu Yu-Tien, Shiba Koichiro, Chien Yung-Shin, Mita Carol, Kawachi Ichiro, Daoud Adel. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. SSM-Population Health 2021;15:100836.

[7] EbrahimiZadeh Javad, Maananejad Alireza, Sadeghi Sajjad, Karlsson Roger L, Mandal Bappaditya, Rangaiah Pramod KB, Perez Mauricio D, Augustine Robin. Time reversal microwave imaging of realistic numerical head phantom for bone flap healing follow-up. In: 2020 IEEE MTT-S international microwave biomedical conference. IMBioC, IEEE; 2020, p. 1–4.

[8] Ravikiran HK, Jayanth J, Sathisha MS, Bindu K. Optimizing sheep breed classification with bat algorithm-tuned CNN hyperparameters. SN Comput Sci 2024;5(2):219.

[9] Abnoosian Karlo, Farnoosh Rahman, Behzadi Mohammad Hassan. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics 2023;24(1):337.

[10] Eberhard Karin. The effects of visualization on judgment and decision-making: a systematic literature review. Manag Rev Q 2023;73(1):167–214.

[11] Thakur Deepak, Gera Tanya, Bhardwaj Vivek, AlZubi Ahmad Ali, Ali Farman, Singh Jaiteg. An enhanced diabetes prediction amidst COVID-19 using ensemble models. Front Public Health 2023;11.

[12] Rastogi R, Bansal M. Diabetes prediction model using data mining techniques. Meas: Sens 2023;25:100605. http://dx.doi.org/10.1016/j.measen.2022.100605.

[13] Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020;8:109513–25. http://dx.doi.org/10.1109/access.2020.2989857.

[14] Alex SA, Jhanjhi NZ, Humayun M, Ibrahim AO, Abulfaraj AW. Deep LSTM model for diabetes prediction with class balancing by SMOTE. Electron (Switz) 2022;11(17):2737. http://dx.doi.org/10.3390/electronics11172737.

[15] Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current techniques for diabetes prediction: Review and case study. Appl Sci (Switz) 2019;9(21):4604. http://dx.doi.org/10.3390/app9214604.

[16] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express 2021;7(4):596–602. http://dx.doi.org/10.1016/j.icte.2021.02.004.

[17] Shin J, Lee J, Ko T, Lee K, Choi Y, Kim HS. Improving machine learning diabetes prediction models for the utmost clinical effectiveness. J Pers Med 2022;12(11):1899. http://dx.doi.org/10.3390/jpm12111899.

[18] Gupta H, Varshney H, Sharma TK, Pachauri N, Verma OP. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Complex Intell Syst 2022;8(4):2095–107. http://dx.doi.org/10.1007/s40747-021-00398-7.

[19] Taha AA, Malebary SJ. A hybrid meta-classifier of fuzzy clustering and logistic regression for diabetes prediction. Comput Mater Continua 2022;71(2):2131–45. http://dx.doi.org/10.32604/cmc.2022.023848.

[20] Dutta Aishwariya, Hasan Md Kamrul, Ahmad Mohiuddin, Awal Md Abdul, Islam Md Akhtarul, Masud Mehedi, Meshref Hossam. Early prediction of diabetes using an ensemble of machine learning models. Int J Environ Res Public Health 2022;19(19):12378.

[21] Khan MZ, Mangayarkarasi R, Vanmathi C, Angulakshmi M. Bio-inspired PSO for improving neural based diabetes prediction system. J ICT Stand 2022;10(2):155–68. http://dx.doi.org/10.13052/jicts2245-800x.1025.

[22] Nanri Akiko, Nakagawa Tohru, Kuwahara Keisuke, Yamamoto Shuichiro, Honda Toru, Okazaki Hiroko, Uehara Akihiko, Yamamoto Makoto, Miyamoto Toshiaki, Kochi Takeshi, et al. Development of risk score for predicting 3-year incidence of type 2 diabetes: Japan epidemiology collaboration on occupational health study. PLoS One 2015;10(11):e0142779.

[23] Alby S, Shivakumar BL. A prediction model for type 2 diabetes risk among Indian women. ARPN J Eng Appl Sci 2016;11(3):2037–43.

[24] Wang J, Wang F, Liu Y, Xu J, Lin H, Jia B, Zuo W, Jiang Y, Hu L, Lin F. Multiple linear regression and artificial neural network to predict blood glucose in overweight patients. Exp Clin Endocrinol Diabetes 2016;124(01):34–8.

[25] Chen Long-Sheng, Cai Sheng-Jhe. Neural-network-based resampling method for detecting diabetes mellitus. J Med Biol Eng 2015;35:824–32.

[26] Choi Sung Eun, Brandeau Margaret L, Basu Sanjay. Dynamic treatment selection and modification for personalised blood pressure therapy using a Markov decision process model: a cost-effectiveness analysis. BMJ Open 2017;7(11).

[27] Saidi Olfa, O'Flaherty Martin, Mansour Nadia Ben, Aissi Wafa, Lassoued Olfa, Capewell Simon, Critchley Julia A, Malouche Dhafer, Romdhane Habiba Ben, funded MEDCHAMPS project EC FP7. Forecasting Tunisian type 2 diabetes prevalence to 2027: validation of a simple model. BMC Public Health 2015;15:1–8.

[28] Anggoro Dimas Aryo, Permatasari Dian. Performance comparison of the kernels of support vector machine algorithm for diabetes mellitus classification. Int J Adv Comput Sci Appl 2023;14(1).

[29] Mansoori Amin, Sahranavard Toktam, Hosseini Zeinab Sadat, Soflaei Sara Saffar, Emrani Negar, Nazar Eisa, Gharizadeh Melika, Khorasanchi Zahra, Effati Sohrab, Ghamsary Mark, et al. Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. Sci Rep 2023;13(1):663.

[30] Alzyoud Mazen, Alazaidah Raed, Aljaidi Mohammad, Samara Ghassan, Qasem M, Khalid Muhammad, Al-Shanableh Najah. Diagnosing diabetes mellitus using machine learning techniques. Int J Data Netw Sci 2024;8(1):179–88.

[31] Zrubka Zsombor, Kertész Gábor, Gulácsi László, Czere János, Hölgyesi Áron, Nezhad Hossein Motahari, Mosavi Amir, Kovács Levente, Butte Atul J, Péntek Márta, et al. The reporting quality of machine learning studies on pediatric diabetes mellitus: Systematic review. J Med Internet Res 2024;26(1):e47430.

[32] Dutta Suparna, Mukherjee Saswati, Nag Medha, Majumdar Sujoy, Goyal Ghanshyam. Risk stratification of cardiovascular disease in type 2 diabetes using LDA and CNN for clinical decision management-a multi-centre study in eastern India. Int J Med Eng Informatics 2024;16(1):1–14.

[33] Lee Jeong Eun, Jeon Hyeon-Ju, Lee O-Joun, Lim Hae Gyun. Diagnosis of diabetes mellitus using high frequency ultrasound and convolutional neural network. Ultrasonics 2024;136:107167.

[34] Kee OT, et al. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. Cardiovasc Diabetol 2023;22(1):1–10. http://dx.doi.org/10.1186/s12933-023-01741-7.

[35] Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol 2023;22(1):1–16. http://dx.doi.org/10.1186/s12933-023-01985-3.

[36] Kumar BP Pradeep, Manoj HM. Comparative assessment of machine learning models for predicting glucose intolerance risk. SN Comput Sci 2024;5(7):894.

[37] Tan Kuo Ren, Seng Jun Jie Benjamin, Kwan Yu Heng, Chen Ying Jie, Zainudin Sueziani Binte, Loh Dionne Hui Fang, Liu Nan, Low Lian Leng. Evaluation of machine learning methods developed for prediction of diabetes complications: A systematic review. J Diabetes Sci Technol 2023;17(2):474–89. http://dx.doi.org/10.1177/19322968211056917.

[38] Febrian ME, Ferdinan FX, Sendani GP, Suryanigrum KM, Yunanda R. Diabetes prediction using supervised machine learning. Procedia Comput Sci 2023;216:21–30. http://dx.doi.org/10.1016/j.procs.2022.12.107.

[39] Whig P, Gupta K, Jiwani N, Jupalle H, Kouser S, Alam N. A novel method for diabetes classification and prediction with Pycaret. Microsyst Technol 2023;29(10):1479–87. http://dx.doi.org/10.1007/s00542-023-05473-2.

[40] Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett 2023;10(1–2):1–10. http://dx.doi.org/10.1049/htl2.12039.

[41] Doğru A, Buyrukoğlu S, Arı M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Med Biol Eng Comput 2023;61(3):785–97. http://dx.doi.org/10.1007/s11517-022-02749-z.

[42] Sotoudeh-Anvari M, Sotoudeh-Anvari A, Sotoudeh-Anvari M, Sotoudeh-Anvari A. Setback in ranking fuzzy numbers: a study in fuzzy risk analysis in diabetes prediction. Artif Intell Rev 2022 56:5 2022;56(5):4591–639. http://dx.doi.org/10.1007/s10462-022-10282-6.

[43] Sonar P, Jaya Malini K. Diabetes prediction using different machine learning approaches. In: Proceedings of the 3rd international conference on computing methodologies and communication, ICCMC 2019. 2019, p. 367–71. http://dx.doi.org/10.1109/iccmc.2019.8819841.

[44] Mahboob Alam T, et al. A model for early prediction of diabetes. Inf Med Unlocked 2019;16:100204. http://dx.doi.org/10.1016/j.imu.2019.100204.

[45] Rahi Parvez, Kang Sandeep Singh. A novel paradigm in cardiovascular disease risk prediction through hybrid machine learning. Sist Inzheneriya I Inf Tekhnologii 2025;7(3(22)):48–65.

[46] Tosan Moein, Nourani Vahid, Kisi Ozgur, Dastourani Mehdi. Evolution of ensemble machine learning approaches in water resources management: a review. Earth Sci Inform 2025;18(2):416.

[47] Gul Ghazala, Korejo Imtiaz Ali, Hakro Dil Nawaz, Alqahtani Haitham, Abbasi Abdullah, Babar Muhammad, Al Rahbi Osama, Ali Najma Imtiaz. Machine learning and ensemble methods for cardiovascular disease prediction: A systematic review of approaches, performance trends, and research challenges. Computers 2026;15(1):25.

[48] Chen Raelynn, Ghosh Attri, Hu Jie, Chen Yong, Moore Jason H, Li Ruowang. Beyond P-values: A multi-metric framework for robust feature selection and predictive modeling. 2025, BioRxiv, 2025-2010.

[49] Houssein Essam H, Ibrahim Ibrahim A, Mostafa Amir, Albarrak Abdullah M, Younan Mina. SMENN-hybrid: an efficient technique combining the synthetic minority oversampling technique with ensemble learning for diabetes prediction. Sci Rep 2025.

[50] Riley Richard D, Collins Gary S, Kirton Laura, Snell Kym IE, Ensor Joie, Whittle Rebecca, Dhiman Paula, Van Smeden Maarten, Liu Xiaoxuan, Alderman Joseph, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. Bmj 2025;388.

[51] Iqbal Munir, Husnain Ali, Ashraf Muhammad. Computational intelligence for predicting the strength of high-performance concrete: Ensemble and non-ensemble machine learning framework. Iran J Sci Technol Trans Civ Eng 2025;1–36.

[52] Wang Zeyu, Hong Yuelan, Huang Luying, Zheng Miaocui, Yuan Hongping, Zeng Ruochen. A comprehensive review and future research directions of ensemble learning models for predicting building energy consumption. Energy Build 2025;115589.

[53] Teodorescu Vlad, Obreja Braşoveanu Laura. Assessing the validity of k-fold cross-validation for model selection: Evidence from bankruptcy prediction using random forest and xgboost. Computation 2025;13(5):127.

[54] Wu Yitong, Su Eileen Lee Ming, Wu Mingyu, Ooi Chia Yee, Holderbaum William. A review of machine learning and deep learning trends in EEG-based epileptic seizure prediction. IEEE Access 2025.

[55] Natesh M, Ranjan Kumar HS, Vinutha K, Tapkire Mayura, Sulthana Shazia, Swetha KR, Bharath KN. Improved diabetes detection through integration of external risk factors and machine learning techniques. SN Comput Sci 2025;6(8):1–18.

[56] Zafar Muhammad Mohsin, Khan Zahoor Ali, Javaid Nadeem, Aslam Muhammad, Alrajeh Nabil. From data to diagnosis: A novel deep learning model for early and accurate diabetes prediction. Healthcare 2025;13(17):2138. http://dx.doi.org/10.3390/healthcare13172138.

[57] Singh Yogendra, Tiwari Mahendra. A comprehensive machine learning approach for early detection of diabetes on imbalanced data with missing and outlier values. SN Comput Sci 2025;6(3):213.

[58] Tran Anh T, Zeevi Tal, Payabvash Seyedmehdi. Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging. BioMedInformatics 2025;5(2):20.

[59] Mahmood Faisal. A benchmarking crisis in biomedical machine learning. Nature Med 2025;31(4). 1060–1060.

[60] Asif Sohaib, Wenhui Yi, ur Rehman Saif, ul ain Qurrat, Amjad Kamran, Yueyang Yi, Jinhai Si, Awais Muhammad. Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision. Arch Comput Methods Eng 2025;32(2):853–83.

[61] Cuzzocrea Alfredo, Folino Francesco, Pontieri Luigi, Sabatino Pietro, Samami Maryam. Toward trustworthy and sustainable clinical decision support by training ensembles of specialized logistic regressors. Neural Comput Appl 2025;1–42.

[62] Zhang Danni, Yang Xingyu, Wang Fangying, Qiu Cifang, Chai Yanfu, Fang Danruo. Comparative analysis of feature extraction methods and machine learning models for predicting osteoporosis prevalence. J Med Syst 2025;49(1):72.

[63] Ghani Burooj, Kalkman Vincent J, Planqué Bob, Vellinga Willem-Pier, Gill Lisa, Stowell Dan. Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification. Sci Rep 2025;15(1):16273.

[64] Kale Mr Mohan K. The role of research design in empirical investigations. IIP: World J Humanit Soc Sci 2025;1(Issue-I (April-June)). 9–9.

[65] Hu Xi, Assaad Rayan H. Improving the predictive analytics of machine-learning pipelines for bridge infrastructure asset management applications: An upstream data workflow to address data quality issues in the national bridge inventory database. J Bridge Eng 2024;29(1):04023103.

[66] Lakshmi S, Maheswaran CP. Effective deep learning based grade prediction system using gated recurrent unit (GRU) with feature optimization using analysis of variance (ANOVA). Automatika 2024;65(2):425–40.

[67] Hakkoum Hajar, Idri Ali, Abnane Ibtissam. Global and local interpretability techniques of supervised machine learning black box models for numerical medical data. Eng Appl Artif Intell 2024;131:107829.
[68] Hasan Md Mahmudul, Watling Christopher N, Larue Grégoire S. Validation and interpretation of a multimodal drowsiness detection system using explainable machine learning. Comput Methods Programs Biomed 2024;243:107925.
[69] Yates Luke A, Aandahl Zach, Richards Shane A, Brook Barry W. Cross validation for model selection: a review with examples from ecology. Ecol Monograph 2023;93(1):e1557.
[70] Abimannan Satheesh, El-Alfy El-Sayed M, Chang Yue-Shan, Hussain Shahid, Shukla Saurabh, Satheesh Dhivyadharsini. Ensemble multifeatured deep learning models and applications: A survey. IEEE Access 2023.
[71] Reza Md Shamim, Hafsha Umme, Amin Ruhul, Yasmin Rubia, Ruhi Sabba. Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. Comput Methods Programs Biomed Update 2023;4:100118.
[72] Dash Arabinda, Sethy Prabira Kumar, Behera Santi Kumari. Maize disease identification based on optimized support vector machine using deep feature of DenseNet201. J Agric Food Res 2023;14:100824.
[73] Abdulazeez Ismail, Abba Sani I, Usman Jamilu, Usman Abdullahi G, Aljundi Isam H. Recovery of brine resources through crown-passivated graphene, silicene, and boron nitride nanosheets based on machine-learning structural predictions. ACS Appl Nano Mater 2023;6(24):23207–21.
[74] Rastogi Deependra, Johri Prashant, Tiwari Varun, Elngar Ahmed A. Multi-class classification of brain tumour magnetic resonance images using multi-branch network with inception block and five-fold cross validation deep learning framework. Biomed Signal Process Control 2024;88:105602.
[75] Modak Sandip Kumar Singh, Jha Vijay Kumar. Diabetes prediction model using machine learning techniques. Multimedia Tools Appl 2023;1–27.
[76] Niktabe Sepideh, Lashkari Arash Habibi, Sharma Dilli Prasad. Detection, characterization, and profiling DoH malicious traffic using statistical pattern recognition. Int J Inf Secur 2023;1–24.
[77] Chengula Tumlumbe Juliana, Mwakalonge Judith, Comert Gurcan, Siuhi Saidi. Improving road safety with ensemble learning: Detecting driver anomalies using vehicle inbuilt cameras. Mach Learn Appl 2023;14:100510.
[78] Nagassou Mahadi, Mwangi Ronald Waweru, Nyarige Euna. A hybrid ensemble learning approach utilizing light gradient boosting machine and category boosting model for lifestyle-based prediction of type-II diabetes mellitus. J Data Anal Inf Process 2023;11(4):480–511.
[79] Shehadeh Ali, Alshboul Odey, Al Mamlook Rabia Emhamed, Hamedat Ola. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Autom Constr 2021;129:103827.
[80] Nipa Nurjahan, Riyad Mahmudul Hasan, Satu Shahriare, Howlader Koushik Chandra, Moni Mohammad Ali, et al. Clinically adaptable machine learning model to identify early appreciable features of diabetes in Bangladesh. Intell Med 2023.
[81] Pathy Abhijeet, Meher Saswat, Balasubramanian P. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. Algal Res 2020;50:102006.
[82] Nagarajan Gayathri, Babu LD Dhinesh. Missing data imputation on biomedical data using deeply learned clustering and L2 regularized regression based on symmetric uncertainty. Artif Intell Med 2022;123:102214.
[83] Datta Debabrata, Bhattacharya Madhubrata, Rajest S Suman, Shynu T, Regin R, Priscila S Silvia. Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting. Int J Bioinform Res Appl 2023;19(3):151–69.
[84] Nasimian Ahmad, Younus Saleena, Tatli Özge, Hammarlund Emma U, Pienta Kenneth J, Rönnstrand Lars, Kazi Julhash U. AlphaML: A clear, legible, explainable, transparent, and elucidative binary classification platform for tabular data. Patterns 2024;5(1).
[85] Shree Ramya, Madagaonkar Suraj, Prateek Lakshmi Aashish, Tony Alan, Rathnamma MV, Venkata Ramana V, Chandrasekaran K. Application of ensemble methods in medical diagnosis. In: International conference on innovations in data analytics. Springer; 2022, p. 355–67.
[86] Rao Patike Kiran, Chatterjee Subarna, Nagaraju K, Khan Surbhi B, Almusharraf Ahlam, Alharbi Abdullah I. Fusion of graph and tabular deep learning models for predicting chronic kidney disease. Diagnostics 2023;13(12):1981.
[87] Mustafa Hiri, Mohamed Chrayah, Nabil Ourdani, Noura Aknin. Machine learning techniques for diabetes classification: A comparative study. Int J Adv Comput Sci Appl 2023;14(9).
[88] Koushik Anil NP, Manoj M, Nezamuddin N. Machine learning applications in activity-travel behaviour research: a review. Transp Rev 2020;40(3):288–311.
[89] Guldogan Emek, Yagin Fatma Hilal, Pinar Abdulvahap, Colak Cemil, Kadry Seifedine, Kim Jungeun. A proposed tree-based explainable artificial intelligence approach for the prediction of angina pectoris. Sci Rep 2023;13(1):22189.
[90] Indrasiri Pubudu L, Halgamuge Malka N, Mohammad Azeem. Robust ensemble machine learning model for filtering phishing URLs: Expandable random gradient stacked voting classifier (ERG-svc). IEEE Access 2021;9:150142–61.
[91] Uddin Nazim, Ahamed Md Khabir Uddin, Uddin Md Ashraf, Islam Md Manwarul, Talukder Md Alamin, Aryal Sunil. An ensemble machine learning based bank loan approval predictions system with a smart application. Int J Cogn Comput Eng 2023;4:327–39.
[92] Pang Keona. A comparative study of explainable machine learning models with Shapley values for diabetes prediction. Healthc Anal 2025;7:100390.
[93] Darolia Aman, Chhillar Rajender Singh. Analyzing three predictive algorithms for diabetes mellitus against the pima Indians dataset. ECS Trans 2022;107(1):2697.