Journal of Intelligent Computing and Health Informatics (JICHI). Vol. 6, No. 1, 2025
ISSN: 2715-6923 e-ISSN: 2721-9186, pp.28~35 | DOI: 10.26714/jichi.v6i1.17299

Machine Learning for Early Non-invasive Diabetes Detection Using Electronic Health Records

Suresh Kumar Arumugam*

Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, Uttarakhand 248002, India
*Corresponding author: suresh79.slm@gmail.com

Jason Patterson

Department of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032, USA
E-mail: jp3477@cumc.columbia.edu

Panagiotis Petridis

Department of Electrical and Computer Engineering, School of Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Sara Masoud

Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48202, USA
E-mail: saramasoud@wayne.edu

Article history :
Received: 13 OCT 2024
Accepted: 21 FEB 2025
Available online: 31 MAR 2025

Research article

Abstract: Early detection of Type 2 Diabetes Mellitus (T2DM) is critical to preventing long-term complications such as cardiovascular disease, nephropathy, and retinopathy. However, conventional diagnostic approaches are often invasive, costly, and unsuitable for population-scale screening. This study proposes a non-invasive, machine learning-based framework for early T2DM detection using electronic health records (EHRs) from a publicly available Kaggle dataset. Key non-invasive features including demographics, vital signs, medication history, and temporal health trends were extracted and used to train six classifiers: random forest (RF), support vector machine (SVM), naïve bayes (NB), alternating decision tree (ADT), random tree (RT), and k-nearest neighbors (KNN). Class imbalance was addressed using the synthetic minority over-sampling technique (SMOTE) at 0%, 150%, and 300% levels. Experimental results show that RF achieved the highest AUC (88.45%) at 150% SMOTE, while SVM demonstrated the best sensitivity gains when temporal features and feature selection were applied. The proposed framework demonstrates the potential of interpretable, EHR-based ML models for scalable, cost-effective diabetes screening and offers a reproducible benchmark for future applications in real-world clinical data.

Keywords: TYPE 2 DIABETES MELLITUS; MACHINE LEARNING; ELECTRONIC HEALTH RECORDS; TEMPORAL FEATURES; NON-INVASIVE DETECTION

1. Introduction

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder characterized by insulin resistance and progressive β-cell dysfunction, resulting in persistent hyperglycemia. The global burden of T2DM continues to escalate, with the World Health Organization (2023) reporting approximately 1.5 million deaths directly attributable to diabetes. Furthermore, the Global Burden of Disease Study 2021 indicates a steady increase in T2DM-related incidence and mortality across nearly all demographic regions over the past decade (Global Burden of Disease Collaborative Network, 2024). Long-term complications include cardiovascular disease, nephropathy, neuropathy, and retinopathy, all of which contribute significantly to morbidity and healthcare expenditure. Therefore, early detection and intervention are critical to reducing the progression and impact of this disease.

Conventional diagnostic methods such as fasting blood glucose (FBG), glycated hemoglobin (HbA1c), and oral glucose tolerance tests (OGTT) are effective but invasive, time-consuming, and impractical for large-scale implementation, particularly in low-resource environments (Lin et al., 2020). This underscores the need for alternative, non-invasive approaches that are both scalable and cost-effective.

Electronic Health Records (EHRs) have emerged as a valuable source of patient information, containing structured data on demographics, clinical observations, vital signs, medical history, and medication usage. When analyzed using machine learning (ML) algorithms, EHRs can uncover complex and nonlinear patterns indicative of T2DM risk. This integration of health informatics and artificial intelligence has the potential to support timely, data-driven clinical decisions and enhance disease prevention strategies.

Recent literature supports this direction. According to a comprehensive review by Kiran et al. (2025), machine learning models built exclusively on structured EHR data demonstrate consistently strong performance in T2DM prediction. Their study emphasizes that such unimodal models are particularly valuable when interpretability, accessibility, and integration into existing healthcare infrastructure are prioritized. Similarly, Hennebelle et al. (2024) proposed a smart healthcare architecture combining ML with cloud computing, with Random Forest (RF) outperforming logistic regression by 6% in prediction accuracy. Bernardini et al. (2020) applied support vector machine (SVM) algorithms to clinical datasets and demonstrated that SVMs are capable of achieving high precision in predicting diabetes, especially when trained on structured medical features.

Despite these advancements, critical challenges persist. These include the class imbalance often present in medical datasets, underutilization of temporal features that reflect disease progression, and the lack of external validation necessary for model generalizability. Additionally, many studies rely on proprietary data, limiting reproducibility and comparability across models.

In this study, we aim to develop and evaluate a non-invasive T2DM detection model using a publicly available EHR dataset. The objectives of this research are as follows:

- To compare the predictive performance of multiple machine learning classifiers, including Random Forest, Naïve Bayes, Support Vector Machine, Alternating Decision Tree, Random Tree, and k-Nearest Neighbors.
- To incorporate temporal features such as trends in body mass index (BMI), blood pressure, and weight to enhance model accuracy.
- To address class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) and evaluate its impact on key performance metrics.
- To validate the model's performance using 10-fold cross-validation for improved reliability.

The remainder of this article is structured as follows. Section 2 describes the materials and methods used, including dataset characteristics, preprocessing steps, and modeling techniques. Section 3 presents and discusses the results of our experiments. Section 4 outlines the study's limitations and Section 5 concludes with key findings and directions for future research.

2. Materials and Methods

This study presents a predictive modeling framework for early non-invasive detection of Type 2 Diabetes Mellitus (T2DM) using electronic health records (EHRs). The workflow includes data preprocessing, temporal feature engineering, machine learning (ML) model training, class balancing with SMOTE, and multi-metric performance evaluation. The complete workflow is illustrated in Fig. 1.

2.1 Dataset

The dataset used was obtained from Kaggle and contains anonymized EHR data spanning from 2009 to 2012. It includes a wide range of non-invasive patient attributes, including demographics, vital signs, medical diagnoses, medication use, and lifestyle indicators. Initially, 529 variables were extracted. After feature selection, 12 key predictors were retained for model development. Table 1 summarizes the retained features and their characteristics.

2.2 Data preprocessing

Before training machine learning models, the dataset was preprocessed to address typical issues found in clinical data, including missing values, outliers, and variations in data scale. These steps were essential to ensure that the models could learn from consistent, high-quality inputs and to minimize the risk of bias or overfitting.

a. Handling missing values.

Missing data were handled using imputation methods tailored to the nature of each variable. For numerical features such as body mass index (BMI), systolic and diastolic blood pressure, and weight, missing values were replaced with the median. This method was chosen for its robustness against skewed distributions and outliers, which are prevalent in real-world health data. For categorical variables, such as smoking status and medication usage, the mode the most frequent category was used. These choices are supported by Tabassum et al. (2022), who emphasize that simple, robust imputation strategies are preferable when working with biomedical datasets to avoid introducing artificial bias and to preserve interpretability.

b. Outlier detection and treatment.

Outliers were identified using the Z-score method, which detects values exceeding ±3 standard deviations from the mean. Rather than eliminating these data points which could reduce statistical power and sample representativeness, they were replaced with the feature's median value. This approach is consistent with current recommendations for clinical predictive modeling, which suggest adjusting extreme values to preserve data completeness while mitigating their impact on model training (Nawaz et al., 2024).

c. Normalization

To ensure feature comparability and improve model performance, Z-score normalization was applied to all continuous variables. This method transforms each feature into a standardized distribution with a mean of zero and a standard deviation of one, using the following in Eq. (1).

$$X' = \frac{X-\mu}{\sigma} \quad (1)$$

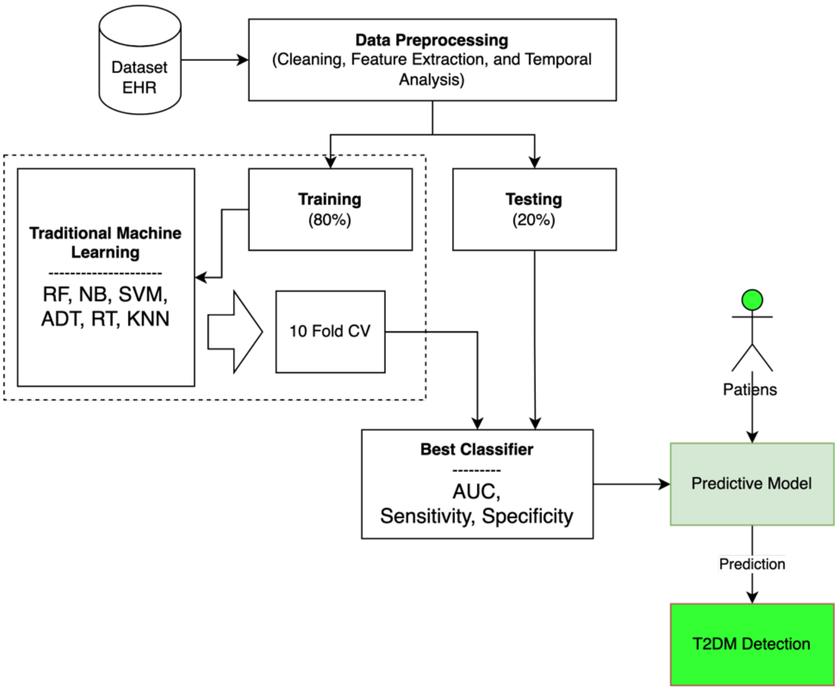
Dataset
EHR
Data Preprocessing
(Cleaning, Feature Extraction, and Temporal Analysis)
Training
(80%)
Testing
(20%)
Traditional Machine Learning
RF, NB, SVM, ADT, RT, KNN
10 Fold CV
Best Classifier
AUC, Sensitivity, Specificity
Patiens
Predictive Model
Prediction
T2DM Detection

Fig 1. The workflow of this study.

Table 1. Summary of key features in dataset.

FEATURES	DESCRIPTION	DATA TYPE	VALUE DOMAIN	COUNT OF DATA
Age	Patient's age at the time of record	Numerical	18–100 years	9.947
Gender	Patient's gender	Categorical	Male, Female	9.947
BMI	Body Mass Index (weight-to-height ratio)	Numerical	10–50 kg/m²	9.460
Systolic_BP	Systolic blood pressure	Numerical	90–200 mmHg	9.320
Diastolic_BP	Diastolic blood pressure	Numerical	50–120 mmHg	9.320
Weight	Patient's weight in kilograms	Numerical	30–200 kg	9.500
Smoking_Status	Smoking history	Categorical	Current, Former, Never	8.750
Hypertension	Presence of hypertension	Categorical	Yes, No	9.947
Dyslipidemia	Presence of abnormal cholesterol levels	Categorical	Yes, No	9.500
Medication_Count	Number of prescribed medications	Numerical	0–15	9.947
Diagnosis_Count	Total number of past medical diagnoses	Numerical	0–50	9.947
Diabetes_Status	Whether the patient was diagnosed with diabetes	Categorical	Yes, No	9.947

where X is the original value, μ is the mean, and σ isthe standard deviation. Z-score normalization was preferred over min–max scaling due to its robustness in handling features with varying ranges and skewed distributions. Moreover, it is well-suited for models such as support vector machines (SVM) and k-nearest neighbors (KNN), which rely on distance or margin-based calculations. Singh and Singh (2021) highlight that normalization techniques significantly influence model performance in healthcare datasets, particularly when features differ in unit scale or variance.

All preprocessing steps were implemented using open-source libraries in Python, including pandas, numpy, and scikit-learn. After preprocessing, the dataset consisted of 9,947 complete and standardized patient records, ready for temporal feature engineering and classifier training.

2.3 Feature engineering

Feature engineering is a crucial step in the development of predictive models, particularly in healthcare, where raw electronic health records (EHRs) often contain heterogeneous, unstructured, or temporally sensitive information. In this study, features were categorized into three groups: temporal clinical indicators, medication-related variables, and demographic and lifestyle attributes.

These were selected based on clinical relevance to type 2 diabetes mellitus (T2DM) and their availability in the dataset.

a. Temporal clinical indicators

To capture longitudinal health dynamics, temporal features were derived from continuous variables such as body mass index (BMI), systolic and diastolic blood pressure, and weight. For each patient, a linear regression was fitted to historical observations, and the slope of the regression line was used as a proxy for temporal trends. This approach enables the model to account for gradual physiological changes rather than relying solely on static values. Studies have shown that incorporating temporal slopes improves early detection of chronic conditions by capturing disease progression patterns not evident in snapshot data (Moglia et al., 2025).

b. Medication-related features

Pharmacological treatment history was included to reflect underlying comorbidities associated with T2DM risk. Features were constructed based on the frequency and category of prescribed medications, including statins, antihypertensives, and ACE inhibitors. Rather than encoding medications as binary indicators, we aggregated prescriptions by class and count to reflect therapeutic intensity. Prior work has demonstrated that such structured medication profiles enhance model performance in EHR-based prediction tasks (Bayramli et al., 2022).

c. Demographic and lifestyle attributes

Demographic variables (e.g., age, gender) and lifestyle-related indicators (e.g., smoking status, hypertension, and dyslipidemia history) were also included. These features represent known risk factors for T2DM and provide essential context for individual susceptibility. For instance, age is one of the most significant predictors of metabolic disorders, and smoking is independently associated with increased insulin resistance and inflammation. Integrating these attributes aligns with findings from large-scale population studies linking lifestyle and comorbidity profiles to diabetes risk (Lee et al., 2025).

d. Feature selection

Following feature construction, irrelevant or redundant features were filtered using the Information Gain Attribute Evaluator, which ranks predictors based on their mutual information with the target variable. Only variables with high information gain were retained. This method is particularly suitable for clinical datasets, where irrelevant features can introduce noise and reduce model robustness. Its use is supported by recent literature on model optimization in healthcare AI (Noroozi et al., 2023).

2.3 Machine learning model training

This study implemented six widely used supervised machine learning (ML) classifiers to identify the most effective model for non-invasive detection of type 2 diabetes mellitus (T2DM) based on electronic health records (EHRs). The selected models represent diverse algorithmic families including probabilistic, tree-based, margin-based, and instance-based approaches to allow comparative performance analysis. Each model was trained under different configurations, including raw and SMOTE-balanced datasets, as well as with and without temporal features and feature selection.

a. Random forest

Random Forest is an ensemble learning method that constructs multiple decision trees using random subsets of features and samples, aggregating their outputs via majority voting. It is robust to overfitting and performs well in high-dimensional datasets with complex, nonlinear relationships. In medical applications, RF has consistently demonstrated high accuracy and resilience to noise and missing values (Fawagreh & Gaber, 2020).

b. Naïve bayes

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, assuming conditional independence between features. Although this assumption rarely holds in clinical data, NB remains effective in many healthcare scenarios due to its simplicity, fast training time, and interpretability (Appasani et al., 2024). For this study, numerical features were discretized where appropriate to improve NB's performance.

c. Support vector machine

SVM is a margin-based classifier that seeks to find an optimal hyperplane separating classes in a high-dimensional feature space. It is particularly effective in datasets with complex decision boundaries and performs well with both linear and nonlinear kernels. In chronic disease detection, SVM is known for its high precision and robustness, especially when integrated with appropriate feature scaling (G et al., 2025).

d. Alternating decision tree

ADT is an extension of traditional decision trees that combines decision and prediction nodes, allowing for multiple contributing paths in classification. It balances accuracy with interpretability and is especially useful in clinical decision support systems where transparency is crucial (Chen et al., 2024).

e. Random tree

The Random Tree model builds a single decision tree using random feature subsets at each node. Although less accurate than ensemble methods like RF, it provides a fast, interpretable baseline for comparison and is useful in evaluating the benefit of ensemble strategies.

f. k-Nearest neighbors

KNN is a non-parametric classifier that assigns class labels based on the majority class among the k closest training instances. It leverages similarity across patients and is effective when relevant features are well-standardized. While computationally intensive at prediction time, KNN can achieve strong performance in clinical settings with limited feature noise and dimensionality reduction (Halder et al., 2024).

g. Hyperparameter tuning and training configuration

Each classifier underwent hyperparameter optimization using grid search combined with 10-fold cross-validation. Parameter ranges were determined empirically based on literature and pilot experiments. For example, the number of estimators in RF ranged from 100 to 500, while SVM kernel types (linear, RBF) were tested along with regularization parameters.

Models were trained using both the original imbalanced dataset and SMOTE-balanced datasets (150% and 300% oversampling). In addition, experiments were conducted with and without temporal features and feature selection, allowing for an ablation-style evaluation of their contributions. All training procedures were implemented using Python's scikit-learn library (v1.3), and results were tracked using reproducible pipelines.

2.4 Model validation and evaluation

To ensure the robustness, reliability, and generalizability of the machine learning models developed in this study, a rigorous validation strategy was implemented. In clinical predictive modeling, model validation is essential to prevent overfitting and to evaluate performance in scenarios that approximate real-world deployment.

a. Cross-validation strategy

All models were validated using a 10-fold cross-validation approach. The dataset was partitioned into 10 equal subsets; in each iteration, one subset was used as the validation set while the remaining nine were used for training. This process was repeated ten times, with each fold serving once as the validation set. The final performance was reported as the mean across all folds. This strategy mitigates variance due to random train-test splits and is a well-established standard in medical machine learning studies (Allgaier & Pryss, 2024).

In addition, a stratified split of 80% training and 20% testing was performed to evaluate the final model on unseen data after training. Stratification ensured that the distribution of diabetic and non-diabetic cases was preserved across splits, which is particularly important in imbalanced clinical datasets.

b. Handling class imbalance

Given that only 19% of instances in the dataset were diabetic cases, Synthetic Minority Over-sampling Technique (SMOTE) was used to address class imbalance. SMOTE was applied at three levels (0%, 150%, and 300%) to generate synthetic examples of the minority class. By augmenting the training data with realistic interpolations, SMOTE reduces the bias toward the majority class and improves sensitivity without simply duplicating instances. Studies have shown that SMOTE enhances recall and F1-score in medical classification tasks with limited positive samples (Hairani et al., 2024).

c. Evaluation metrics

Due to the imbalanced nature of the dataset and the clinical implications of false positives and false negatives, multiple evaluation metrics were used:

- Accuracy. Measures overall correctness but can be misleading in imbalanced datasets.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

- Sensitivity (recall or true positive rate, TPR). Critical for minimizing false negatives, especially important in diabetes detection, where undetected cases may lead to delayed treatment (Gurcan & Soylu, 2024).

$$TPR = \frac{TP}{TP+FN} \tag{3}$$

- Specificity (true negative rate, TNR). Helps ensure that non-diabetic individuals are not misclassified, reducing unnecessary clinical intervention.

$$TNR = \frac{TN}{TN+FP} \tag{4}$$

- Precision (positive predictive value, PPV). Reflects the proportion of predicted diabetic cases that are actually correct.

$$PPV = \frac{TP}{TP+FP} \tag{5}$$

- F1-Score. Balances the trade-off between sensitivity and precision, particularly useful for imbalanced classes.

$$F1 - score = \frac{PPV \times TPR}{PPV+TPR} \tag{6}$$

- AUC-ROC (area under the receiver operating characteristic curve). AUC quantifies the model's ability to distinguish between diabetic and non-diabetic cases across various thresholds. AUC is widely adopted in clinical ML evaluation as it remains unaffected by class distribution and provides an aggregate view of model discrimination capability (Diallo et al., 2025).

All metrics were computed for each SMOTE configuration and model variant (with and without temporal features and feature selection), providing a comprehensive view of model behavior under different conditions.

3. Results and Discussion

This section presents the performance of six machine learning models under various configurations, based on 10-fold cross-validation results. The models were evaluated using AUC, sensitivity, and specificity metrics, as detailed in Table 2 and Table 3.

3.1 AUC performance across models and configurations

As shown in Table 2, random forest (RF) consistently achieved the highest AUC across all SMOTE levels. The best AUC performance was recorded at 150% SMOTE with original data (OD), yielding 88.45%, followed closely by RF at 300% SMOTE (88.20%). RF also remained stable across data configurations with temporal features (TF) and feature selection (FS), indicating its robustness.

Support vector machine (SVM) demonstrated significant improvement in AUC when class balancing was applied. The AUC increased from 67.59% (OD, 0%

SMOTE) to 77.27% (OD, 300% SMOTE), and further improved to 71.69% under the OD+TF+FS configuration, despite the smaller impact from temporal features. These results suggest that SVM is highly responsive to balanced data but less affected by additional engineered features.

Naïve bayes (NB) exhibited stable AUC performance in the 83–84% range under 0% SMOTE but declined under oversampling, particularly at 150% (e.g., 74.78%, OD) and 300% (e.g., 72.95%, OD). This performance drop is likely due to the synthetic variance introduced by SMOTE, which may distort probability assumptions in NB.

KNN and ADT models produced modest AUC values, generally ranging from 63% to 66%, but remained more consistent across feature variations. Random tree (RT) showed the lowest AUC overall, reflecting its known tendency to overfit and its lack of ensemble robustness.

3.2 Sensitivity and specificity trade-off

As summarized in Table 3, RF not only yielded the highest AUC but also maintained excellent specificity across most configurations, up to 99.26% (OD, 0% SMOTE), with relatively low sensitivity (14.19% in the same setting). However, sensitivity improved markedly with SMOTE, reaching 40.45% (OD, 300% SMOTE), albeit with a slight decline in specificity (95.31%). This aligns with prior findings that RF favors specificity unless class imbalance is directly addressed (Zhu et al., 2018).

Table 2. Area under the curve (AUC) results from 10-fold cross-validation across SMOTE levels and feature configurations.

AUC (%)							
SMOTE	DATA	RF	NB	SVM	ADT	RT	KNN
0%	OD	87.38	83.56	67.59	83.81	63.11	66.48
	OD+TF	87.13	83.54	67.93	83.80	64.66	65.01
	OD+TF+FS	87.52	83.72	67.76	83.80	65.17	65.61
150%	OD	88.45	74.78	75.66	83.14	64.29	64.64
	OD+TF	87.32	73.77	68.38	82.81	62.95	61.62
	OD+TF+FS	87.52	81.45	69.71	83.37	65.59	63.83
300%	OD	88.20	72.95	77.27	82.55	64.37	65.68
	OD+TF	87.13	70.72	68.94	81.71	63.17	61.70
	OD+TF+FS	87.24	81.03	71.69	82.50	65.05	64.10

This table presents the AUC values (in percentage) for six machine learning classifiers: random forest (RF), naïve bayes (NB), support vector machine (SVM), alternating decision tree (ADT), random tree (RT), and k-nearest neighbors (KNN) evaluated using original data (OD), with and without temporal features (TF) and feature selection (FS). The performance is reported under three SMOTE levels: 0%, 150%, and 300%. AUC represents the overall discriminative ability of the models across varying thresholds, providing a robust summary of classification performance in imbalanced clinical data settings.

Table 3. Sensitivity and specificity results (mean ± SD) from 10-fold cross-validation for all classifiers under varying SMOTE levels and feature settings.

PERFORMANCE	SMOTE	DATA	RF	NB	SVM	ADT	RT	KNN
Specificity (%)	0%	OD	99.26±0.06	78.75±0.13	97.33±0.12	95.96±0.24	87.84±0.57	92.65±0.19
		OD+TF	99.25±0.09	78.70±0.13	97.15±0.15	95.96±0.24	88.10±0.54	92.84±0.27
		OD+TF+FS	97.37±0.13	79.12±0.10	98.50±0.13	96.27±0.25	88.5±0.33	92.59±0.6
	150%	OD	97.07±0.05	76.83±0.19	88.00± 0.3	87.44±0.96	82.69±0.13	85.40±0.30
		OD+TF	99.06±0.09	77.40±0.74	95.50±0.29	94.02±0.45	85.30±0.48	88.91±0.29
		OD+TF+FS	96.83 ± 0.4	83.72±0.11	93.07±0.35	91.35±0.63	84.85±0.28	85.05±0.38
	300%	OD	95.31 ± 0.4	58.10±0.40	80.57±0.31	80.65±0.95	81.09±0.5	82.95±0.19
		OD+TF	98.95 ± 0.4	60.58±1.93	94.50±0.41	93.64±0.64	85.25±0.27	88.08±0.46
		OD+TF+FS	96.13 ± 0.4	84.06±0.26	89.69±0.51	88.54±0.50	84.2 ± 0.46	82.96±0.12
Sensitivity (%)	0%	OD	14.19±0.29	67.98±2.26	35.86±2.51	32.01±2.88	33.95±2.49	27.34 ± 2.49
		OD+TF	12.84±0.35	68.42±2.21	36.73±2.51	32.01±2.88	34.33±2.49	25.75 ± 2.78
		OD+TF+FS	20.5±0.59	68.67±2.09	36.28±2.55	32.01±2.88	37.15±2.86	26.31 ± 2.28
	150%	OD	32.54±0.44	64.39±3.10	61.34±2.33	51.55±3.65	43.77±2.96	41.88 ± 2.80
		OD+TF	16.31±0.32	62.89±3.08	39.28±3.14	36.41±4.22	37.07±3.63	31.97 ± 2.97
		OD+TF+FS	29.45±0.31	58.17±2.38	44.37±2.43	43.47±4.09	42.22±3.16	40.80 ± 2.32
	300%	OD	40.45±0.51	76.08±2.57	72.00±2.38	63.58±3.31	45.69±2.86	46.27 ± 2.81

PERFORMANCE	SMOTE	DATA	RF	NB	SVM	ADT	RT	KNN
		OD+TF	16.31±0.73	69.89±3.02	41.39±3.39	33.56±4.57	36.93±3.39	33.02 ± 2.77
		OD+TF+FS	31.55±0.47	57.98±2.73	53.70±2.68	50.38±3.07	41.85±2.77	43.41 ± 2.56

This table reports the average sensitivity and specificity values (in percentage, with standard deviation) for six classifiers under the same experimental settings as Table 2. Sensitivity indicates the model's ability to correctly identify diabetic cases, while specificity reflects its accuracy in identifying non-diabetic individuals. Results are organized across three SMOTE levels (0%, 150%, and 300%) and three data configurations (OD, OD+TF, OD+TF+FS). This table highlights the trade-offs between sensitivity and specificity, and the influence of oversampling and feature engineering on class-specific prediction performance.

SVM displayed the greatest sensitivity improvement, jumping from 35.86% (OD, 0%) to 72.00% (OD, 300%), with corresponding specificity declining from 97.33% to 80.57%. This illustrates the classic trade-off between recall and precision in imbalanced settings, where oversampling improves minority detection but may introduce false positives (Ilham et al., 2024).

NB had the highest baseline sensitivity, achieving 67.98% (OD, 0%), but its specificity remained lower (78.75%), particularly under aggressive SMOTE (58.10%, OD, 300%). ADT and RT achieved balanced metrics under moderate oversampling but lagged in sensitivity, generally not exceeding 63.58% (ADT, OD, 300%).

Interestingly, temporal features did not universally improve sensitivity. For instance, RF and SVM had better performance without TF in many SMOTE settings, while KNN and ADT gained marginal benefit from TF and FS. This suggests that the value of engineered features is model-dependent, and their effectiveness may vary based on model architecture and class distribution.

3.3 Clinical relevance and implications

From a clinical perspective, high sensitivity is crucial to reduce undiagnosed cases of T2DM, especially in primary detection scenarios. While RF offers the best balance between specificity and AUC, SVM may be preferable when recall is the priority such as in early-risk alerts or population-level detection systems. However, caution is warranted due to the corresponding drop in specificity and the risk of overdiagnosis.

Models such as ADT and RT, despite their lower performance, offer greater interpretability, which can support clinician trust and explainable AI integration. Meanwhile, KNN's reliance on patient similarity metrics may be useful in cohort-matching applications, though computational cost remains a concern.

Overall, these findings support the feasibility of implementing non-invasive, EHR-based predictive models using open-access datasets and widely available ML techniques. Careful tuning of class balancing and feature engineering is essential to adapt models to specific clinical contexts.

5. Conclusion

This study proposed a machine learning-based framework for the early and non-invasive detection of type 2 diabetes mellitus (T2DM) using structured electronic health record (EHR) data. By combining temporal health features, feature selection, and synthetic class balancing (SMOTE), six widely used classifiers were evaluated to identify the most effective predictive model. Random Forest (RF) achieved the highest area under the curve (AUC) and specificity, while Support Vector Machine (SVM) showed the greatest improvement in sensitivity under aggressive oversampling.

Temporal feature engineering, particularly trends in BMI, blood pressure, and weight significantly enhanced model sensitivity in SVM, KNN, and ADT classifiers. Feature selection using information gain further improved generalizability and reduced redundancy, particularly benefiting distance- and probability-based models. The experiments also revealed that the effectiveness of these enhancements was model-specific, highlighting the need for tailored configurations based on clinical priorities such as sensitivity or interpretability.

The proposed framework demonstrates that predictive detection for T2DM is feasible using routinely collected EHR data, even in the absence of invasive laboratory markers. This can support population-level risk stratification and early intervention, especially in low-resource healthcare environments. Future research should focus on validating the models using external datasets, improving explainability, and integrating the system into real-world clinical workflows for broader adoption.

Author Contributions

Suresh Kumar Arumugam conceived and designed the study, implemented the machine learning models, and led data preprocessing and manuscript drafting. Jason Patterson contributed to experimental design, model evaluation, and methodological refinement. Panagiotis Petridis supported algorithmic optimization and parameter tuning, ensuring technical robustness. Sara Masoud conducted the literature review, interpreted results from a biomedical informatics perspective, and enhanced the clarity and cohesion of the manuscript. All authors critically revised, reviewed, and approved the final version of the manuscript.

Acknowledgements

The authors would like to express their sincere appreciation to Graphic Era (Deemed to be University), Columbia University Medical Center, Aristotle University of Thessaloniki, and Wayne State University for their academic support and research environment that made this study possible. The collaboration among institutions significantly contributed to the interdisciplinary quality and international scope of this research.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

References

Allgaier, J., & Pryss, R. (2024). Cross-Validation Visualized: A Narrative Guide to Advanced Methods. *Machine Learning and Knowledge Extraction*, *6*(2), 1378–1388. https://doi.org/10.3390/make6020065

Appasani, D., Bokkisam, C. S., & Surendran, S. (2024). An Incremental Naive Bayes Learner for Real-time Health Prediction. *Procedia Computer Science*, *235*, 2942–2954. https://doi.org/10.1016/j.procs.2024.04.278

Bayramli, I., Castro, V., Barak-Corren, Y., Madsen, E. M., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2022). Predictive structured–unstructured interactions in EHR models: A case study of suicide prediction. *Npj Digital Medicine*, *5*(1), 15. https://doi.org/10.1038/s41746-022-00558-0

Bernardini, M., Romeo, L., Misericordia, P., & Frontoni, E. (2020). Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 235–246. https://doi.org/10.1109/JBHI.2019.2899218

Chen, Z., Tang, J., & Song, D. (2024). Modeling landslide susceptibility using alternating decision tree and support vector. *Terrestrial, Atmospheric and Oceanic Sciences*, *35*(1), 12. https://doi.org/10.1007/s44195-024-00074-6

Diallo, R., Edalo, C., & Awe, O. O. (2025). *Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score* (pp. 283–312). https://doi.org/10.1007/978-3-031-72215-8_12

Fawagreh, K., & Gaber, M. M. (2020). Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach. *Computing*, *102*(5), 1187–1198. https://doi.org/10.1007/s00607-019-00785-6

G, K., K P, I., Hasin A, J., M, L. F. J., Siluvai, S., & G, K. (2025). Support Vector Machines: A Literature Review on Their Application in Analyzing Mass Data for Public Health. *Cureus*. https://doi.org/10.7759/cureus.77169

Global Burden of Disease Collaborative Network. (2024, April 3). *Global Burden of Disease Study 2021: Results*. Institute for Health Metrics and Evaluation.

Gurcan, F., & Soylu, A. (2024). Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers*, *16*(19), 3417. https://doi.org/10.3390/cancers16193417

Hairani, H., Widiyaningtyas, T., & Dwi Prasetya, D. (2024). Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies. *JOIV : International Journal on Informatics Visualization*, *8*(3), 1310. https://doi.org/10.62527/joiv.8.3.2283

Halder, R. K., Uddin, M. N., Uddin, Md. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1), 113. https://doi.org/10.1186/s40537-024-00973-y

Hennebelle, A., Dieng, Q., Ismail, L., & Buyya, R. (2024). SmartEdge: Smart Healthcare End-to-End Integrated Edge and Cloud Computing System for Diabetes Prediction Enabled by Ensemble Machine Learning. *2024 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 127–134. https://doi.org/10.1109/CloudCom62794.2024.00031

Ilham, A., Kindarto, A., Fathurohman, A., Khikmah, L., Dias Ramadhani, R., Abdunnasir Jawad, S., April Liana, D., Amylia. AR, A., Kareem Oleiwi, A., & Mutiar, A. (2024). CFCM-SMOTE: A Robust Fetal Health Classification to Improve Precision Modeling in Multiclass Scenarios. *International Journal of Computing and Digital Systems*, *15*(1), 471–486. https://doi.org/10.12785/ijcds/160137

Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscionek, B., & Russell, D. (2025). Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1557467

Lee, H., Hwang, S. H., Park, S., Choi, Y., Lee, S., Park, J., Son, Y., Kim, H. J., Kim, S., Oh, J., Smith, L., Pizzol, D., Rhee, S. Y., Sang, H., Lee, J., & Yon, D. K. (2025). Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts from South Korea, Japan, and the UK: a model development and validation study. *EClinicalMedicine*, *80*, 103069. https://doi.org/10.1016/j.eclinm.2025.103069

Lin, H.-C., Kuo, Y.-C., & Liu, M.-Y. (2020). A health informatics transformation model based on intelligent cloud computing – exemplified by type 2 diabetes mellitus with related cardiovascular diseases. *Computer Methods and Programs in Biomedicine*, *191*(2), 105409. https://doi.org/10.1016/j.cmpb.2020.105409

Moglia, V., Johnson, O., Cook, G., de Kamps, M., & Smith, L. (2025). Artificial intelligence methods applied to longitudinal data from electronic health records for prediction of cancer: a scoping review. *BMC Medical Research Methodology*, *25*(1), 24. https://doi.org/10.1186/s12874-025-02473-w

Nawaz, A., Khan, S. S., & Ahmad, A. (2024). Ensemble of Autoencoders for Anomaly Detection in Biomedical Data: A Narrative Review. *IEEE Access*, *12*, 17273–17289. https://doi.org/10.1109/ACCESS.2024.3360691

Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports*, *13*(1), 22588. https://doi.org/10.1038/s41598-023-49962-w

Singh, N., & Singh, P. (2021). Exploring the effect of normalization on medical data classification. *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, 1–5. https://doi.org/10.1109/AIMV53313.2021.9670938

Tabassum, S., Abedin, N., Maruf, R. I., Taufiq Ahmed, M., & Ahmed, A. (2022). Improving Health Status Prediction by Applying Appropriate Missing Value Imputation Technique. *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, 345–348. https://doi.org/10.1109/LifeTech53646.2022.9754794

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data. *IEEE Access*, *6*, 4641–4652. https://doi.org/10.1109/ACCESS.2018.2789428