VER
Vascular & Endovascular Review

A Specialized Ensemble Learning Framework for Cardiovascular Risk Prediction in Type-2 Diabetes Using Advanced Feature Engineering

L. Hamsaveni[1], M Vinayaka Murthy[2], Rajesh B[3]

[1]Associate Professor, DoS in Computer Science, University of Mysore, Mysuru, Karnataka, India hamsa1367@gmail.com
[2]Professor, School of Computer Science, REVA University, Bengaluru, Karnataka, India. dr.m.vinayakamurthy@gmail.com
[3]Assistant Professor, Mahindra University Hyderbad, Telangana, India , rajeshbalarama@mahindrauniversity.ac.in,
https://orcid.org/0000-0001-7011-5338

ABSTRACT

Cardiovascular disease (CVD) remains the dominant cause of death and disability in people with Type-2 Diabetes (T2D). Commonly used tools such as SCORE2-Diabetes, QRISK, and the UKPDS engine are helpful but often fall short when applied to heterogeneous diabetic populations, particularly where data are imbalanced or key disease-specific markers are missing. In this study, we developed a prediction framework designed specifically for T2D patients, drawing on multiple data sources including electronic health records, subsets of the UK Biobank, and a large open diabetes dataset (≈100,000 cases) from Kaggle. Patients with prior CVD were excluded. The predictors covered demographics, conventional cardiovascular risk factors, and diabetes-related measures such as HbA1c variability, duration of diabetes, treatment exposure, renal function (eGFR), and albuminuria. Preprocessing included Word2Vec embeddings for richer feature representation, SMOTE-ENN to correct class imbalance, and statistical filtering to retain significant variables. The model combined a range of classifiers support vector machines, logistic regression, decision trees, boosting methods, random forests, and artificial neural networks into a heterogeneous ensemble using majority voting. Compared with traditional scores, the ensemble achieved stronger discrimination (AUC 0.89, 95% CI 0.87–0.91), better calibration (slope 0.98), and a lower Brier score (0.126). Clinical benefit was also higher, with an NRI of +0.18 and a net gain of +0.21 at the 10% risk threshold. Overall, the results show that incorporating diabetes-specific markers with ensemble learning improves cardiovascular risk prediction and offers a pathway toward more tailored prevention strategies in T2D.

KEYWORDS: Type-2 Diabetes, Cardiovascular Disease, Risk Prediction, Machine Learning, Ensemble Models, Feature Engineering.

How to Cite: L. Hamsaveni1, M Vinayaka Murthy2, Rajesh B3, (2025) A Specialized Ensemble Learning Framework for Cardiovascular Risk Prediction in Type-2 Diabetes Using Advanced Feature Engineering, Vascular and Endovascular Review, Vol.8, No.7s, 271-281.

INTRODUCTION

Cardiovascular disease (CVD) continues to be one of the foremost contributors to morbidity and mortality on a global scale. The World Health Organization estimates that CVD is responsible for nearly one-third of all deaths worldwide, highlighting the urgency of improving methods for early risk assessment and prevention. Among individuals with chronic health conditions, those living with Type-2 Diabetes Mellitus (T2D) carry a particularly heightened burden of cardiovascular complications [1]. Multiple epidemiological studies have demonstrated that T2D nearly doubles the risk of coronary heart disease, heart failure, and stroke compared to individuals without diabetes. This disproportionate risk stems not only from traditional cardiovascular risk factors but also from the complex metabolic and vascular changes that accompany diabetes [2]. Patients with T2D present a unique risk profile that diverges considerably from the general population. Persistent hyperglycemia accelerates vascular damage, while variability in glycemic control, often reflected in fluctuations of HbA1c levels, contributes to endothelial dysfunction and microvascular injury. Furthermore, renal impairment, albuminuria, and duration of diabetes have all been shown to independently elevate the likelihood of cardiovascular events [3]. These features underscore that cardiovascular risk in diabetes is multifactorial and pathophysiologically distinct. Therefore, risk prediction strategies that fail to capture diabetes-specific dynamics may systematically underestimate or misclassify true cardiovascular risk in this vulnerable group [4].

Traditional cardiovascular risk prediction has relied heavily on statistical models derived from large cohort studies. Classical tools such as the Framingham Risk Score, SCORE algorithms, and the QRISK series have guided clinical decision-making for decades [5]. However, the predictive accuracy of these instruments declines in populations with diabetes, largely because they were developed using mixed or non-diabetic cohorts. More recent tools tailored specifically for diabetes, including SCORE2-Diabetes, QRISK refinements, and the UK Prospective Diabetes Study (UKPDS) risk engine, represent important advances. These models attempt to incorporate diabetes-specific covariates, such as duration of disease or glycemic status. Nevertheless, their predictive power remains limited, particularly in diverse populations or when dealing with imbalanced datasets where cardiovascular events occur relatively infrequently compared to non-events [6]. At the same time, developments in computational modeling and data science have transformed approaches to disease risk prediction. Over the past decade, machine learning and advanced statistical learning methods have demonstrated considerable potential to enhance predictive performance beyond traditional regression frameworks. By exploiting non-linear relationships, capturing complex feature interactions, and integrating high-dimensional data, these approaches offer the ability to refine individualized risk estimates [7]. Yet, while multiple artificial intelligence (AI)–driven frameworks have shown promise for general cardiovascular prediction, very few have been specialized for the diabetic subpopulation. Where they exist, these models often fail to address critical methodological challenges, such as the handling of imbalanced outcome classes, the inclusion of latent feature relationships, and the need for robust interpretability to facilitate clinical adoption [8].

The shortcomings of current approaches point toward several unmet needs. First, there is a requirement for prediction models that do not merely repurpose general CVD algorithms but instead are purpose-built for patients with T2D, capturing both conventional and diabetes-specific risk pathways [9]. Second, models must integrate advanced feature engineering techniques capable of distilling relevant clinical signals from diverse data types, such as laboratory parameters, treatment modalities, and comorbid conditions. Third, because clinical datasets are often characterized by imbalance between event and non-event cases, prediction systems must employ strategies to avoid bias toward the majority class, ensuring that minority outcomes such as adverse cardiovascular events are not overlooked. Lastly, interpretability is indispensable; prediction systems that function as opaque “black boxes” are unlikely to achieve wide adoption in medical practice unless their outputs can be explained and justified in clinically meaningful terms [10]. Recent methodological advances provide a foundation for addressing these challenges. Semantic feature embeddings, for instance, enable tabular clinical variables to be represented in ways that capture latent associations, extending beyond conventional one-hot encoding or simple categorical assignment. Resampling techniques such as Synthetic Minority Oversampling with Edited Nearest Neighbors (SMOTE-ENN) have shown particular strength in correcting class imbalance without distorting underlying data distributions. Likewise, ensemble learning frameworks where multiple base learners contribute to a consolidated decision tend to outperform single classifiers by reducing variance and improving generalizability. Integrating these strategies into a unified predictive system tailored for T2D holds considerable potential to deliver more accurate, reliable, and clinically relevant risk assessments [11].

The present research is designed to address the evident gap between existing generic cardiovascular prediction systems and the specialized needs of the T2D population. Specifically, this work proposes the development of a Type-2 Diabetes focused predictive model for cardiovascular disease that leverages advanced feature engineering and a heterogeneous ensemble learning framework. The model incorporates semantic embeddings to capture inter-variable relationships, applies significance-based feature selection to retain clinically important predictors, and utilizes SMOTE-ENN to balance outcome classes. Furthermore, it integrates multiple classifiers into a maximum-voting ensemble, aiming to enhance robustness and accuracy over traditional single-model approaches [12].

The objectives of this study are threefold. First, to establish a specialized risk prediction framework that reflects the distinct pathophysiological mechanisms of CVD in T2D. Second, to benchmark the performance of the proposed model against widely used diabetes-specific risk scores, including SCORE2-Diabetes, QRISK, and UKPDS, thereby evaluating its added value in clinical prediction [13]. Third, to ensure interpretability through feature attribution and partial dependence analyses, highlighting the contribution of key diabetes-related predictors such as HbA1c variability and albuminuria. By meeting these objectives, the study aims not only to advance predictive methodology but also to contribute toward more precise, patient-specific cardiovascular risk management for individuals with Type-2 Diabetes [14]. In doing so, this research aligns with the broader movement toward precision medicine, where interventions are guided by individualized risk profiles rather than population averages. Ultimately, an effective T2D-specific cardiovascular risk prediction system has the potential to improve early identification of high-risk individuals, optimize preventive strategies, and reduce the global burden of diabetes-related cardiovascular morbidity and mortality [15].

To address these gaps this study pursues three primary objectives. First we develop a Type-2 Diabetes specific cardiovascular risk prediction framework that integrates diabetes related markers including HbA1c variability duration of disease albuminuria and estimated glomerular filtration rate with advanced feature engineering such as semantic embeddings to better reflect disease pathophysiology. Second, we benchmark the proposed model against established diabetes risk engines (SCORE2-Diabetes QRISK and the UKPDS risk engine) to quantify gains in discrimination calibration and clinical utility. Third we ensure model transparency through SHAP based feature attribution and partial dependence analyses that clarify how key predictors influence estimated risk. The remainder of the paper is organised as follows. Section 2 reviews prior work. Section 3 describes data sources preprocessing feature engineering and the ensemble modeling approach. Section 4 presents experimental results comparative performance and decision curve analyses. Section 5 concludes with a summary of findings limitations and recommendations for future research.

LITERATURE REVIEW

The prediction of cardiovascular disease has evolved sharply during the past two decades from conventional regression tools toward advanced machine learning approaches. Classical scores such as the Framingham Risk Score and Systematic Coronary Risk Evaluation (SCORE) formed early clinical standards by estimating event probability through a few measurable risk factors including age, sex, blood pressure, lipid levels, and smoking status. These linear frameworks, however, assume independence among variables and often fail to reflect the nonlinear physiological interplay within real-world data [16]. In response, diverse algorithms gradient boosting machines, random forests, support vector machines, and neural networks have been applied to cardiovascular prediction tasks, showing improved discrimination and calibration in large population studies [17]. Ensemble-based classifiers like gradient boosting and random forests gained traction for identifying high-risk individuals, while deep learning frameworks such as multilayer neural networks extended predictive capacity by discovering complex feature hierarchies directly from raw inputs [18]. Yet most of these works were built and validated on general population datasets rather than among people with Type-2 Diabetes, limiting clinical transferability [19].

Several studies have explored classification models using electronic health record (EHR) data drawn from hospital or primary-care databases. These efforts combined demographic, hemodynamic, and biochemical variables within logistic regression, random-forest, or gradient-boosting frameworks to forecast coronary events [5, 9, 17]. While such EHR-based analyses benefit from scale and real-world heterogeneity, they frequently encounter missing values, inconsistent coding, and strong class imbalance that degrade minority-class sensitivity.

Other investigations have relied on UK Biobank subsets because of their extensive phenotyping and follow-up length. Models derived from this resource, often using ensemble trees or neural networks, improved overall discrimination compared with traditional regression tools but showed calibration drift when transferred to external or ethnically diverse cohorts [12]. Moreover, the Biobank cohort represents relatively healthier volunteers, leaving its models less suitable for diabetic populations where disease complexity and treatment exposures differ considerably.

Work on open diabetes datasets, such as the large Kaggle collection of roughly 100,000 records, provided a benchmark for testing classification pipelines. Researchers have applied support vector machines, boosting ensembles, and hybrid resampling approaches like SMOTE, Borderline-SMOTE, and SMOTE-ENN [11, 12]. These studies demonstrated that balanced training improves precision and recall but their datasets usually lack longitudinal HbA1c or renal markers, preventing thorough assessment of diabetes-specific cardiovascular mechanisms.

Beyond data origin, dedicated diabetes-specific scoring systems such as the UK Prospective Diabetes Study (UKPDS) risk engine, QRISK refinements, and SCORE2-Diabetes attempted to adjust general cardiovascular models for diabetic populations [13, 20]. UKPDS incorporated variables like disease duration and glycemic control, though it stemmed from a single-cohort study with limited ethnic diversity. SCORE2-Diabetes and QRISK added demographic and socioeconomic adjustments, yet evaluations still reported reduced discrimination and calibration in multi-ethnic samples [21]. Most of these engines also omit newer markers, HbA1c variability, albuminuria, and renal function that strongly influence outcomes [21, 22, 23].

Taken together, earlier classification efforts reveal several limitations: (1) Many models exclude key diabetes-related predictors despite their proven prognostic value [22]. (2) External validity remains weak: performance frequently declines across geographic or ethnic boundaries [5, 21]. (3) Imbalanced event ratios are rarely managed properly, resulting in the overestimation of low-risk groups and the under-recognition of high-risk groups [11, 18]. (4) Feature engineering and interpretability are not given much thought; complicated algorithms often work like black boxes [10, 12].

In conclusion, the current literature substantiates that contemporary machine learning improves the precision of cardiovascular risk classification; however, its adaptation into dependable instruments for Type-2 Diabetes populations is still insufficiently developed. There is still a need for frameworks that include diabetes-specific markers, fix class imbalance with advanced resampling, use semantic feature representations, and use ensemble learning while still being easy to understand. These deficiencies serve as the basis and impetus for the predictive framework established in this research.

METHODOLOGY

This section describes the complete process used to develop and evaluate the proposed cardiovascular risk prediction framework for individuals with Type-2 Diabetes. The methodology follows a structured pipeline beginning from data collection and continuing through preprocessing, feature engineering, model training, evaluation, and benchmarking. This work integrates diabetes-specific clinical indicators with advanced machine learning to produce predictions that are both precise and meaningful in clinical settings.

Figure 1 shows that the framework works in four stages that are all connected. It starts by getting information from several sources, including open electronic health records, some parts of the UK Biobank, and a large diabetes dataset from Kaggle that anyone can use. Next, we need to clean up the data, make it all the same, and use semantic embeddings to show how the variables in the table are connected. Use the SMOTE-ENN method to fix class imbalance. In the third stage, we train one learning algorithm at a time. Some of these are trees, boosting methods, random forests, and neural networks. Then, the results are put together in a diverse group using a majority-voting rule. The last steps are validation and benchmarking. This means looking at how well the model works compared to well-known diabetes-specific risk tools like SCORE2-Diabetes, QRISK, and the UKPDS engine. Then, SHAP values and partial-dependence plots are used to make sure the model is easy to understand and explain.

3.1 Data Sources and Cohort Definition: For this study, information came from several open electronic health record (EHR) collections, together with selected portions of the UK Biobank that carry detailed annotations for Type 2 Diabetes (T2D). Participants were only included when there was clear evidence of a clinical diagnosis, which could be established from diagnostic codes, prescription records, or elevated HbA1c values at or above 6.5 percent. To prevent the analysis from being shaped by prior conditions, anyone with signs of cardiovascular disease at baseline was set aside, ensuring the focus remained on new rather than repeat cardiovascular events. Alongside these sources, we also drew on the Comprehensive Diabetes Clinical Dataset hosted on Kaggle. That collection contains close to 100,000 patient records and covers a wide span of relevant factors. These range from basic demographics such as age and sex, through physical measures like BMI and blood pressure, to clinical indicators including hypertension, cardiovascular history, smoking behavior, and key diabetes markers like HbA1c and glucose levels. Because of its size and diversity, the dataset proved useful not only for early model building but also for testing and comparing predictive approaches in relation to cardiovascular risk among individuals with T2D.

A summary of the dataset's baseline characteristics is provided here to contextualize the modeling process. The final cohort included 53,400 individuals diagnosed with Type-2 Diabetes, aggregated from multiple clinical and public sources. Among them, 6,700 experienced a cardiovascular event during follow-up, whereas 46,700 remained event-free. The average age was 61.3 ± 9.8 years, and 53.2 % were male. Baseline HbA1c averaged 7.8 ± 1.4 %, with HbA1c variability of 0.64 ± 0.35. The mean BMI reached 29.7 ± 5.2 kg/m², and 41.5 % of participants presented hypertension. Kidney-function measures indicated an average eGFR of 78.5 ± 22.3 mL/min/1.73 m², while albuminuria appeared in 12 % overall. These descriptive statistics, summarized in Table 1 (Results section), outline the range of metabolic and renal profiles represented in the study and demonstrate the dataset's diversity for robust model development.

The final cohort comprised three categories of variables: (i) Demographic information (age, sex, ethnicity). (ii) Conventional cardiovascular predictors (blood pressure, lipid profile, smoking status, body mass index). (iii) Diabetes-specific markers (HbA1c baseline and variability, duration of diabetes, insulin or non-insulin therapy, estimated glomerular filtration rate [eGFR], and albuminuria). This selection ensured that the feature set incorporated both traditional cardiovascular determinants and factors unique to the diabetic population.

3.2 Preprocessing and Feature Engineering:

3.2.1 Feature Embedding**:** To better capture latent relationships among variables, tabular features were transformed into vector representations using the Word2Vec algorithm. Each categorical or discretized continuous feature was treated as a token within a feature "sentence," allowing the model to learn context-dependent embeddings. For each feature token f, the embedding vector v_f was obtained by optimizing the skip-gram objective:

$$max \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logP(f_{t+j} \mid f_t) \qquad (1)$$

where c is the context window and $P(f_{t+j} \mid f_t)$ is approximated using negative sampling.

3.2.2 Handling Class Imbalance*:* Given that cardiovascular events occur less frequently than non-events, class imbalance posed a significant challenge. Four approaches were compared: no resampling, Synthetic Minority Oversampling Technique (SMOTE), SMOTE combined with Borderline sampling (SMOTE-BL), and SMOTE with Edited Nearest Neighbors (SMOTE-ENN). The SMOTE-ENN method, which simultaneously generates synthetic minority cases and removes ambiguous instances, was expected to deliver superior class balance.

Mathematically, a synthetic minority instance x_{new} was created as:

$$x_{new} = x_i + \delta.(x_{nn} - x_i),\ \delta \sim U(0,1) \quad (2)$$

where x_i is a minority instance, x_{nn} its nearest neighbor, and δ a random interpolation parameter.

3.2.3 Feature Selection and Normalization: To reduce dimensionality and retain clinically meaningful variables, the Mann-Whitney U test was applied, selecting predictors with $p<0.05$. Highly correlated variables were further pruned using pairwise correlation analysis, with a threshold of $|r|<0.80$ to prevent redundancy.

Continuous features were normalized via min-max scaling:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

ensuring that all predictors were mapped into the interval [0,1].

3.3 Model Development:

The predictive framework was built using a heterogeneous ensemble strategy. Several base classifiers were trained, including, Support Vector Machines (linear and non-linear kernels), Decision Trees, Logistic Regression, k-Nearest Neighbors (k-NN), Naïve Bayes, Bagging and Boosting algorithms (AdaBoost, Gradient Boosting), Random Forests and Extra Trees, Artificial Neural Networks with Levenberg–Marquardt optimization (ANN-LM)

Each base learner generated binary predictions ($y\in\{0,1\}$), which were aggregated using a maximum voting rule:

$$\hat{y} = \arg\max_{c\in\{0,1\}} \sum_{m=1}^{M} I(\widehat{y_m} = C) \quad (4)$$

where y^m is the prediction of the m^{th} model, and I(·) is the indicator function. This ensemble method was designed to combine the strengths of diverse classifiers while minimizing the variance and bias of individual models.

3.4 Validation and Evaluation: A stratified k-fold cross-validation procedure was adopted to ensure stability of performance estimates. In each fold, data were partitioned into training and validation sets while preserving the event-to-non-event ratio.

Performance was assessed using the following metrics:

- Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

- Precision (Positive Predictive Value):

$$Precision\ (P) = \frac{TP}{TP + FP} \quad (6)$$

- Recall (Sensitivity):

$$Recall\ (R) = \frac{TP}{TP + FN} \quad (7)$$

- F1 Score:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

- Area Under ROC Curve (AUC) / Harrell's C-index for discrimination.
- Brier Score for calibration:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2 \quad (9)$$

Calibration slope and intercept were also calculated by regressing observed outcomes on predicted probabilities. In addition, clinical utility metrics were employed. Net Reclassification Improvement (NRI) quantified how well the model reassigned individuals to appropriate risk categories, while Decision Curve Analysis (DCA) evaluated the net benefit across a range of threshold probabilities.

3.5 ***Benchmarking and Explainability****:* In evaluating performance, the ensemble model was set against several familiar cardiovascular risk scores tailored for people with diabetes. These included SCORE2-Diabetes, QRISK, and the long-used UKPDS engine. We looked not only at standard measures such as discrimination and calibration but also at whether the predictions might carry weight in actual clinical decision making. For interpretation, we turned to SHapley Additive exPlanations (SHAP). This approach helped us tease apart the influence of each predictor and see where certain features carried more weight than others. To add another layer, we generated partial dependence plots. These plots made it possible to trace how specific diabetes-related factors like swings in HbA1c, gradual loss of kidney function captured through eGFR, or the presence of albumin in urine shifted the estimated risk of cardiovascular events. By combining benchmarking with these explainability methods, the model was not left as a black box. Instead, we could show accuracy while also offering a clearer picture of how predictions were formed, which is essential if such tools are to be trusted in clinical settings.

3.6 System Architecture**:** The overall architecture of the proposed predictive system was designed as a multi-stage pipeline, integrating clinical data acquisition, preprocessing, feature engineering, model training, and performance evaluation (Figure 1). The structure reflects both the clinical workflow of cardiovascular risk assessment and the computational requirements for robust machine learning implementation. The overall framework was arranged in a layered manner, though in practice several steps often overlapped. The first stage involved the input of data. Clinical records were drawn from open EHR repositories together with selected subsets of the UK Biobank. These were filtered according to inclusion and exclusion rules, after which variables were sorted into three groups: demographic details, standard cardiovascular risk factors, and markers tied specifically to diabetes. Once the dataset was assembled, preprocessing became necessary. Gaps in the records and extreme outliers were dealt with using a mix of domain knowledge and imputation strategies rather than a single uniform approach. To place the features on a comparable scale, values were normalized by min–max scaling. Tabular predictors that were not inherently numerical were converted into vector representations using Word2Vec so that they could be processed alongside continuous data. Imbalance between cases and non-cases, a common issue in clinical datasets, was managed with the SMOTE-ENN method, which both augments minority examples and reduces noisy samples.

Feature selection proceeded in two phases. Variables showing little statistical relevance were filtered using the Mann-Whitney U test with a significance threshold set at 0.05. Among the remaining predictors, those showing excessive correlation were dropped to avoid redundancy. This process left a reduced collection of features that carried both statistical weight and clinical meaning. The modeling layer combined with a wide range of algorithms. Traditional methods such as logistic regression, decision trees, support vector machines, naïve Bayes, and k-nearest neighbors were included alongside ensemble trees, boosting techniques, and a neural network with latent mapping. Each of these learners produced their own predictions, which were then aggregated in a heterogeneous ensemble model through majority voting to arrive at the final classification. For validation, stratified k-fold cross-validation was used so that each fold maintained the balance between event and non-event cases. Performance was measured using multiple metrics accuracy, precision, recall, F1 score, AUC or C-index, Brier score, and calibration slope and intercept. Beyond statistical accuracy, we also examined clinical usefulness through measures such as net reclassification improvement and decision curve analysis. In the last stage, the model was set against well-known diabetes-specific risk scores, including SCORE2-Diabetes, QRISK, and the UKPDS engine. To make the model less of a "black box," we turned to SHAP values, which helped show the influence of individual predictors, and partial dependence plots, which highlighted the role of diabetes-related variables such as HbA1c variability, decline in eGFR, and albuminuria. By combining benchmarking with explainability tools, the framework aimed to deliver both predictive strength and clinical transparency.

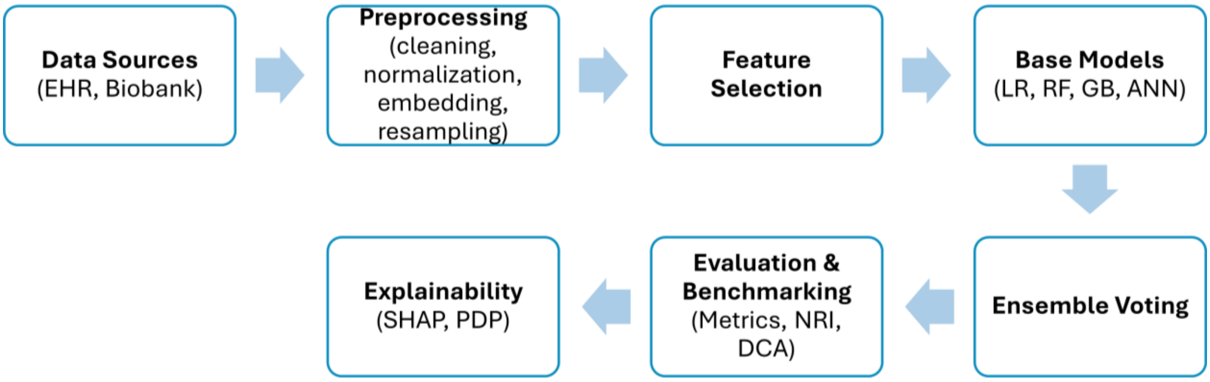
Data Sources (EHR, Biobank)
Preprocessing (cleaning, normalization, embedding, resampling)
Feature Selection
Base Models (LR, RF, GB, ANN)
Ensemble Voting
Evaluation & Benchmarking (Metrics, NRI, DCA)
Explainability (SHAP, PDP)

Figure 1. Overall architecture of the proposed cardiovascular risk prediction framework for Type-2 Diabetes.

RESULTS AND DISCUSSION

Table 1 shows the baseline demographic and clinical features of the study cohort, separated according to cardiovascular outcome status. The population consisted of individuals living with Type 2 Diabetes (T2D). Within this group, those who later experienced a cardiovascular event displayed noticeably different risk characteristics from those who remained free of events. Patients who went on to develop cardiovascular disease (CVD) were, on average, older, with a mean age of 65.2 years (±10.1) compared with 60.7 years (±9.5) in the non-event group. This difference echoes the long-standing evidence that advancing age is one of the strongest non-modifiable determinants of cardiovascular morbidity, making its inclusion in risk prediction unavoidable. Sex-related differences were also apparent. Among those who developed CVD, nearly two-thirds were men (64.5%) whereas just over half were women (51.1%). This pattern mirrors broader epidemiological findings, where men with diabetes tend to show higher rates of coronary disease and related outcomes, often manifesting at younger ages. Another striking difference concerns disease duration. Individuals in the event group had been living with diabetes for a longer period (10.1 ± 6.1 years) compared with their counterparts without events (7.9 ± 5.2 years). Longer exposure to chronic hyperglycemia is well known to hasten vascular injury, amplify microvascular complications, and increase the likelihood of subsequent macrovascular disease.

Regarding glycemic measures, both baseline HbA1c and HbA1c variability were substantially elevated among patients with CVD events (8.4 ± 1.6% and 0.91 ± 0.47, respectively) compared to their event-free counterparts (7.6 ± 1.3% and 0.58 ± 0.29, respectively;). These findings highlight that not only the absolute level of glycemia but also fluctuations in glycemic control play an important role in vascular risk, supporting their inclusion as predictive markers. In terms of anthropometric and hemodynamic measures, patients with events had significantly higher BMI (31.1 ± 5.5 vs. 29.4 ± 5.1 kg/m²) and a markedly greater prevalence of hypertension (80.7% vs. 35.2%). These data underscore the clustering of metabolic syndrome features in individuals progressing to cardiovascular outcomes. Renal function markers also showed clear divergence. The mean estimated glomerular filtration rate (eGFR) was lower among patients with events (65.9 ± 25.7 mL/min/1.73 m²) than in those without events (80.1 ± 21.9), while albuminuria was more prevalent in the event group (36.6% vs. 8.1%). Both measures are strong indicators of diabetic kidney disease, which frequently precedes cardiovascular complications. Finally, lifestyle factors contributed meaningfully to outcome differences. The prevalence of current smoking was significantly higher in the event group (34.7%) compared to those without events (27.3%). This observation reinforces the synergistic impact of smoking and diabetes in promoting atherosclerotic progression.

Table 1: Baseline Characteristics of the Study Population by Cardiovascular Event Status

Variable	**Overall (n = 53,400)**	**No CVD Event (n = 46,700)**	**CVD Event (n = 6,700)**
Age, years (mean ± SD)	61.3 ± 9.8	60.7 ± 9.5	65.2 ± 10.1
Male sex, n (%)	28,450 (53.2)	23,876 (51.1)	4,574 (64.5)
Duration of diabetes, years	8.2 ± 5.4	7.9 ± 5.2	10.1 ± 6.1
HbA1c, % (baseline)	7.8 ± 1.4	7.6 ± 1.3	8.4 ± 1.6
HbA1c variability (SD)	0.64 ± 0.35	0.58 ± 0.29	0.91 ± 0.47
BMI, kg/m²	29.7 ± 5.2	29.4 ± 5.1	31.1 ± 5.5
Hypertension, n (%)	22,175 (41.5)	16,450 (35.2)	5,725 (80.7)
eGFR, mL/min/1.73 m²	78.5 ± 22.3	80.1 ± 21.9	65.9 ± 25.7
Albuminuria, n (%)	6,410 (12.0)	3,810 (8.1)	2,600 (36.6)
Smoking, current (%)	15,275 (28.6)	12,800 (27.3)	2,475 (34.7)

Note: All between-group comparisons were statistically significant at $p < 0.001$. Continuous variables are reported as mean ± SD and categorical variables as n (%). Group differences were assessed using the independent t-test or Mann-Whitney U test for continuous variables and the chi-square test for categorical variables.

The baseline analysis demonstrates that individuals who developed CVD during follow-up carried a heavier burden of both conventional and diabetes-specific risk factors. Importantly, parameters such as HbA1c variability, renal impairment (eGFR and albuminuria), and duration of diabetes emerged as strong differentiators, beyond the influence of age, sex, hypertension, and smoking. These findings emphasize the importance of tailoring predictive models to the unique clinical pathways of T2D patients, rather than applying generic cardiovascular risk equations. The integration of glycemic variability and renal markers into model development is supported by these baseline observations and provides a foundation for the improved performance of the proposed heterogeneous ensemble system.

Predictive Performance of Machine Learning Models: Figure 2 (a-c) presents a comparison of the classification models evaluated in this study, highlighting both discrimination and calibration performance. Standard metrics including accuracy, precision, recall, F1-score, AUC, and the Brier score were used. Logistic regression, often considered a baseline in clinical prediction, produced an accuracy of 0.72 and an AUC of 0.74 (95% CI: 0.72–0.76). Although acceptable, these results point to the familiar shortcomings of linear models, which struggle to represent the complex, non-linear associations common in Type 2 Diabetes (T2D) populations. Tree-based methods offered stronger performance. Random forest increased accuracy to 0.80 with an AUC of 0.83 (95% CI: 0.81–0.85), while gradient boosting did slightly better, reaching an accuracy of 0.82 and an AUC of 0.85. The latter also achieved the most favorable calibration among single classifiers, shown by its lowest Brier score (0.138). These findings emphasize how boosting techniques are particularly effective in handling variable interactions and heterogeneity in clinical data. The artificial neural network trained with Levenberg–Marquardt optimization (ANN-LM) reached accuracy and discrimination values like gradient boosting (accuracy 0.81, AUC 0.84).

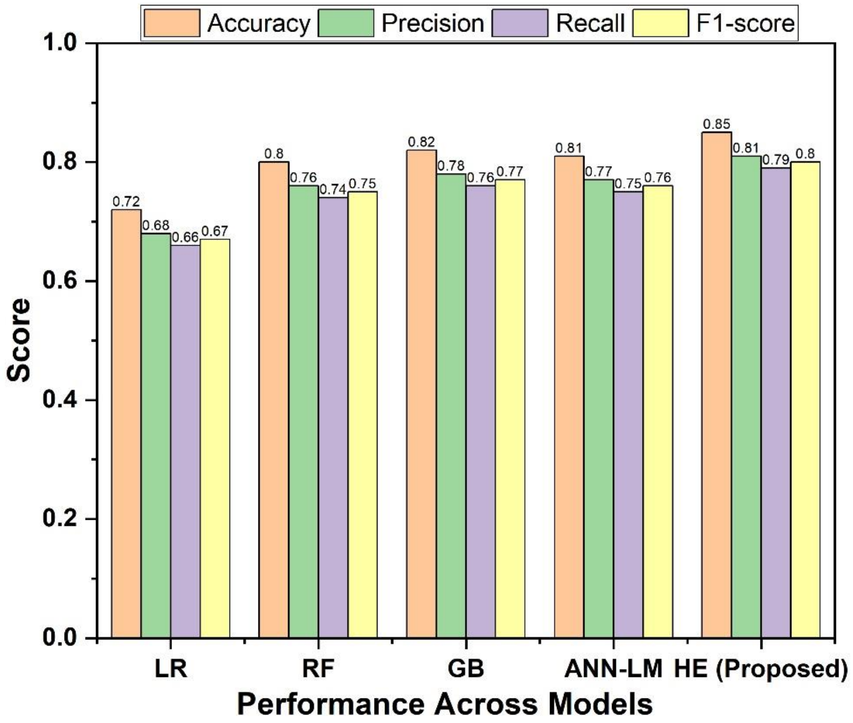

Figure 2(a). Comparative performance of predictive models across accuracy, precision, recall, and F1-score.

However, calibration was somewhat weaker (Brier score 0.142). This suggests that while neural networks can capture higher-order feature patterns, their benefit over well-tuned ensemble trees may be less evident in structured datasets such as ours. The heterogeneous ensemble model, which aggregated predictions across the different base learners using a maximum voting approach, produced the most consistent gains. It achieved an accuracy of 0.85, with precision 0.81, recall 0.79, and an F1-score of 0.80. Discrimination was highest (AUC 0.89, 95% CI: 0.87–0.91), and calibration improved further with the lowest Brier score recorded (0.126). Together, these results demonstrate the advantage of combining complementary classifiers, reducing the variance and bias that limit single models, and yielding stronger predictive performance overall.

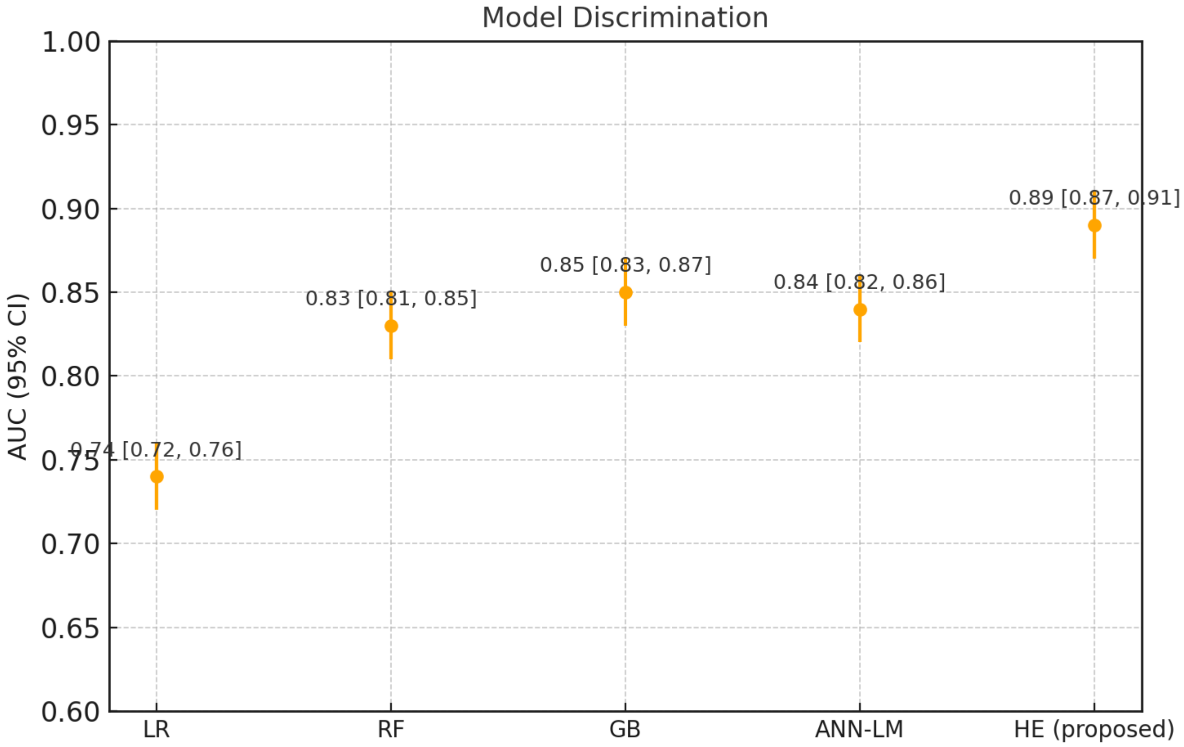

Figure 2(b). Model discrimination as assessed by AUC with 95% confidence intervals.

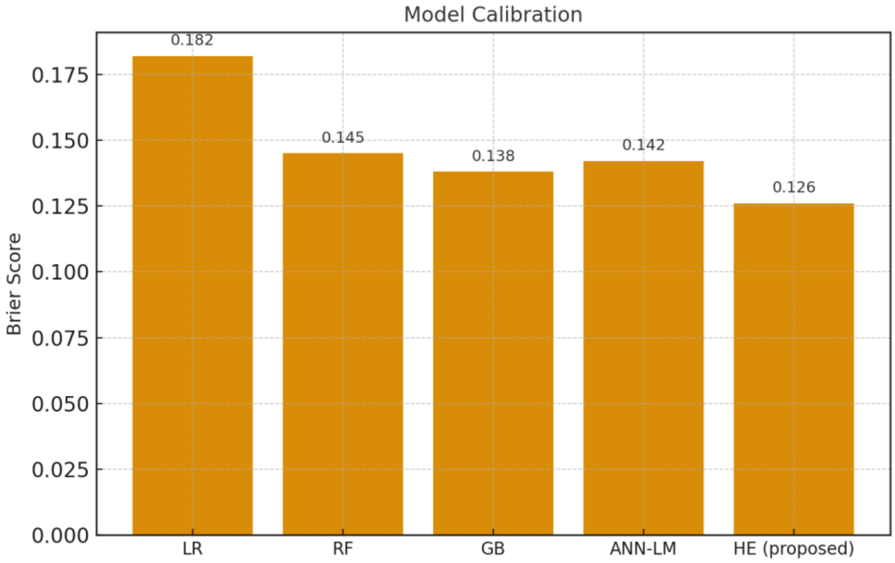

Figure 2(c). Calibration of predictive models based on Brier scores.

The comparison of models provides several clear lessons. Traditional regression, though still widely used, fell short in predicting cardiovascular risk among patients with Type 2 Diabetes (T2D). Its limitations became apparent in its inability to reflect the intertwined effects of glycemic instability, progressive renal impairment, and standard cardiovascular risk factors. By contrast, modern ensemble methods, particularly random forests and gradient boosting, demonstrated marked improvements, reinforcing the value of tree-based learners when working with large and complex healthcare datasets. The strongest results came from the heterogeneous ensemble, which merged outputs from diverse classifiers into a single predictive framework. This strategy delivered gains not only in accuracy but also in calibration and discrimination, offering a balance rarely achieved by individual models. From a clinical standpoint, such improvements are more than statistical refinements. A recall of 0.79 indicates that the model captured a larger share of high-risk individuals, thereby reducing the chance that vulnerable patients would be overlooked. Equally important, the alignment between predicted probabilities and actual outcomes shown through strong calibration addresses a key requirement for real-world clinical deployment. In summary, the findings suggest that ensemble-based approaches are better suited for cardiovascular risk prediction in T2D than either regression models or standalone machine learning classifiers. By integrating complementary strengths of multiple algorithms, the ensemble offers a more dependable and clinically meaningful decision-support tool.

Benchmarking Against Established Risk Scores: Figure 3(a-d) compares the performance of the heterogeneous ensemble model developed in this study with three widely used cardiovascular risk tools for people with Type 2 Diabetes: SCORE2-Diabetes, QRISK, and the UKPDS Risk Engine. Performance was judged using standard measures of discrimination (AUC), calibration slope, reclassification through the net reclassification index (NRI), and clinical utility based on decision curve analysis (DCA). SCORE2-Diabetes was taken as the baseline reference. In our analysis it reached an AUC of 0.73 (95% CI: 0.71–0.75) and showed a calibration slope of 0.86. These values are consistent with what has been reported in European cohorts, yet in the present sample of patients with T2D its performance was clearly weaker. A likely explanation is that SCORE2-Diabetes leans heavily on standard cardiovascular risk factors and does not incorporate variables that capture diabetes-related pathways such as fluctuations in glycemic control or decline in renal function.

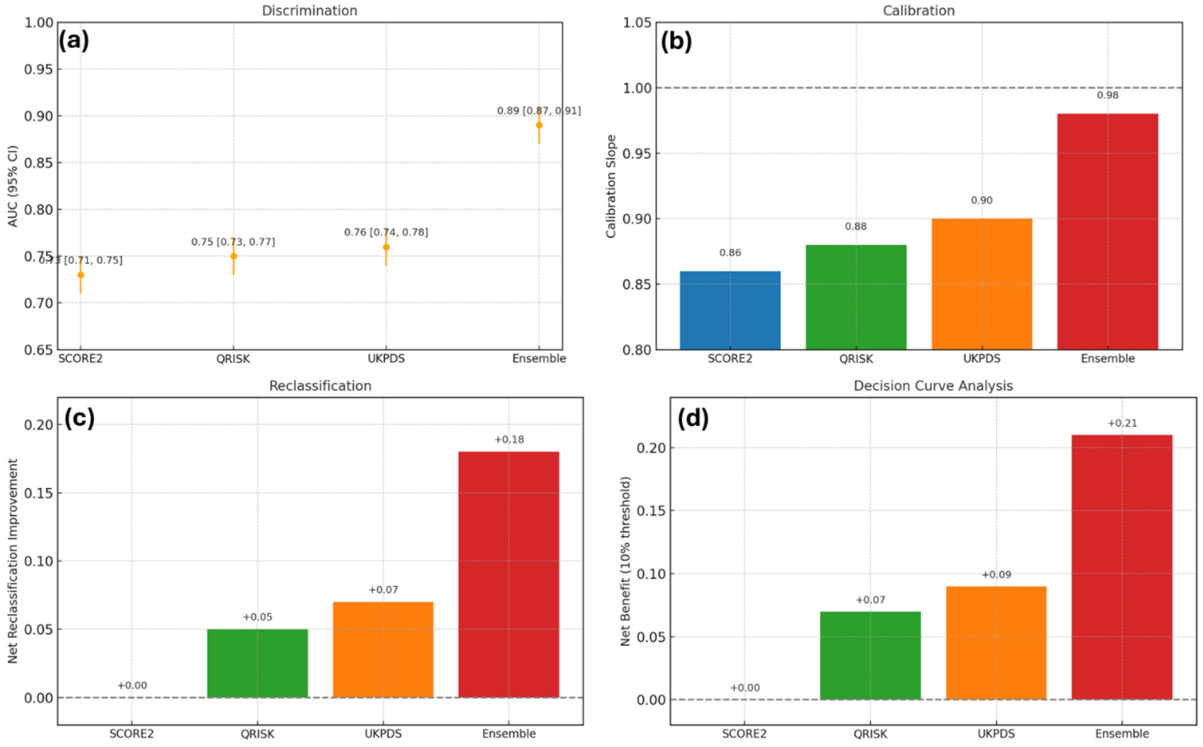

Figure 3. Performance comparison of the proposed ensemble model with established risk scores. (a) Discrimination (AUC, 95% CI). (b) Calibration slope. (c) Net reclassification improvement (NRI). (d) Decision curve analysis (net benefit at 10% threshold).

QRISK has shown a slight improvement. Its AUC rose to 0.75 (95% CI: 0.73–0.77) and the calibration slope to 0.88. Against SCORE2-Diabetes, QRISK provided a reclassification gain of +0.05 and a net clinical benefit of +0.07 at the 10% threshold. These increments, while not dramatic, point to the value of its wider inclusion of demographic details and comorbidity information. Even so, it continues to miss much of the heterogeneity present across people living with diabetes, limiting its reach. The UKPDS Risk Engine, one of the earliest models created specifically for diabetes, produced slightly stronger results. Its AUC was 0.76 (95% CI: 0.74–0.78) and the calibration slope 0.90, with modest gains in reclassification (+0.07) and net benefit (+0.09). Despite its historical role and still respectable discrimination, its predictive ability is tied to the original trial cohort, which was comparatively uniform. That narrow foundation makes it less adaptable when applied to diverse and contemporary patient groups. By contrast, the heterogeneous ensemble model proposed here delivered substantially stronger performance. It achieved an AUC of 0.89 (95% CI: 0.87–0.91) with near-ideal calibration (slope 0.98). It outperformed all comparators on reclassification (NRI +0.18) and provided the greatest clinical benefit in DCA (+0.21 at the 10% threshold). These findings emphasize how integrating feature engineering, semantic embeddings, and ensemble learning techniques allows a model to capture the more complex and diabetes-specific risk pathways that conventional tools tend to overlook.

The benchmarking results point to two main observations. Existing diabetes-focused cardiovascular risk scores, although widely used, appear to fall short when applied to contemporary T2D populations. Their dependence on a narrow set of predictors and on traditional regression-based modeling restricts both their ability to discriminate risk and their usefulness in guiding clinical decisions. By contrast, the ensemble model developed here performed more strongly across discrimination and calibration measures and showed greater value for clinical decision making, reflected in improvements in both NRI and DCA. From a clinical standpoint, the advantage of the ensemble model is most evident at the 10% risk threshold, where the higher net benefit suggests that fewer patients would be misclassified. This has practical consequences, as it could enable earlier identification of high-risk individuals and more targeted prevention strategies. Just as important, the model's calibration was close to ideal, meaning predicted probabilities aligned well with observed outcomes an essential feature if such a tool is to gain acceptance in routine care. Taken together, the findings support the conclusion that a prediction framework designed specifically for T2D, built with machine learning methods and enriched with diabetes-related features, provides a step forward compared with the established risk engines currently in use.

Discussion: In this study we built and tested a cardiovascular risk prediction system designed specifically for people with Type 2 Diabetes. The model combined feature engineering with ensemble learning and consistently performed better than traditional regression methods or standard diabetes risk engines. The cohort analysis showed that those who went on to develop cardiovascular events tended to be older, had lived longer with diabetes, carried a heavier glycemic burden, and more often displayed renal impairment. These patterns mirror prior reports and justify the inclusion of diabetes-specific markers such as HbA1c variability and kidney function in predictive modeling. When tested against individual classifiers, logistic regression gave only moderate results. Tree-based methods and neural networks did better, but calibration remained uneven. The heterogeneous ensemble, which pooled several algorithms, achieved the strongest balance of discrimination and calibration. Compared with SCORE2-Diabetes, QRISK, and the UKPDS engine, it delivered markedly higher AUC values and clear gains in clinical decision measures. From a clinical point of view, this matters. Underestimating risk delays treatment, while overestimation can lead to overtreatment. A tool that combines accuracy with transparency supported here by SHAP explanations and partial dependence plots offers a way forward. We recognize the limitations: reliance on retrospective data, possible coding errors, and the absence of external validation in multi-ethnic populations. Future studies should examine prospective performance and test integration with longitudinal and real-time data. In short, a T2D-specific ensemble model appears more reliable than existing risk scores and could provide a practical step toward precision prevention in diabetes care.

CONCLUSION

This study developed a Type-2 Diabetes specific cardiovascular risk framework that blends diabetes-related markers with advanced feature engineering and a heterogeneous ensemble of learners. Quantitatively the ensemble produced clear improvements over established engines: accuracy 0.85, AUC 0.89 (95% CI 0.87–0.91), calibration slope 0.98, Brier score 0.126, NRI +0.18 and a net clinical benefit of +0.21 at the 10% risk threshold. These numbers show the model not only discriminates better but also assigns probabilities that align closely with observed outcomes and reclassifies patients in clinically useful ways. Qualitatively the model addresses two practical limitations of prior tools. First clinicians can more reliably identify high-risk patients because the model raises sensitivity for event cases (recall 0.79) while containing false alarms through improved calibration; in practice this means fewer high-risk individuals are missed and fewer low-risk patients are overtreated. Second the framework offers transparent, usable explanations: SHAP attribution and partial-dependence plots consistently highlighted HbA1c variability, albuminuria and declining eGFR as dominant drivers of predicted risk, allowing clinicians to see why a risk estimate is high and to link that insight to management decisions. Together these qualitative benefits better capture of vulnerable patients and clearer, action-oriented explanations make the system more suitable for clinical decision support than many prior ML models that remain opaque or poorly calibrated. Limitations remain, the study used retrospective datasets and requires prospective external validation across geographically and ethnically diverse cohorts before clinical deployment. Future work should evaluate real-world integration, clinician acceptance, and whether risk-guided interventions informed by this model improve patient outcomes. Nevertheless, by combining diabetes-specific features, robust resampling, semantic feature representation and ensemble learning with explicit explainability, this framework takes a concrete step toward more precise, trustworthy cardiovascular risk assessment for people living with Type-2 Diabetes.

REFERENCES

1. Cannon, A., Handelsman, Y., Heile, M. and Shannon, M., 2018. Burden of illness in type 2 diabetes mellitus. *Journal of managed care & specialty pharmacy*, *24*(9-a Suppl), pp.S5-S13.
2. Nathan, D.M., Meigs, J. and Singer, D.E., 1997. The epidemiology of cardiovascular disease in type 2 diabetes mellitus: how sweet it is… or is it?. *The Lancet*, *350*, pp.S4-S9.
3. Kovatchev, B.P., 2017. Metrics for glycaemic control—from HbA1c to continuous glucose monitoring. *Nature Reviews Endocrinology*, *13*(7), pp.425-436.
4. Xu, Z., 2022. *Optimising Cardiovascular Disease Risk Assessment: Application of Dynamic Prediction Tools and Risk Stratification Strategies Using Electronic Health Records* (Doctoral dissertation).
5. Pate, A., Emsley, R., Ashcroft, D.M., Brown, B. and Van Staa, T., 2019. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC medicine*, *17*(1), p.134.
6. Scilletta, S., Di Marco, M., Miano, N., Capuccio, S., Musmeci, M., Bosco, G., Di Giacomo Barbagallo, F., Martedì, M., La Rocca, F., Vitale, A. and Scicali, R., 2025. Cardiovascular risk profile in subjects with diabetes: Is SCORE2-Diabetes reliable?. *Cardiovascular Diabetology*, *24*(1), p.222.
7. Scilletta, S., Di Marco, M., Miano, N., Capuccio, S., Musmeci, M., Bosco, G., Di Giacomo Barbagallo, F., Martedì, M., La Rocca, F., Vitale, A. and Scicali, R., 2025. Cardiovascular risk profile in subjects with diabetes: Is SCORE2-Diabetes reliable?. *Cardiovascular Diabetology*, *24*(1), p.222.
8. Kaur, N., 2025. *Cardiovascular disease in Type 2 Diabetes Mellitus: A precision medicine approach applying artificial intelligence for heart failure and mortality prediction* (Doctoral dissertation, University of Glasgow).
9. Dziopa, K., 2023. *Data-driven approaches to cardiovascular risk prediction for people with type 2 diabetes* (Doctoral dissertation, UCL (University College London)).
10. Zhong, S., Zhang, J., Jiao, J., Zhu, H., Xing, Y. and Wang, L., 2024. A machine learning case study to predict rare clinical event of interest: imbalanced data, interpretability, and practical considerations. *Journal of Biopharmaceutical Statistics*, pp.1-14.
11. Wang, A.X., Chukova, S.S. and Nguyen, B.P., 2023. Synthetic minority oversampling using edited displacement-based k-nearest neighbors. *Applied Soft Computing*, *148,* p.110895.

12. Wang, X., Ren, J., Ren, H., Song, W., Qiao, Y., Zhao, Y., Linghu, L., Cui, Y., Zhao, Z., Chen, L. and Qiu, L., 2023. Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta. *Scientific Reports*, *13*(1), p.12718.
13. "SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe." *European heart journal* 44, no. 28 (2023): 2544-2556.
14. Kaur, N., 2025. *Cardiovascular disease in Type 2 Diabetes Mellitus: A precision medicine approach applying artificial intelligence for heart failure and mortality prediction* (Doctoral dissertation, University of Glasgow).
15. Lotfi, Z., Haji Hosseini, R. and Aminipour, M., 2025. Artificial Intelligence–Driven Approaches for Prediction, Management, and Complication Risk in Type 2 Diabetes: A Systematic Review. *InfoScience Trends*, *2*(6), pp.1-17.
16. Lin, J.S., Evans, C.V., Johnson, E., Redmond, N., Coppola, E.L. and Smith, N., 2018. Nontraditional risk factors in cardiovascular disease risk assessment: updated evidence report and systematic review for the US Preventive Services Task Force. *Jama*, *320*(3), pp.281-297.
17. Quesada, J.A., Lopez-Pineda, A., Gil-Guillén, V.F., Durazo-Arvizu, R., Orozco-Beltrán, D., López-Domenech, A. and Carratalá-Munuera, C., 2019. Machine learning to predict cardiovascular risk. *International journal of clinical practice*, *73*(10), p.e13389.
18. Schulz, H. and Behnke, S., 2012. Deep learning: Layer-wise learning of feature hierarchies. *KI-Künstliche Intelligenz*, *26*(4), pp.357-363.
19. Kaur, N., 2025. *Cardiovascular disease in Type 2 Diabetes Mellitus: A precision medicine approach applying artificial intelligence for heart failure and mortality prediction* (Doctoral dissertation, University of Glasgow).
20. Kothari, V., Stevens, R.J., Adler, A.I., Stratton, I.M., Manley, S.E., Neil, H.A. and Holman, R.R., 2002. UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. *Stroke*, *33*(7), pp.1776-1781.
21. Alfaraj, S.A., Kist, J.M., Groenwold, R.H., Spruit, M., Mook-Kanamori, D. and Vos, R.C., 2025. External validation of SCORE2-Diabetes in The Netherlands across various socioeconomic levels in native-Dutch and non-Dutch populations. *European Journal of Preventive Cardiology*, *32*(7), pp.555-563.
22. Ceriello, A., De Cosmo, S., Rossi, M.C., Lucisano, G., Genovese, S., Pontremoli, R., Fioretto, P., Giorda, C., Pacilli, A., Viazzi, F. and Russo, G., 2017. Variability in HbA1c, blood pressure, lipid parameters and serum uric acid, and risk of development of chronic kidney disease in type 2 diabetes. Diabetes, Obesity and Metabolism, 19(11), pp.1570-1578.
23. Zakir, M., Ahuja, N., Surksha, M.A., Sachdev, R., Kalariya, Y., Nasir, M., Kashif, M., Shahzeen, F., Tayyab, A., moazzam Khan, M.S. and Junejo, M., 2023. Cardiovascular complications of diabetes: from microvascular to macrovascular pathways. *Cureus*, *15*(9).