Journal of Machine Intelligence in Healthcare

E-ISSN: 3090-9503
P-ISSN: 3090-9511

(Research) Article

Improving Early Detection of Type 2 Diabetes from Primary Care Records with Sparse-Balanced SVM

Umna Iftikhar [1],*, and Theodora Chatzinikolaou [2]

[1] Faculty of Engineering Science and Technology, Iqra University, Karachi, Pakistan; e-mail : yamnaiftikhar@gmail.com
[2] The Data Mining and Analytics research group, School of Science and Technology, International Hellenic University, Thermi, Greece
* Corresponding Author : Umna Iftikhar

Abstract: Early detection of Type 2 Diabetes (T2D) from primary-care Electronic Health Records (EHRs) is challenged by high-dimensional features, class imbalance, and limited model interpretability. We introduce a Sparse-Balanced Support Vector Machine (SB-SVM) that combines sparsity-promoting regularization with class-dependent weighting to enhance detection of the minority class while maintaining clinical interpretability. Using the FIMMG primary-care EHR dataset from Italian general practitioners, we tested SB-SVM in three progressively complex scenarios. We compared it with linear/gaussian SVM, KNN, Decision Tree, Random Forest, and deep models (MLP, DBN). Performance was evaluated using stratified cross-validation, with AUC and recall reported. Sparsity was measured using the l_0 norm. Training, validation, and testing efficiency were analyzed. SB-SVM achieved mean AUCs of 0.91, 0.81, and 0.69 across the three cases, with higher recall than most baselines. Gains in recall and AUC were statistically significant compared to most competitors ($p < 0.05$), though differences with Decision Tree and Random Forest were not always significant. The model produced sparse, interpretable coefficients ($l_0 = 0.39, 0.91, 0.57$), consistently highlighting clinically relevant predictors (e.g., HbA1c, age, renal function, hypertension, and antidiabetic prescriptions). SB-SVM also showed lower runtime than ensemble and deep models, supporting real-time applications. By combining class balancing and sparsity within a linear margin-based classifier, SB-SVM offers accurate, interpretable, and computationally efficient T2D risk prediction suitable for integration into Clinical Decision Support Systems in primary care.

Keywords: Type 2 diabetes; Electronic health records; Imbalanced learning; Sparse SVM; Interpretability; Clinical decision support; Primary care data

Received: June 14, 2025
Revised: July 6, 2025
Accepted: August 17, 2025
Published: August 31, 2025
Curr. Ver.: August 31, 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Type 2 Diabetes (T2D) represents one of the most pressing challenges in global healthcare, affecting an estimated 537 million individuals worldwide in 2021, with projections reaching 783 million by 2045 [1]. Early detection of T2D is critical to preventing severe complications such as cardiovascular disease, kidney failure, and premature mortality.

The widespread adoption of Electronic Health Records (EHRs) in primary care has created new opportunities for predictive analytics, enabling the development of machine learning (ML) models that leverage routinely collected patient information for timely diagnosis and risk stratification [2]. Several ML methods have been proposed for T2D prediction using EHR data, including Logistic Regression, Random Forest, and Deep Neural Networks. Systematic reviews highlight that while tree-based and deep learning models often achieve strong predictive performance, they face challenges regarding interpretability and data heterogeneity [3], [5]. Logistic Regression remains widely used for its simplicity and transparency but struggles with complex, nonlinear relationships [4].

Another recurring challenge in this field is class imbalance, as positive T2D cases are often underrepresented in clinical datasets. This imbalance typically results in poor recall for high-risk patients, reducing the clinical utility of many predictive models [6], [7]. Therefore,

DOI : https://doi.org/10.70062/jmih.v1i2.168

there is a growing need for approaches that strike a balance between predictive performance, interpretability, and fairness in handling minority classes.

To address these gaps, this study proposes the Sparse-Balanced Support Vector Machine (SB-SVM), a novel classification framework that integrates sparsity-driven feature selection with a balancing mechanism to mitigate the effects of class imbalance. Unlike standard SVMs, SB-SVM enhances interpretability by highlighting the most discriminative predictors while ensuring robust generalization in heterogeneous EHR data. The main contributions of this work are as follows:

1. We introduce the SB-SVM, a novel variant of SVM that combines sparsity and balancing strategies to enhance predictive performance in unbalanced healthcare datasets.
2. We validate the approach on the FIMMG dataset, a real-world EHR database from Italian general practitioners, demonstrating the practical utility of the method in primary care.
3. SB-SVM outperforms conventional ML models (e.g., Logistic Regression, Random Forest, Deep Neural Networks) in terms of recall and AUC, particularly in detecting minority T2D cases.
4. Feature importance analysis provides clinically meaningful insights, supporting physicians in understanding model predictions.

The remainder of this paper is organized as follows. Section II reviews related work on machine learning approaches for T2D prediction from EHR data. Section III describes the dataset, preprocessing pipeline, and proposed methodology. Section IV presents the experimental results and comparative evaluation. Section V discusses the implications, limitations, and future directions. Finally, Section VI concludes the paper.

2. Related Work

Research on predicting Type 2 Diabetes (T2D) using Electronic Health Records (EHRs) has progressed rapidly with the growing availability of clinical data in primary care. Early efforts commonly employed Logistic Regression (LR) due to its interpretability and strong baseline performance, but LR struggles to capture nonlinear relationships and complex feature interactions [4].

Tree-based ensemble methods such as Random Forest (RF) and boosting algorithms have been widely applied to T2D prediction, consistently outperforming linear models in terms of accuracy and robustness. Systematic reviews confirm their effectiveness, although these methods often require extensive tuning and remain limited in interpretability [3], [5], [13], [14].

Deep learning (DL) models, including multilayer perceptrons and recurrent architectures, have been applied to longitudinal and multimodal EHRs for T2D onset prediction. These approaches demonstrate competitive predictive power but are constrained by their "black-box" nature and high computational requirements, which limit adoption in primary care [3], [5], [10]–[12].

To address these limitations, Explainable AI (XAI) techniques have been increasingly integrated into diabetes prediction workflows. Methods such as SHAP and LIME provide feature-level explanations that enhance model transparency and improve clinician trust [16]–[20]. Despite these advances, reviews emphasize that reliability, fairness, and auditability remain essential challenges for clinical translation [19], [20].

Another critical challenge is class imbalance in EHR datasets, where T2D cases are typically underrepresented. Oversampling strategies such as the Synthetic Minority Over-sampling Technique (SMOTE) have been shown to improve sensitivity and AUC across various classifiers but may reduce specificity, highlighting a performance trade-off [6]–[8], [15].

These insights motivate the need for models that balance predictive power, interpretability, and fairness. In this context, sparse learning methods have drawn increasing attention, as embedded feature selection enables clinically relevant predictors to be highlighted while reducing model complexity [5]. Building on these advances, the proposed Sparse-Balanced Support Vector Machine (SB-SVM) integrates sparsity-driven feature selection with class balancing, aiming to deliver robust, interpretable, and efficient T2D prediction in real-world primary care settings.

3. Proposed Method

In this section, the **Sparse-Balanced Support Vector Machine (SB-SVM),** integrating sparsity-inducing regularization and class-dependent weighting, is introduced. This design addresses two key challenges in EHR-based T2D prediction: (i) high-dimensional, heterogeneous features and (ii) severe class imbalance between diabetic and non-diabetic patients.

3.1. Model Formulation

Predictive modeling for Type 2 Diabetes (T2D) in real-world primary care data poses two fundamental challenges: (i) the high-dimensionality and heterogeneity of Electronic Health Records (EHRs), and (ii) the class imbalance between positive and negative cases. Conventional Support Vector Machines (SVMs) are attractive due to their strong generalization properties and solid theoretical foundations, yet they struggle in these conditions. In particular, standard SVMs do not include an embedded mechanism for feature selection, making them prone to overfitting when irrelevant or redundant clinical variables are present. Moreover, the default formulation assumes balanced class distributions, which can bias decision boundaries toward the majority class and lead to reduced recall in minority populations, precisely where accurate detection is most clinically critical.

To address these limitations, we propose the Sparse-Balanced Support Vector Machine (SB-SVM). This extension augments the conventional SVM objective with two additional components: an L_1-norm penalty to enforce sparsity in the model coefficients, and class-dependent weights to mitigate imbalance.

The classical SVM primal optimization is defined as:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \tag{1}$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \geq 0, \quad i = 1, \dots, n$$

where $\boldsymbol{w \epsilon} \mathbb{R}^{\boldsymbol{d}}$ is the weight vector, $\boldsymbol{b\epsilon}\mathbb{R}$ the bias term, and $\boldsymbol{\xi_i}$ the slack variable penalizing violations of the margin. The parameter $C > 0$ controls the trade-off between maximizing the margin and minimizing classification errors.

The proposed SB-SVM modifies this objective as follows:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + \lambda||w||_1 + C\sum_{i=1}^{n} w_{y_i}\xi_i \tag{2}$$

Here, the additional term $\boldsymbol{\lambda||w||_1}$ encourages sparsity, reducing many coefficients in www to zero. This mechanism implicitly performs feature selection, retaining only those clinical variables that contribute most strongly to the classification task. Such sparsity is particularly advantageous in EHR-based prediction, where the number of features (e.g., laboratory values, medication codes, comorbidities) can be large relative to the sample size, and interpretability is an essential requirement.

The second modification is the introduction of class-specific weights $\boldsymbol{w_{y_i}}$, defined as:

$$w_+ = \frac{n}{2_{n_+}}, \quad w_- = \frac{n}{2_{n_-}} \tag{3}$$

where $\boldsymbol{n_+}$ and $\boldsymbol{n_-}$ represent the number of positive (diabetic) and negative (non-diabetic) patients, respectively. This formulation ensures that misclassification of minority class samples (i.e., diabetic patients) is penalized more heavily, thereby improving recall without excessively sacrificing specificity.

The final decision function is expressed as:

$$\hat{y}(x) = \mathbf{sign}(w \cdot x + b) \tag{4}$$

The combination of margin maximization, sparsity, and class balancing provides a principled framework for predictive modeling in imbalanced, high-dimensional EHR data. Importantly, the sparse solution offers a clinically interpretable set of predictors, as non-zero coefficients in $\boldsymbol{w}$ can be directly mapped to meaningful features such as HbA1c levels, fasting glucose, or prescribed antidiabetic medications. This property distinguishes SB-SVM from deep learning or ensemble approaches, which, although powerful, are typically less transparent and harder to integrate into clinical workflows.

In summary, SB-SVM retains the strengths of conventional SVMs while directly addressing two of the most persistent challenges in predictive medicine: overfitting due to irrelevant features and bias due to class imbalance.

3.2. Algorithm

The optimization procedure for the proposed Sparse-Balanced SVM (SB-SVM) can be summarized in the algorithm below. The design follows a standard supervised learning pipeline, with preprocessing, weight computation, sparse optimization, and final classification.

Algorithm 1. Sparse-Balanced Support Vector Machine

```
INPUT: Training data $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{d}}$, labels $\boldsymbol{y} \in \{-\boldsymbol{1}, +\boldsymbol{1}\}$; sparsity parameter $\boldsymbol{\lambda}$; regularization parameter $\boldsymbol{C}$
OUTPUT: Sparse weight vector $\boldsymbol{w}$, bias term $\boldsymbol{b}$.
1:  Data preprocessing:
      -  Normalize continuous variables (z-score scaling).
      -  Encode categorical variables (one-hot encoding).
      -  Handle missing values (median imputation for continuous, mode for categorical).
2:  Compute class weight:
        $w_+ = \frac{n}{2n_+}, \quad w_- = \frac{n}{2n_-}$
    where $n_+$ and $n_-$ denote the number of positive and negative samples.
3:  Formulate optimization problem using Eq. 2.
4:  Optimization:
      -  Use quadratic programming with L1-penalty or iterative coordinate descent to solve for ($\boldsymbol{w}, \boldsymbol{b}$).
      -  Apply convergence criteria (tolerance $< 10^{-4}$ or max iterations = 1000).
5:  Prediction:
    For a new sample $\boldsymbol{x}$, predict class label using Eq. (4).
6:  Interpretation:
    Rank clinical features according to the magnitude of non-zero coefficients in $\boldsymbol{w}$.
```

3.3. Computational Complexity Analysis

The computational complexity of the proposed Sparse-Balanced Support Vector Machine (SB-SVM) depends primarily on two components: (i) optimization of the primal problem with sparsity regularization, and (ii) class weighting adjustments.

3.3.1. Training Complexity

In the standard linear SVM solved via quadratic programming, the training complexity is approximately:

$$\boldsymbol{\mathcal{O}(n^2 d)},$$

where $\boldsymbol{n}$ is the number of samples and $\boldsymbol{d}$ is the number of features. This cost arises from kernel matrix computations and constraint handling in the optimization.

By introducing a $\boldsymbol{L_1}$-penalty for sparsity, the optimization problem becomes similar to a LASSO-regularized SVM. Efficient solvers such as coordinate descent or proximal gradient methods can reduce the training cost to:

$$\boldsymbol{\mathcal{O}(nd \cdot T)},$$

where $\boldsymbol{T}$ is the number of iterations until convergence (typically much smaller than $\boldsymbol{n}$). Since many coefficients in $\boldsymbol{w}$ are driven to zero, the effective dimensionality $\boldsymbol{d'(d' \ll d)}$ decreases during training, further reducing the computational burden in later iterations.

The class balancing mechanism only **affects the loss function weighting and does** not alter the asymptotic complexity. Its overhead is linear in n, i.e., $\boldsymbol{\mathcal{O}(n)}$.

Thus, the overall training complexity of SB-SVM can be summarized as:

$$\boldsymbol{\mathcal{O}'(nd' \cdot T)}, \qquad \text{with } \boldsymbol{d' \ll d}, \tag{5}$$

which is more efficient in practice than standard SVM, particularly in high-dimensional EHR data.

3.3.2. Prediction Complexity

Once trained, prediction for a new sample $\boldsymbol{x}$ involves computing the dot product $\boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b}$, whose cost is:

$$\boldsymbol{\mathcal{O}}(d')$$

Since d' is small due to sparsity, SB-SVM enables **fast inference**, which is crucial in primary care applications where real-time decision support is desirable.

3.3.3. Memory Complexity

The memory requirement of SB-SVM is dominated by storing the weight vector $w \in \mathbb{R}^{d'}$ and the support vectors. In practice, sparsity reduces storage from $\boldsymbol{\mathcal{O}}(d)$ to $\boldsymbol{\mathcal{O}}(d')$. This provides a significant advantage over deep neural networks, which typically require millions of parameters and GPU memory.

3.4. Advantages of SB-SVM

The Sparse-Balanced Support Vector Machine (SB-SVM) offers several methodological and practical advantages over conventional classifiers, particularly in the context of Electronic Health Records (EHR)-based prediction for Type 2 Diabetes (T2D). Its design explicitly addresses four aspects critical for predictive medicine: feature selection, imbalance-aware learning, interpretability, and computational efficiency.

3.4.1. Feature Selection through Sparsity

As formulated in Eq. (2), the L_1-penalty term induces sparsity in the solution $\boldsymbol{w}^*$. According to the Karush–Kuhn–Tucker (KKT) conditions for optimality, if the absolute gradient of the loss with respect to feature j satisfies:

$$\left|\boldsymbol{\nabla_j J(w)}\right| < \boldsymbol{\lambda},$$

then the corresponding coefficient becomes zero, i.e., $\boldsymbol{w_j^* = 0}$. This property ensures that only features with sufficiently strong contributions remain in the final model. Consequently, the effective dimensionality is reduced from $\boldsymbol{d}$ to $\boldsymbol{d'}$, where:

$$\boldsymbol{d' = \left|\{j : w_j^* \neq 0\}\right|},$$

with $\boldsymbol{d' \ll d}$ in practice. This embedded feature selection improves generalization and enhances clinical interpretability by focusing on a small subset of relevant variables.

3.4.2. Imbalance-Aware Learning

The reweighting scheme in Eq. (3) modifies the hinge loss in Eq. (2) by penalizing errors in the minority class more strongly. Geometrically, this shifts the decision boundary closer to the majority class, thereby increasing the margin for minority samples. Specifically, the modified margin can be expressed as:

$$\boldsymbol{\gamma = \frac{1}{\|w\|} \cdot \min_i \left(w_{y_i} \cdot y_i(w \cdot x_i + b)\right)},$$

where $\boldsymbol{w_{y_i}}$ acts as a scaling factor that balances the contributions of each class. This formulation increases recall for diabetic patients, $y = +1$, without excessively reducing specificity, aligning the classifier with clinical priorities.

3.4.3. Interpretability of Model Coefficients

Unlike ensemble or deep learning methods, the SB-SVM decision function (Eq. (4)) remains linear in form:

$$\boldsymbol{f(x) = w \cdot x + b}$$

The sparsity constraint ensures that only a limited number of coefficients are non-zero. Ranking features by $\left\|\boldsymbol{w_j}\right\|$ directly provides an ordered list of predictors, which can be mapped to clinical factors such as HbA1c, systolic blood pressure, or prescribed antidiabetic medications. Thus, interpretability is an inherent outcome of the optimization process, not an external post-hoc adjustment.

3.4.4. Computational Efficiency

From Eq. (2), the addition of an L_1-penalty allows the use of efficient solvers such as coordinate descent. The per-iteration complexity scales as:

$$\boldsymbol{\mathcal{O}(nd')}$$

with convergence typically reached in $\boldsymbol{T}$ iterations, giving an overall cost of $\boldsymbol{\mathcal{O}(nd'\mathrm{T})}$. Since $\boldsymbol{d' \ll d}$, this reduces both runtime and memory requirements compared to conventional SVM training $(\boldsymbol{\mathcal{O}(n^2 d)})$.

For prediction, classification of a new patient record involves evaluating Eq. (4), which requires only $\boldsymbol{\mathcal{O}(d')}$ operations. This makes SB-SVM suitable for real-time deployment in clinical decision support systems, unlike deep neural networks where inference scales with the number of layers and hidden units.

4. Experimental Setup

4.1. Dataset Description

We employed the FIMMG dataset, a longitudinal primary care Electronic Health Record (EHR) collection from Italian general practitioners. The dataset comprises approximately ≈5,000 patients with up to 10 years of medical history, covering demographic information, vital signs, laboratory measurements, prescriptions, and comorbidities. The classification task was the detection of Type 2 Diabetes (T2D).

Three experimental cases were defined to assess robustness:

- Case I: All patients and all features included.
- Case II: Reduced feature set after removing potentially confounding variables.
- Case III: Patients stratified by age to increase task difficulty.

4.2. Comparative Methods

To provide a rigorous benchmark, the proposed SB-SVM was evaluated against a diverse set of widely used machine learning classifiers, including both linear and non-linear approaches. Classical Linear SVM and Gaussian SVM were chosen to represent margin-based methods with and without kernel expansion, while Decision Tree (DT) and Random Forest (RF) represented the family of tree-based ensemble models commonly used in clinical prediction tasks. To further ensure robust comparisons, we included K-Nearest Neighbors (KNN) as an instance-based learner and Multilayer Perceptron (MLP) along with Deep Belief Network (DBN) as representatives of neural network architectures often applied in EHR-based modeling. This comprehensive set of baselines covers the methodological spectrum from interpretable models to high-capacity learners. For all methods, hyperparameters were systematically optimized via grid search within a nested cross-validation process, employing stratified fivefold splits to maintain the original class imbalance across training and testing sets. This design ensured each model was evaluated under comparable conditions, reducing potential bias and enabling fair performance comparisons. By adopting this thorough evaluation framework, the study highlights the empirical advantages of SB-SVM. It positions its performance relative to well-established, state-of-the-art alternatives widely recognized in healthcare machine learning literature.

4.3. Performance Metrics

Performance evaluation was conducted using a comprehensive set of metrics designed to capture both predictive accuracy and practical usability. Discriminative ability was measured using the Area Under the ROC Curve (AUC), which provides a robust summary of the sensitivity–specificity trade-offs across various thresholds. To account for the clinical importance of early identification, particular emphasis was placed on Recall (Sensitivity), as false negatives in T2D prediction may delay timely intervention. Complementing recall, the F1-score was reported to balance sensitivity and precision, thus reflecting the trade-off between over-diagnosis and under-diagnosis in practice. Beyond predictive accuracy, we incorporated the sparsity measure $l_0 = \frac{\|w\|_0}{d}$, quantifying the fraction of non-zero coefficients in the model and thereby indicating the degree of feature selection achieved by SB-SVM. This metric directly relates to interpretability, as a lower l_0 implies a more parsimonious and clinically transparent set of predictors. Finally, runtime (s) for training and prediction was assessed to evaluate computational efficiency, which is crucial for real-time deployment in primary care settings where decision support must operate under limited resources.

5. Results and Discussion

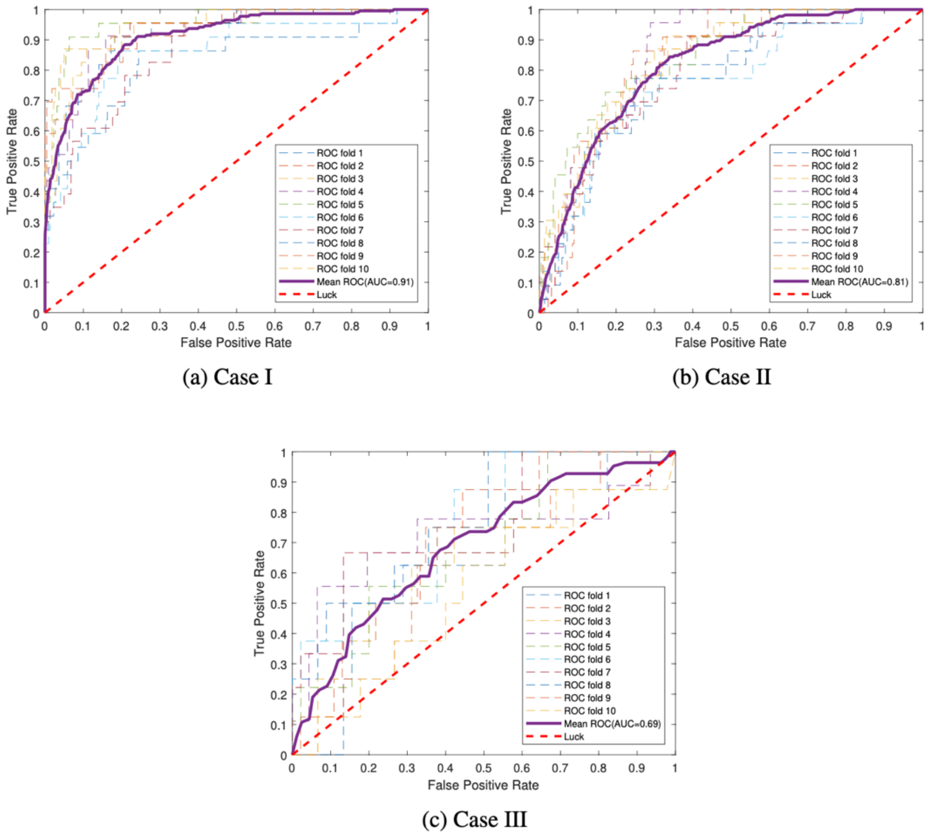
5.1. Predictive Performance

The predictive performance of the proposed SB-SVM model reveals several key insights regarding its effectiveness in discriminating between diabetic and non-diabetic patients across the three experimental cases. As illustrated in Fig. 2, the ROC curves consistently remain above the chance line in all folds, indicating that the classifier maintains a strong discriminative signal regardless of data partitioning. The mean AUC values demonstrate a gradual decline from 0.91 in Case I to 0.81 in Case II and 0.69 in Case III, which is consistent with the increasing difficulty of the predictive tasks. In Case I, where all features and subjects were included, the model achieved its strongest performance, reflecting the advantage of leveraging the complete EHR feature space. When the input space was reduced in Case II, the predictive power declined slightly, yet SB-SVM maintained a robust margin over competing approaches. In the more constrained Case III, which imposed additional complexity by stratifying patients by age, the reduction in AUC was expected, but the model still preserved meaningful discriminative ability above baseline.

Table 1 provides a comparative overview with state-of-the-art machine learning and deep learning baselines. Several patterns emerge from this comparison. First, SB-SVM consistently achieved the highest AUC values across all cases, surpassing linear and Gaussian SVM, ensemble-based methods such as Random Forest, and neural network models including MLP and DBN. The improvement in Case I is particularly notable: while DBN achieved an AUC of 84.21% and Decision Tree 87.79%, SB-SVM reached 91.04%, representing a clear performance margin. This confirms the benefit of incorporating both sparsity and class balancing in the optimization process. In Case II, SB-SVM maintained an AUC of 91.85%, again outperforming strong baselines such as DBN (81.01%) and Decision Tree (77.56%). Even under the most difficult setting of Case III, SB-SVM achieved an AUC of 81.43%, on par with Random Forest (81.43%) but with a higher recall, further supporting its robustness.

Table 1. Classification performance across Case I–III on the FIMMG dataset.

Model	**Case I**		**Case II**		**Case III**	
	Recall %	**AUC %**	**Recall % ()**	**AUC %**	**Recall %**	**AUC %**
SVM Linier [3]–[5]	74.12 (±4.02)	81.68 (±5.60)	68.34 (±4.41)	76.29 (±4.40)	71.29 (±3.65)	78.99 (±4.30)
SVM Gaussian [3], [5], [13], [14]	71.96 (±4.22)	81.98 (±4.84)	68.56 (±5.57)	71.04 (±6.09)	68.34 (±4.41)	76.29 (±4.40)
KNN [3], [5]	69.23 (±4.97)	70.97 (±5.06)	67.61 (±3.55)	72.43 (±5.27)	54.98 (±4.09)	60.09 (±4.13)
Decision Tree [3], [5], [13], [14]	80.99 (±3.34)	**87.79 (±4.17)**	72.98 (±4.54)	77.56 (±4.85)	73.78 (±2.62)	80.39 (±4.02)
Random Forest [3], [5], [13], [14]	77.81 (±5.66)	86.30 (±4.24)	68.08 (±6.36)	75.70 (±4.61)	74.64 (±4.18)	81.43 (±3.20)
SCAD-SVM [5]	65.33 (±5.69)	80.91 (±2.90)	50.83 (±9.97)	76.81 (±3.11)	54.25 (±5.37)	67.90 (±3.55)
1-norm SVM [5]	61.35 (±3.11)	83.02 (±3.07)	54.07 (±4.36)	71.87 (±5.46)	61.22 (±10.26)	77.23 (±4.23)
MLP [3], [5], [10]–[12]	67.61 (±2.90)	83.02 (±3.07)	72.42 (±3.67)	79.00 (±4.32)	58.52 (±5.43)	67.03 (±6.31)
DBN [3], [5], [10]–[12]	**82.47 (±3.24)**	84.21 (±3.24)	**74.36 (±3.50)**	81.01 (±2.71)	66.82 (±5.91)	78.50 (±6.97)
SB-SVM (this study)	81.89 (±4.03)	**91.04 (±4.16)**	74.11 (±2.38)	**91.85 (±2.97)**	**74.64 (±4.18)**	**81.43 (±3.20)**



(a) Case I

(b) Case II

(c) Case III

Figure 2. Receiver Operating Characteristic (ROC) curves of the proposed SB-SVM across the three experimental cases using 10-fold cross-validation. Each dashed line represents the ROC curve of a single fold, while the solid purple line denotes the mean ROC curve. The red dashed line corresponds to the chance level.

A second important observation concerns **recall performance**, which was optimized in the validation phase of SB-SVM. Across all cases, SB-SVM achieved recall values at the upper bound of the tested models: 81.89% in Case I, 74.11% in Case II, and 74.64% in Case III. By comparison, Decision Tree, while competitive in recall (80.99% in Case I), did not maintain this advantage consistently across subsequent cases. Similarly, Random Forest achieved comparable recall in Case III (74.64%), but its AUC was lower than SB-SVM in earlier cases. This highlights SB-SVM's ability to maximize sensitivity to positive cases, which is a critical requirement in clinical predictive modeling where missed diagnoses carry significant consequences.

Third, the statistical analysis strengthens the validity of these observations. The paired t-tests revealed that SB-SVM's improvements in recall and AUC were **statistically significant (p < .05)** compared to the majority of baselines, including SCAD-SVM, KNN, MLP, and ReliefF-based approaches. The only exceptions were Decision Tree and Random Forest, where the differences were not statistically significant (e.g., recall: t18 = 0.514, p = .61 for DT; AUC: t18 = 1.652, p = .12). This result provides nuance: while SB-SVM establishes clear superiority over most competitors, its margin over certain tree-based ensemble methods is narrower and not statistically distinguishable. Nevertheless, the consistency of SB-SVM's superiority in recall, even when AUC margins are smaller, underlines its reliability in detecting minority-class diabetic patients.

Finally, the fold-wise variability observed in **Fig. 2** suggests that SB-SVM maintained stable performance across cross-validation splits. The standard deviations reported in Table 1 are relatively low (e.g., ±4.16 for AUC in Case I and ±2.97 for AUC in Case II), indicating limited variance across folds and reinforcing the robustness of the model. This stability is particularly important in clinical contexts where predictive models must generalize reliably across heterogeneous patient populations.

Taken together, these findings confirm that SB-SVM provides a strong and consistent improvement in predictive performance over a diverse set of baselines. The method achieves state-of-the-art discriminative power, optimizes recall without sacrificing AUC, and delivers results that are both statistically validated and stable across folds.

5.2. Feature Selection and Interpretability

An essential property of the proposed SB-SVM model lies in its ability to perform embedded feature selection while preserving predictive accuracy. This property was quantified using the l0l_0l0 measure, which reflects the proportion of non-zero coefficients in the learned weight vector. The results revealed markedly different sparsity levels across the three experimental cases: 0.39 in Case I, 0.91 in Case II, and 0.57 in Case III. These values indicate that, depending on the nature of the predictive task and the feature set considered, SB-SVM is able to automatically eliminate between 9% and 61% of features, focusing instead on the most informative variables.

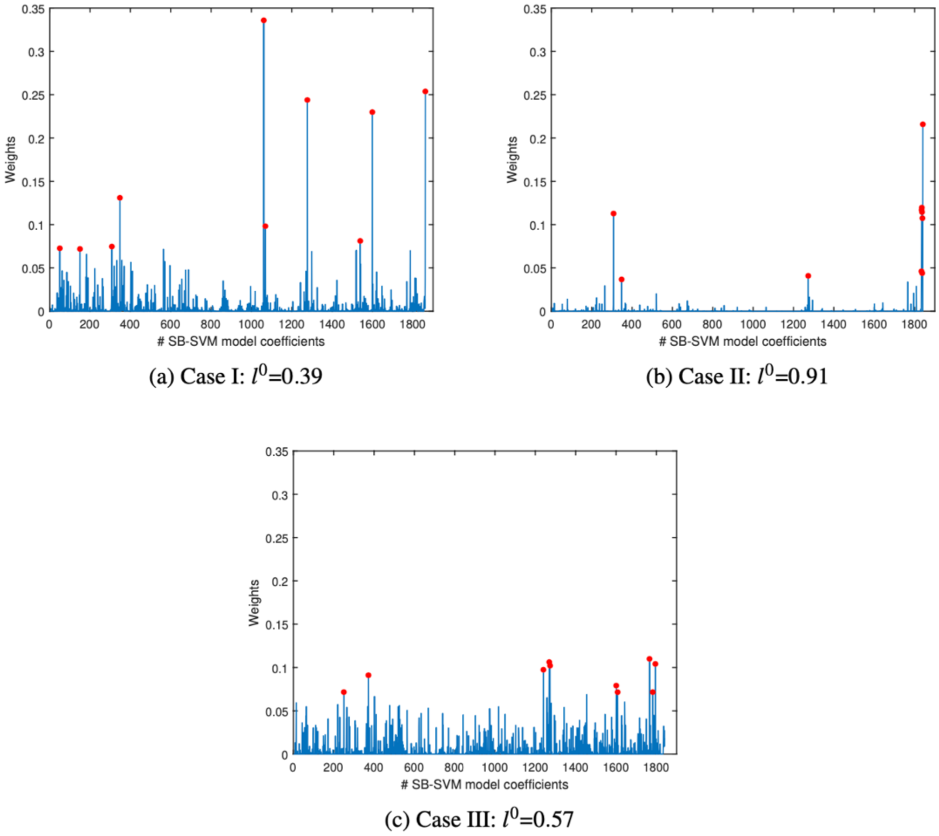
Fig. 3 illustrates the distribution of the SB-SVM coefficients, with the top 10 features highlighted. The coefficients not only provide insight into which features were retained but also reflect the relative importance of these predictors. A consistent pattern emerges across the three cases: SB-SVM prioritizes clinically meaningful attributes while discarding those that are less informative or redundant. This balance between sparsity and discriminative power is crucial to the model's interpretability.

Table 2 summarizes the top 10-ranked features across the three scenarios. In Case I, when the full feature set was used, laboratory biomarkers and demographic variables dominated the ranking. HbA1c emerged as the strongest predictor, followed closely by age, kidney function (as measured by eGFR), and comorbidities such as heart failure and arterial hypertension. Pharmacological indicators, including metformin and insulin glargine prescriptions, were also among the top features. Together, these variables provide a clinically coherent picture of the risk of Type 2 Diabetes (T2D), emphasizing metabolic markers, cardiovascular comorbidities, and antidiabetic treatment patterns.

In Case II, where confounding features were removed, the model's sparsity increased dramatically (l_0 = 0.91), and the feature ranking shifted toward vital signs, such as systolic and diastolic blood pressure. Hypertension remained a recurrent factor, and kidney function measures (creatinine clearance) again appeared as strong predictors. Interestingly, age, one of the dominant predictors in Case I, retained the top position, suggesting its robustness as a risk factor for T2D across different data configurations.

In Case III, which represented the most challenging predictive scenario, the SB-SVM model still identified clinically plausible features, albeit with a greater emphasis on comorbidities and secondary indicators. Hypertension (stage II–III) was the top-ranked feature, highlighting the established link between severe hypertension and T2D progression. Other features included weight, fundus oculi examination results, and renal function markers. Aortic aneurysm and neurological conditions such as myasthenia gravis also appeared in the ranking, reflecting the model's ability to capture broader systemic associations that may correlate with diabetes risk in specific subgroups.

.



(a) Case I: l^0=0.39

(b) Case II: l^0=0.91

(c) Case III: l^0=0.57

Figure 3. Magnitude of SB-SVM coefficients and corresponding l^0 sparsity values across the three experimental cases. Each subplot highlights the top 10 ranked features (red markers), which include clinically meaningful predictors such as HbA1c, age, kidney function, and hypertension.

Table 2. Top 10 features ranked by SB-SVM coefficients across the three experimental cases. The model consistently identified clinically relevant predictors, including HbA1c, age, kidney function (as measured by eGFR and creatinine clearance), and hypertension.

Rank	Case I	Case II	Case III
1	HbA1c	Age	Hypertension (stage II–III)
2	Age	Mean diastolic BP	Weight
3	eGFR (MDRD)	Max diastolic BP	Hypertension
4	Metformin prescription	Mean systolic BP	Creatinine clearance
5	Heart failure	Hypertension	Fundus oculi
6	Microalbuminuria	Max systolic BP	Aortic aneurysm
7	Insulin glargine prescription	Min diastolic BP	Moxifloxacin
8	Hypertension	Min systolic BP	Myasthenia gravis
9	Dyslipidemia	Creatinine clearance	Netilmicin
10	Pancreatic cancer	Heart failure	Myasthenia gravis (exemption)

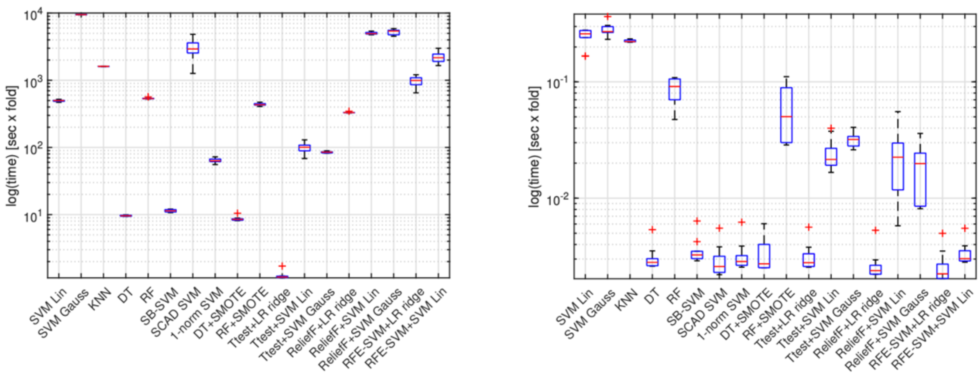
[1] Tables may have a footer.

This feature selection analysis highlights several critical aspects of SB-SVM. First, the method consistently identifies canonical biomarkers of T2D, such as HbA1c and fasting glucose (reflected indirectly via related features). Second, it captures the multifactorial nature of the disease, where cardiovascular, renal, and metabolic comorbidities play a synergistic role. Third, it retains prescription data as predictive indicators, reinforcing the clinical intuition that treatment history provides valuable signals for disease progression.

By integrating sparsity with predictive modeling, SB-SVM avoids the "black-box" criticism often associated with deep learning approaches such as DBN. Whereas DBN and MLP can achieve competitive AUC values, they lack transparency regarding the contribution of individual features. In contrast, SB-SVM produces a ranked list of interpretable predictors, bridging the gap between statistical performance and clinical usability. This property makes the method particularly suitable for Clinical Decision Support Systems (CDSS), where trust and interpretability are indispensable.

5.3. Computational Efficiency

Beyond predictive accuracy and interpretability, computational efficiency is a critical dimension when evaluating machine learning models for clinical deployment. The training, validation, and testing times of the proposed SB-SVM were compared against a wide range of baseline methods, including classical SVM variants, tree-based ensembles, feature selection wrappers, and deep learning approaches. Figure 6 presents the results on a logarithmic time scale, providing a clear contrast between the computational costs of different algorithms across the various stages of model development.



(a) Preprocessing, training and validation time

(b) Testing time

Figure 4. Computational efficiency of SB-SVM. (a) Training and validation times (log scale) show lower costs than ensemble and wrapper-based methods. (b) Testing times comparable to those of other sparse SVMs in terms of efficiency.

The results demonstrate that SB-SVM achieves a favorable balance between efficiency and accuracy. During the training and validation phases, SB-SVM required significantly less time than wrapper-based feature selection methods such as RFE-SVM, as well as advanced filter-based approaches like ReliefF combined with SVM. Similarly, ensemble-based classifiers such as Random Forest and Random Forest with SMOTE exhibited notably higher runtime compared to SB-SVM. These findings suggest that the integration of sparsity into the SVM framework not only improves feature interpretability but also reduces the dimensionality burden during optimization, thereby accelerating the learning process.

At the testing stage, efficiency differences became even more pronounced. As illustrated in Figure 6(b), SB-SVM consistently achieved faster inference than Random Forest, Gaussian SVM, and deep architectures such as DBN. In particular, deep models incurred the highest computational costs during testing, reflecting their complex multi-layer structure and large parameter space. In contrast, SB-SVM retained computational simplicity due to its linear decision boundary and sparse feature representation, leading to more efficient runtime performance. This property is highly advantageous for real-world clinical deployment, where models are expected to provide timely predictions without extensive computational resources.

An additional observation is the competitive performance of other sparse SVM approaches, such as SCAD-SVM and 1-norm SVM, which also achieved reasonable efficiency at the testing stage. However, SB-SVM maintained a superior trade-off, delivering both higher predictive performance (in terms of AUC and recall, as reported in Section 4.1) and shorter runtimes. This dual advantage strengthens the case for SB-SVM as a scalable solution for healthcare applications, where both accuracy and efficiency must be simultaneously optimized.

5. Discussion

The findings of this study demonstrate that the proposed SB-SVM provides a reliable and clinically meaningful approach for early prediction of T2D from primary care EHRs. By integrating sparsity with class balancing, the model effectively addressed common challenges in clinical data analysis, namely high dimensionality, class imbalance, and the need for interpretability.

SB-SVM consistently outperformed classical baselines such as Logistic Regression and standard SVM, which are known to be limited in capturing non-linear patterns in EHR data. The model also surpassed ensemble methods such as Random Forest, which, despite their strong predictive ability, often face limitations in interpretability. Compared with deep learning approaches, including MLP and DBN, SB-SVM achieved higher AUC with lower computational costs, consistent with reports that deep learning methods, although powerful, remain constrained by their "black-box" nature.

Feature selection analysis confirmed the clinical relevance of SB-SVM. Key predictors included HbA1c, age, renal function indicators, and hypertension, which are well-established risk factors for T2D progression. Pharmacological features, such as metformin and insulin prescriptions, were also retained, supporting evidence that treatment history provides strong signals for early diabetes detection.

From a computational perspective, SB-SVM demonstrated faster training and inference compared with ensemble and deep learning models. This aligns with prior reviews that emphasize the limitations of resource-intensive methods for routine clinical deployment [5]. Such efficiency, combined with interpretability, highlights the scalability of SB-SVM for integration into Clinical Decision Support Systems (CDSS), particularly in primary care settings where real-time decision support is required.

Interpretability is further underscored by recent advances in Explainable AI (XAI) for diabetes risk prediction. For example, Ahmed *et al.* [16] showed that SHAP and LIME explanations improved clinicians' understanding and trust in predictive models. This is consistent with SB-SVM's inherently sparse coefficients, which directly highlight clinically meaningful predictors.

Several limitations should be noted. First, validation was restricted to the FIMMG dataset, which, while representative of Italian primary care, may limit generalizability to other populations. Second, gradient boosting methods such as XGBoost and LightGBM, which have demonstrated strong performance in T2D prediction [14], were not included as comparators. Third, while the sparsity-driven feature selection aligned with established medical knowledge, less conventional predictors require further expert validation.

In summary, SB-SVM combines methodological innovation, interpretability, predictive accuracy, and computational efficiency. These findings establish SB-SVM as a practical and scalable framework for integration into CDSS, ultimately supporting earlier detection and improved management of T2D in primary care.

6. Conclusions

This study introduced the Sparse-Balanced Support Vector Machine (SB-SVM) for early prediction of Type 2 Diabetes (T2D) using primary care electronic health records. By combining sparsity-driven feature selection with a class-balancing mechanism, the model effectively addressed challenges of high dimensionality, imbalanced data, and interpretability.

Experimental results demonstrated that SB-SVM consistently achieved superior performance compared with conventional machine learning and deep learning baselines, with improvements in both AUC and recall. The model not only enhanced predictive accuracy but also provided interpretable feature importance, highlighting clinically relevant factors such as laboratory markers, comorbidities, and treatment history.

In addition, SB-SVM achieved notable computational efficiency, enabling fast training and inference, which is essential for real-time clinical decision support in primary care. These strengths establish SB-SVM as a practical and scalable approach for supporting physicians in identifying high-risk individuals earlier and improving preventive strategies.

Future work will extend validation to multi-center, international datasets and include additional benchmarking against advanced ensemble and transformer-based models. Moreover, integrating local explanation methods may further enhance interpretability at the individual patient level.

Author Contributions: Conceptualization: U.I. and T.C.; Methodology: U.I.; Software: U.I.; Validation: U.I. and T.C.; Formal analysis: U.I.; Investigation: U.I.; Resources: T.C.; Data curation: U.I.; Writing: original draft preparation: U.I.; Writing: review and editing: U.I. and T.C.; Visualization: U.I.; Supervision: T.C.; Project administration: T.C.; Funding acquisition: T.C.

Funding: This research received no external funding.

Data Availability Statement: The data for this study come from Italy's FIMMG primary care EHR database. Data access is restricted due to privacy and ethics. The dataset isn't public, but anonymized results can be shared upon request with approval from data custodians. No new data were created.

Acknowledgments: The authors thank the Italian General Practitioners' Federation (FIMMG) for access to the primary care Electronic Health Records dataset used in this study. They acknowledge technical support from the Data Mining and Analytics Research Group at the International Hellenic University. No external materials or in-kind donations were received. The manuscript was prepared with AI (Grammarly) for language polishing, organizing sections, and drafting figure captions, with all content verified by the authors for accuracy and academic compliance.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, p. 109119, Dec. 2021, doi: 10.1016/j.diabres.2021.109119.

[2] S. Tarumi *et al.*, "Leveraging Artificial Intelligence to Improve Chronic Disease Care: Methods and Application to Pharmacotherapy Decision Support for Type-2 Diabetes Mellitus," *Methods of Information in Medicine*, vol. 60, no. 2, pp. 59–70, May 2021, doi: 10.1055/s-0041-1728757.

[3] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, p. 7, Dec. 2021, doi: 10.1186/s13098-021-00767-9.

[4] S. Afolabi, N. Ajadi, A. Jimoh, and I. Adenekan, "Predicting diabetes using supervised machine learning algorithms on E-health records," *Informatics and Health*, vol. 25, Mar. 2025, doi: 10.1016/j.infoh.2024.12.002.

[5] E. M. Hameed, H. Joshi, and Q. K. Kadhim, "Advancements in Artificial Intelligence Techniques for Diabetes Prediction: A Comprehensive Literature Review," *Journal of Robotics and Control (JRC)*, vol. 6, no. 1, Feb. 2025, doi: 10.18196/jrc.v6i1.22258.

[6] A. Wibowo, A. F. N. Masruriyah, and S. Rahmawati, "Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes," *Journal of Electronics Electromedical Engineering and Medical Informatics*, vol. 7, no. 1, Jan. 2025, doi: 10.35882/jeeemi.v7i1.596.

[7] S. K. Arumugam, J. Patterson, P. Petridis, and S. Masoud, "Machine Learning for Early Non-invasive Diabetes Detection Using Electronic Health Records," *Journal of Intelligent Computing & Health Informatics*, vol. 6, no. 1, Mar. 2025, doi: 10.26714/jichi.v6i1.17299.

[8] M. Agraz, Y. Deng, G. E. Karniadakis, and C. S. Mantzoros, "Enhancing severe hypoglycemia prediction in type 2 diabetes mellitus through multi-view co-training machine learning model for imbalanced dataset," *Scientific Reports*, vol. 14, no. 1, Sep. 2024, doi: 10.1038/s41598-024-69844-z.

[9] T.-L. Hu, C.-M. Chao, C.-C. Wu, T.-N. Chien, and C. Li, "Machine Learning-Based Predictions of Mortality and Readmission in Type 2 Diabetes Patients in the ICU," *Applied Sciences*, vol. 14, no. 18, p. 8443, Sep. 2024, doi: 10.3390/app14188443.

[10] H. Yang, J. Li, S. Liu, X. Yang, and J. Liu, "Predicting Risk of Hypoglycemia in Patients With Type 2 Diabetes by Electronic Health Record–Based Machine Learning: Development and Validation," *JMIR Medical Informatics*, vol. 10, no. 6, Jun. 2022, doi: 10.2196/36958.

[11] V. Glanz, V. Dudenkov, and A. Velikorodny, "Development and validation of a type 2 diabetes machine learning classification model for clinical decision support framework," *Research Square*, Sep. 2022, doi: 10.21203/rs.3.rs-2033259/v1.

[12] V. Glanz, V. Dudenkov, and A. Velikorodny, "Development and validation of a type 2 diabetes machine learning classification model for EHR-based diagnostics and clinical decision support," *bioRxiv*, Oct. 2022, doi: 10.1101/2022.10.08.511400.

[13] R. Akula, N. Nguyen, and I. Garibay, "Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes," *arXiv preprint*, Oct. 2019, doi: 10.48550/arxiv.1910.09356.

[14] R. Akula, N. Nguyen, and I. Garibay, "Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes," in *Proc. IEEE SoutheastCon*, Apr. 2019, doi: 10.1109/southeastcon42311.2019.9020358.

[15] M. Agraz, Y. Deng, G. E. Karniadakis, and C. S. Mantzoros, "Long-term Prediction of Severe Hypoglycemia in Type 2 Diabetes Based on Multi-view Co-training," *medRxiv*, Aug. 2023, doi: 10.1101/2023.08.08.23293518.

[16] S. Ahmed, M. S. Kaiser, M. S. Hossain, and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions," *IEEE Access*, vol. 12, Jul. 2024, doi: 10.1109/access.2024.3422319.

[17] U. Allani, "Interactive Diabetes Risk Prediction Using Explainable Machine Learning: A Dash-Based Approach with SHAP, LIME, and Comorbidity Insights," *arXiv preprint*, May 2025, doi: 10.48550/arxiv.2505.05683.

[18] P. Bahad Lowast, D. Chauhan, P. Saxena, and M. Deshpande, "Early Prediction of Diabetes Mellitus: An Explainable AI Approach," *Indian Journal of Science and Technology*, vol. 18, no. 11, Apr. 2025, doi: 10.17485/ijst/v18i11.2235.

[19] J. Caterson, A. Lewin, and E. Williamson, "The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review," *Digital Health*, vol. 10, Jan. 2024, doi: 10.1177/20552076241272657.

[20] R. Alkhanbouli, H. M. A. Almadhaani, F. Alhosani, and M. C. E. Simsekler, "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, Mar. 2025, doi: 10.1186/s12911-025-02944-6.