diabetology

MDPI

Article

Towards Robust Risk-Based Screening of Early-Stage Diabetes: Machine Learning Models with Union Features Selection and External Validation

Pasa Sukson [1], Watcharaporn Cholamjiak [2], Nontawat Eiamniran [2] and Mallika Khwanmuang [1,*]

[1] School of Medicine, University of Phayao, Phayao 56000, Thailand; pasa.su@up.ac.th
[2] School of Science, University of Phayao, Phayao 56000, Thailand; watcharaporn.ch@up.ac.th (W.C.); nontawatskip@gmail.com (N.E.)
* Correspondence: mallika.kh@up.ac.th

Abstract

Background/Objectives: Early-stage diabetes often presents with subtle symptoms, making timely screening challenging. This study aimed to develop an interpretable and robust machine learning framework for early-stage diabetes risk prediction using integrated statistical and machine learning–based feature selection, and to evaluate its generalizability using real-world hospital data. Methods: A Union Feature Selection approach was constructed by combining logistic regression significance testing with ReliefF and MRMR feature importance scores. Five machine learning models—Decision Tree, Naïve Bayes, SVM, KNN, and Neural Network—were trained on the UCI Early Stage Diabetes dataset (N = 520) under multiple feature-selection scenarios. External validation was performed using retrospective hospital records from the University of Phayao (N = 60). Model performance was assessed using accuracy, precision, recall, and F1-score. Results: The union feature-selection approach identified four core predictors—polyuria, polydipsia, gender, and irritability—with additional secondary features providing only marginal improvements. Among the evaluated models, Naïve Bayes demonstrated the most stable external performance, achieving 85% test accuracy, balanced precision, recall, and F1-score, along with a moderate AUC of 0.838, indicating reliable discriminative ability in real-world hospital data. In contrast, SVM, KNN, and Neural Network models, despite exhibiting very high internal validation performance (>96%) under optimally selected ML features, showed marked performance decline during external validation, highlighting their sensitivity to distributional shifts between public and clinical datasets. Conclusions: The combined statistical–ML feature selection method improved interpretability and stability in early-stage diabetes prediction. Naïve Bayes demonstrated the strongest generalizability and is well suited for real-world screening applications. The findings support the use of integrated feature selection to develop efficient and clinically relevant risk assessment tools.

Keywords: early-stage diabetes; machine learning; feature selection; external validation; risk prediction

Check for updates

Academic Editor: Peter Clifton

Received: 15 November 2025
Revised: 19 December 2025
Accepted: 24 December 2025
Published: 26 December 2025

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Diabetes mellitus (DM) is a chronic, progressive metabolic disorder characterized by persistent hyperglycemia resulting from defects in insulin secretion, insulin action, or both. It has emerged as one of the most significant global health challenges, affecting an estimated 537 million adults worldwide in 2021, with projections to reach 643 million by

2030 and 783 million by 2045 [1]. The disease contributes substantially to global morbidity and mortality, being a major cause of cardiovascular diseases, neuropathy, nephropathy, and retinopathy, thereby imposing enormous economic and social burdens [2–4]. Early detection and timely intervention in the prediabetic or early diabetic stage have been shown to significantly mitigate long-term complications and improve patient outcomes through targeted lifestyle modifications, pharmacological therapy, and preventive care [5,6].

Conventional diabetes screening typically relies on clinical and lifestyle indicators such as age, obesity, family history, hypertension, smoking, and diet patterns [7]. Although such risk scores—such as the Finnish Diabetes Risk Score (FINDRISC) or the American Diabetes Association (ADA) screening and diagnostic criteria—are simple and widely used in clinical and population-based settings, they often exhibit only moderate sensitivity and specificity, particularly across heterogeneous populations with differing genetic, lifestyle, and environmental backgrounds [8–11]. This limitation arises from the multifactorial nature of diabetes, which involves complex interactions among genetic predispositions, metabolic dysregulation, inflammatory processes, and environmental exposures [12–14]. Consequently, conventional statistical models may inadequately capture nonlinear relationships and higher-order interactions critical to disease onset, thereby reducing predictive robustness and transferability [15].

Recent advancements in Machine Learning (ML) have opened new avenues for developing accurate, data-driven diabetes risk prediction models. ML algorithms—such as Random Forest (RF), Gradient Boosting (GB), Support Vector Machines (SVM), and Deep Neural Networks (DNN)—excel in identifying subtle and nonlinear interactions among high-dimensional features, outperforming conventional logistic regression and regression tree models in various medical contexts [16–18]. Studies have demonstrated the utility of ML for both type 2 diabetes onset prediction [19], early-stage detection [20], and risk stratification based on lifestyle, metabolic, and biochemical indicators [21–23]. Furthermore, interpretable ML frameworks incorporating explainability metrics such as SHAP (SHapley Additive exPlanations) and feature importance analysis have facilitated better clinical translation and trustworthiness in model deployment [24].

Despite these advances, ensuring robustness, stability, and generalizability remains a methodological challenge in clinical ML research. Overfitting and data imbalance can lead to inflated performance in internal validation yet poor generalization on external datasets [25]. To overcome this, feature selection techniques play a vital role in model interpretability and reproducibility. The Union Feature Selection approach, which aggregates results from multiple selection methods (e.g., LASSO, mutual information, tree-based importance), has been shown to improve feature stability, reduce bias, and enhance cross-dataset consistency [26–28]. Complementarily, External Validation—using independent real-world datasets—provides the most stringent assessment of a model's generalizability, a step increasingly emphasized by the TRIPOD and PROBAST guidelines for clinical prediction models [29–31]. Recent advances in machine learning for tabular data further highlight the importance of rigorous external evaluation, as performance gains observed in complex models may not consistently translate to unseen clinical datasets without careful validation and feature interpretability considerations [32].

Previous studies on diabetes risk prediction have predominantly focused on optimizing predictive accuracy using individual machine learning algorithms or single feature-selection techniques. Many approaches rely on complex ensemble models or high-dimensional feature spaces, which often yield excellent internal validation performance but suffer from limited interpretability and reduced generalizability when applied to external clinical datasets. In addition, several studies evaluate model performance exclusively on

a single dataset, without independent external validation, thereby increasing the risk of overfitting and optimistic bias.

In contrast, the present study adopts a methodologically distinct strategy by explicitly integrating statistical significance testing with complementary machine learning–based feature selection methods to enhance feature stability and clinical interpretability. Moreover, the proposed framework places strong emphasis on external validation using independent real-world hospital data, allowing direct assessment of model robustness under distributional shifts. This methodological design addresses key limitations of prior approaches and provides a more realistic evaluation of screening performance in practical clinical and community settings.

Therefore, the present study aims to develop and externally validate a robust ML framework for early-stage diabetes risk screening by combining the publicly available UCI Early Stage Diabetes Risk Prediction dataset with retrospective hospital data from the University of Phayao Hospital. By employing Union Feature Selection and cross-domain validation, this study seeks to build a model that is not only accurate but also interpretable and clinically applicable, supporting evidence-based decision-making for diabetes prevention and management in both community and hospital settings.

In contrast to recent collaborative or federated learning approaches that primarily focus on improving feature representation across large, distributed, and heterogeneous datasets, the proposed framework emphasizes feature stability, interpretability, and real-world generalizability for early-stage diabetes screening. While collaborative learning methods are powerful in data-rich environments, they often require substantial infrastructure, complex model aggregation, and may sacrifice feature-level transparency. In this study, the union feature-selection strategy integrates statistical significance with complementary machine learning–based importance measures to identify robust and clinically meaningful predictors that remain stable across datasets. Combined with external validation using independent hospital records, this approach demonstrates strong robustness and practical applicability for early-stage screening in real-world clinical and community settings, where data availability is limited and interpretability is essential.

2. Materials and Methods

2.1. Study Design

This study employed a retrospective analytical design integrating a publicly available dataset with real-world clinical data to develop and validate machine learning models for early-stage diabetes risk screening. The statistical and preliminary analyses were conducted using amovi version 2.7 and R version 4.5, ensuring transparent and reproducible workflows [33–35].

2.2. Dataset Description

The primary dataset used for initial modeling was the UCI Early Stage Diabetes Risk Prediction Dataset [36], which consists of 520 individual records. Each record includes 17 predictor variables related to demographic, clinical, and lifestyle factors, and one binary outcome variable representing the presence ("Positive", n = 320) or absence ("Negative", n = 200) of early-stage diabetes. Although the dataset is not perfectly balanced, the class distribution remains within an acceptable range for supervised learning and was handled consistently across all analyses. The variables cover a wide range of symptoms and risk indicators, including age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity.

All categorical variables were systematically transformed into binary numerical representations to ensure consistency across statistical analyses and machine learning models, while the age variable was retained in its original continuous numerical form. A unified encoding scheme was applied to both datasets, as summarized in Table 1, with no missing values observed in any variables.

Table 1. Binary Encoding Scheme for Categorical Variables.

Variable	Encoding
Gender	Male = 1, Female = 0
Polyuria	Yes = 1, No = 0
Polydipsia	Yes = 1, No = 0
Sudden weight loss	Yes = 1, No = 0
Weakness	Yes = 1, No = 0
Polyphagia	Yes = 1, No = 0
Genital thrush	Yes = 1, No = 0
Visual blurring	Yes = 1, No = 0
Itching	Yes = 1, No = 0
Irritability	Yes = 1, No = 0
Delayed healing	Yes = 1, No = 0
Partial paresis	Yes = 1, No = 0
Muscle stiffness	Yes = 1, No = 0
Alopecia	Yes = 1, No = 0
Obesity	Yes = 1, No = 0
Class label	Positive = 1, Negative = 0

In addition, retrospective hospital records were obtained from the University of Phayao Hospital between 2023 and 2025 to serve as an external validation cohort. This dataset comprised 60 anonymized patient records, including 30 individuals clinically diagnosed with early-stage diabetes and 30 non-diabetic controls matched by age and sex where possible. Extracted variables encompassed demographic characteristics, lifestyle factors (e.g., smoking and alcohol use), medical history, and laboratory findings consistent with the predictors used in the UCI dataset. All patient identifiers were removed prior to analysis to ensure compliance with ethical standards and institutional data protection policies.

In this study, early-stage diabetes was defined as individuals who were newly diagnosed based on standard laboratory criteria at their first clinical detection and had not yet received any pharmacological or lifestyle treatment. Importantly, the proposed machine learning models were not designed to replace laboratory diagnosis. Instead, they function as symptom-based screening tools that utilize readily available demographic and clinical symptom information to identify individuals likely to be in this early diagnostic window prior to treatment initiation.

2.3. Descriptive Statistical Analysis

Descriptive statistical feature extraction was performed as an initial step to characterize predictor distributions and statistical relevance prior to machine learning model development, using jamovi (version 2.7) to ensure transparency and reproducibility. For continuous variables, summary statistics and the Shapiro–Wilk test were applied to assess central tendency, variability, and normality. For categorical variables, group-wise frequencies and chi-square tests were used to evaluate associations with the outcome. Multicollinearity was assessed via variance inflation factors (VIFs) derived from logistic regression, with all predictors showing acceptable VIF values. These statistical insights informed clinically meaningful feature identification, subsequent machine learning–based feature ranking, and the construction of compact, interpretable feature subsets.

2.3.1. Age Distribution

- Sample size (N): 520.
- Mean ± SD: 48.0 ± 12.2 years.
- Median: 47.5 years.
- Range: 16–90 years.
- Skewness: 0.329 (slightly right-skewed).
- Kurtosis: −0.192 (approximately mesokurtic).
- Normality test (Shapiro–Wilk): W = 0.983, $p < 0.001$, indicating deviation from a normal distribution.

2.3.2. Categorical Variables

The frequency analysis indicated that most clinical symptoms had unequal class distributions between diabetic ("Positive") and non-diabetic ("Negative") groups:

Table 2 presents the most frequent category, corresponding percentages, and statistical associations with diabetes classification derived from the binomial logistic regression model. Variables such as polyuria, polydipsia, polyphagia, genital thrush, irritability, and partial paresis were identified as significant predictors ($p < 0.05$), whereas delayed healing, alopecia, and obesity showed no significant associations ($p > 0.05$). These results indicate that polyuria, polydipsia, irritability, genital thrush, polyphagia, and partial paresis were key differentiating features between diabetic and non-diabetic individuals.

Table 2. Descriptive summary and statistical significance of categorical predictors in the UCI Early Stage Diabetes dataset (N = 520).

Variable	Most Frequent Category	VIF	Key Observation
Gender	Female: 37.0%, Male: 63.0%	1.98	Slight male predominance overall, but higher diabetes positivity in females (33.3%).
Polyuria	Yes (61.5%)	2.37	Strongly associated with diabetes ($\chi^2 = 67.88$, $p < 0.001$).
Polydipsia	Yes (61.6%)	1.89	Highly significant predictor ($p < 0.001$).
Sudden weight loss	Yes (41.8%)	1.55	Not statistically significant ($p = 0.729$).
Weakness	Yes (58.6%)	1.82	Weak effect ($p = 0.126$).
Polyphagia	Yes (45.5%)	1.53	Significant predictor ($p = 0.025$).
Genital thrush	Yes (22.3%)	1.60	Strongly significant ($p < 0.001$).
Itching	Yes (48.6%)	2.92	Highly significant negative association ($p < 0.001$).
Irritability	Yes (24.3%)	1.43	Highly significant predictor ($p < 0.001$).
Partial paresis	Yes (43.1%)	1.48	Significant predictor ($p = 0.027$).
Other variables (delayed healing, alopecia, obesity)	Non-significant ($p > 0.05$).		

2.4. Workflow

From Figure 1, the research workflow begins with defining the study objectives and collecting two datasets: a public early-stage diabetes dataset (N = 520) for model development and a retrospective hospital dataset (30 positive, 30 negative) for external validation. After data preprocessing, exploratory data analysis (EDA) is performed to understand variable distributions and clinical patterns.

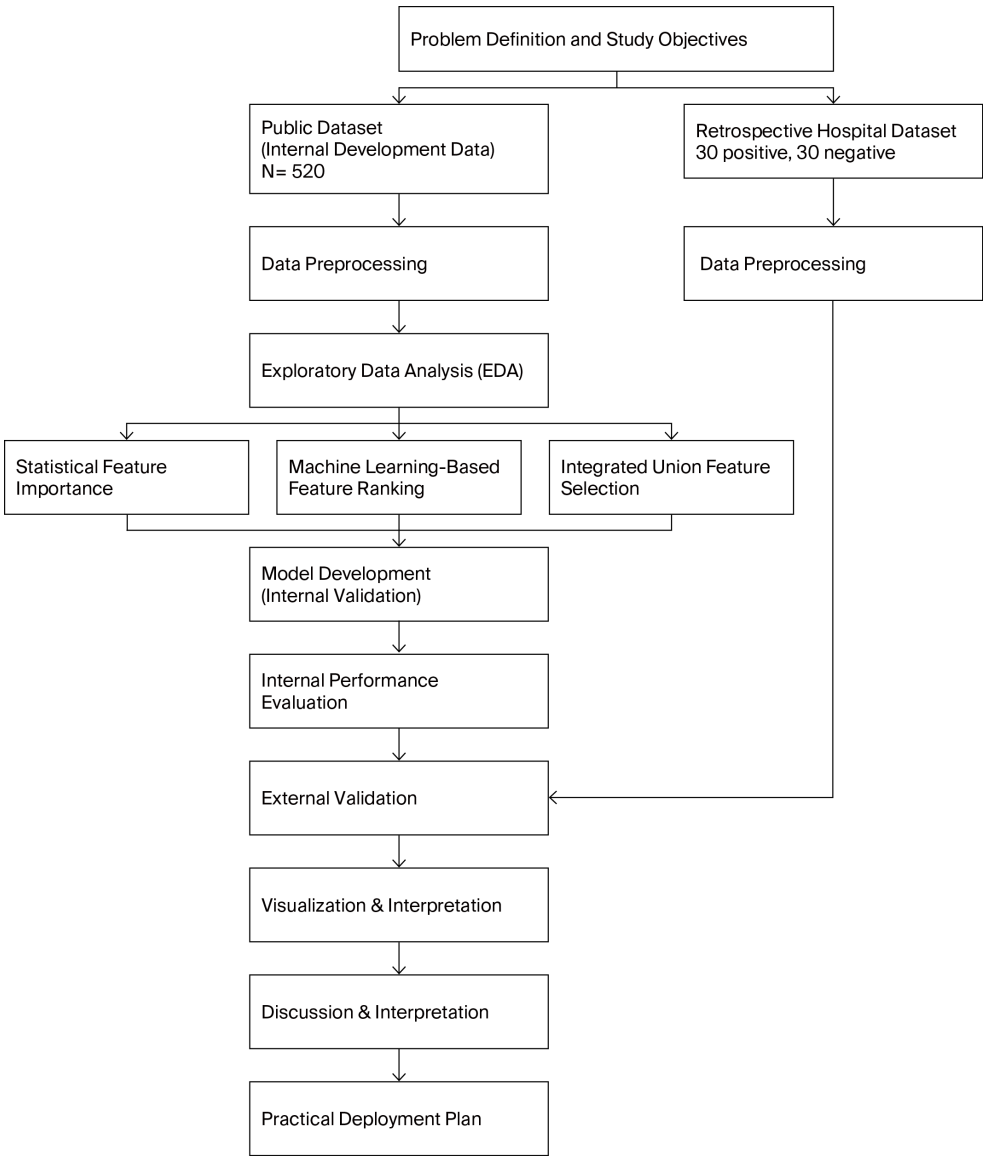


Figure 1. Research workflow of the proposed early-stage diabetes screening framework.

Feature selection is conducted through three complementary approaches—statistical significance testing, machine-learning–based feature ranking, and an integrated union feature-selection strategy—to identify the most informative predictors. These selected features are used to develop and internally validate multiple machine learning models.

Internal performance evaluation identifies the best-performing algorithms, which are then further assessed using the external hospital dataset to evaluate real-world generalizability. The results are visualized and interpreted through confusion matrices, ROC curves, and model comparisons. Finally, insights from the analyses inform the discussion, interpretation, and formulation of a practical deployment plan for real-world early-stage diabetes screening.

3. Results

3.1. Descriptive and Statistical Analysis of Predictors

The model includes 17 predictors covering demographic, clinical, and symptomatic variables. Odds ratios (OR) and 95% confidence intervals (CI) represent the likelihood of being classified as "Positive" for early-stage diabetes.

Significant predictors are defined at $p < 0.05$.

Following the initial full-model analysis in Table 3, non-significant predictors were systematically removed to improve parsimony and model interpretability. Four reduced logistic regression models were compared:

(1) Reduced model excluding Gender;
(2) Reduced model including only significant predictors;
(3) Reduced model with additional refinement on Itching and interaction terms;
(4) Reduced model excluding weakly correlated variables.

Table 3. Binomial logistic regression results of the full model for early-stage diabetes classification (N = 520).

Predictor	β (Estimate)	SE	Z	*p*-Value	OR	95% CI (Lower–Upper)
Intercept	2.747	1.076	2.55	0.011	15.59	1.89–128.31
Age	−0.051	0.025	−2.02	0.044	0.95	0.90–1.00
Gender (Male vs. Female)	−4.351	0.598	−7.27	<0.001	0.013	0.004–0.042
Polyuria (Yes vs. No)	4.440	0.705	6.30	<0.001	84.74	21.27–337.59
Polydipsia (Yes vs. No)	5.070	0.829	6.12	<0.001	159.24	31.37–808.36
Sudden weight loss	0.190	0.548	0.35	0.728	1.21	0.41–3.54
Weakness	0.817	0.537	1.52	0.126	2.26	0.79–6.48
Polyphagia (Yes vs. No)	1.194	0.534	2.24	0.025	3.30	1.16–9.39
Genital thrush (Yes vs. No)	1.864	0.553	3.37	<0.001	6.45	2.18–19.07
Visual blurring	0.916	0.651	1.41	0.160	2.50	0.70–8.96
Itching (Yes vs. No)	−2.803	0.673	−4.17	<0.001	0.061	0.016–0.227
Irritability (Yes vs. No)	2.341	0.591	3.96	<0.001	10.39	3.27–33.05
Delayed healing	−0.392	0.550	−0.71	0.476	0.68	0.23–1.99
Partial paresis (Yes vs. No)	1.159	0.525	2.21	0.027	3.19	1.14–8.92
Muscle stiffness	−0.729	0.580	−1.26	0.209	0.48	0.15–1.50
Alopecia	0.150	0.620	0.24	0.808	1.16	0.34–3.92
Obesity	−0.289	0.544	−0.53	0.595	0.75	0.26–2.18

From Table 4, the best-performing reduced model (AIC = 200, AUC = 0.980) retained nine predictors—Age, Gender, Polyuria, Polydipsia, Polyphagia, Genital thrush, Itching, Irritability, and Partial paresis—which collectively yielded excellent predictive accuracy and generalizability. This outcome suggests that most of the discriminative power can be maintained even after removing non-significant or redundant variables, thereby improving the interpretability and clinical applicability of the model for early-stage diabetes screening.

Table 4. Comparative performance metrics of the full and reduced logistic regression models for early-stage diabetes prediction.

Model	AIC	McFadden R^2	Accuracy (%)	Sensitivity	Specificity	AUC	Notes
Full model	206	0.752	93.0	0.944	0.915	0.982	Baseline full model
Reduced (Gender only)	283	0.637	89.9	0.878	0.920	0.956	Slight drop in AUC
Reduced (Significant only)	200	0.741	93.8	0.947	0.930	0.980	Best performance
Reduced (No Gender + significant)	295	0.600	87.1	0.891	0.850	0.948	Gender excluded
Reduced (Itching-refined + Gender)	223	0.704	91.0	0.934	0.880	0.973	Balanced, interpretable

3.2. Feature Selection and Model Development Using Machine Learning

To identify the most relevant and stable predictors for early-stage diabetes classification, two complementary feature selection algorithms—ReliefF and Minimum Redundancy–Maximum Relevance (MRMR)—were employed. The ReliefF algorithm evaluates each feature based on its ability to differentiate between neighboring instances of different outcome classes, effectively capturing local interactions and non-linear relationships among variables. In contrast, the MRMR algorithm optimizes feature selection by maximizing the relevance of each variable to the target class while minimizing redundancy among correlated predictors, thereby improving model interpretability and generalizability. By integrating the strengths of both methods, the study aimed to enhance feature stability, reduce multicollinearity, and identify the most discriminative subset of predictors for subsequent model development.

From Figure 2 and Table 5, the comparative analysis between the ReliefF and MRMR algorithms revealed both consistency and complementarity in identifying the most relevant predictors for early-stage diabetes classification.

- Both algorithms consistently identified polyuria and polydipsia as the top-ranked predictors, with average importance scores of 0.246 and 0.132, respectively, confirming their strong association with early hyperglycemic symptoms.
- Gender (0.124) and irritability (0.107) were also highly ranked, particularly under the MRMR algorithm, emphasizing the demographic and behavioral diversity influencing diabetes risk.
- Clinical symptoms such as genital thrush (0.064), sudden weight loss (0.062), partial paresis (0.057), and visual blurring (0.042) were moderately important, indicating potential secondary manifestations of metabolic dysregulation.
- Features with lower averaged relevance, including polyphagia, alopecia, obesity, muscle stiffness, delayed healing, age, itching, and weakness, exhibited limited discriminative power and may be excluded in model refinement phases to enhance generalizability.

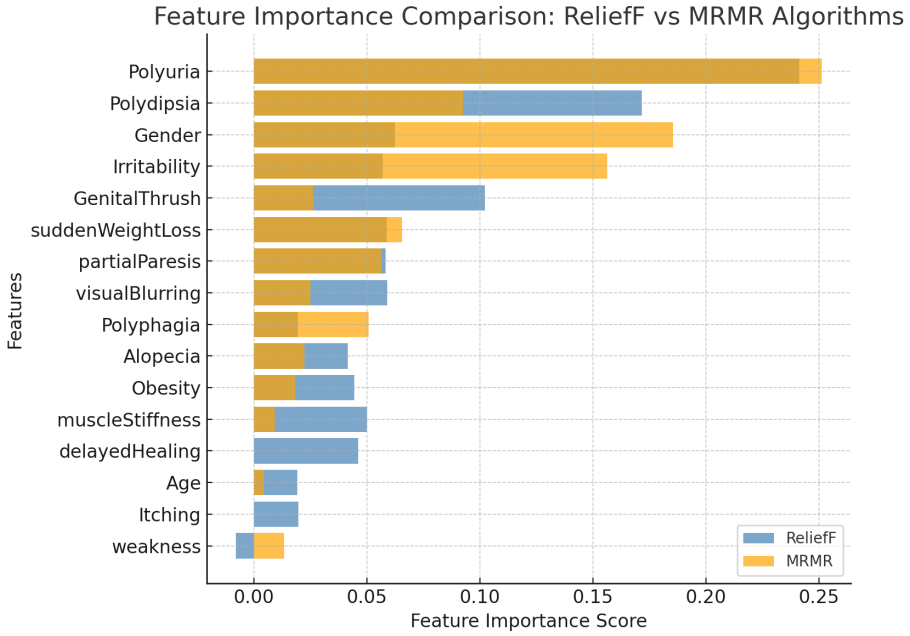


Figure 2. Comparative feature importance ranking obtained using the ReliefF and MRMR algorithms. Blue bars represent ReliefF scores, while dark yellow bars represent MRMR scores. The ReliefF method emphasizes symptom-level discriminative features, whereas MRMR enhances global feature independence and stability.

Table 5. Comparison of feature importance scores obtained from ReliefF and MRMR algorithms. The table presents individual importance values and their average scores used for subsequent union feature selection.

Feature	ReliefF_Score	MRMR_Score	Average_Score
Polyuria	0.2412	0.2511	0.24615
Polydipsia	0.1717	0.0925	0.1321
Gender	0.0624	0.1854	0.1239
Irritability	0.057	0.1564	0.1067
GenitalThrush	0.1022	0.0261	0.06415
suddenWeightLoss	0.0588	0.0655	0.06215
partialParesis	0.0583	0.0563	0.0573
visualBlurring	0.0589	0.0251	0.042
Polyphagia	0.0195	0.0506	0.03505
Alopecia	0.0415	0.0224	0.03195
Obesity	0.0445	0.0182	0.03135
muscleStiffness	0.0501	0.0092	0.02965
delayedHealing	0.046	0	0.023
Age	0.0192	0.0044	0.0118
Itching	0.0196	0	0.0098
weakness	−0.008	0.0133	0.00265

Overall, the integration of ReliefF and MRMR results strengthened feature stability by combining local relevance detection (ReliefF) with redundancy minimization (MRMR).

Table 6, each case corresponds to a different feature inclusion threshold (Case 1 > 0.04, Case 2 > 0.03, Case 3 > 0.02, and Case 4 = all features).

Table 6. Classification of feature selection scenarios based on averaged importance scores derived from ReliefF and MRMR algorithms.

Case	Criteria (Average Score)	No. of Features	Purpose
1	>0.04	8	Minimal, high-impact features for compact model
2	>0.03	11	Balanced, interpretable model
3	>0.02	13	Comprehensive model for clinical settings
4	All features	16	Baseline reference model

Table 7, all models achieved high accuracy (≥90%). The Naive Bayes and SVM consistently performed best, with SVM in Case 3 (97.31%) offering the optimal balance between accuracy and training time, while Naive Bayes in Case 4 (97.69%) reached the highest overall accuracy but required longer computation.

Table 7. Performance comparison of five machine learning models (Tree, SVM, KNN, Efficient Linear, and Neural Network) across four feature selection scenarios.

	Model Type	Training Time (Sec)	Accuracy %	Micro Precision %	Micro Recall %	Micro F1 Score %
Case 1	Tree	616.3994	92.69	56.70	91.99	92.84
	Naive Bayes	645.5716	89.81	83.94	89.21	89.28
	SVM	587.6629	93.65	64.35	93.26	93.34
	KNN	31.1410	93.65	64.35	93.05	93.72
	Neural Network	277.0775	93.46	60.18	92.87	93.47

Table 7. *Cont.*

	Model Type	Training Time (Sec)	Accuracy %	Micro Precision %	Micro Recall %	Micro F1 Score %
Case 2	Tree	13.3348	93.27	66.48	92.51	93.78
	Naive Bayes	40.3338	90.00	83.48	89.62	89.16
	SVM	37.2573	95.77	68.75	95.39	95.72
	KNN	19.6883	95.96	71.13	96.07	96.57
	Neural Network	365.3940	94.81	58.57	94.41	94.66
Case 3	Tree	17.7640	94.23	94.23	94.23	94.23
	Naive Bayes	27.2720	90.00	90.00	90.00	90.00
	SVM	26.5675	97.31	97.31	97.31	97.31
	KNN	14.9959	94.81	94.81	94.81	94.81
	Neural Network	226.7519	95.19	95.19	95.19	95.19
Case 4	Tree	18.3220	93.65	59.16	93.01	93.81
	Naive Bayes	64.9094	89.42	83.48	89.04	88.50
	SVM	115.2744	97.31	67.33	97.24	97.06
	KNN	80.2074	97.12	74.31	96.91	97.00
	Neural Network	203.9738	96.73	68.75	96.51	96.59

3.3. Union Feature Selection and Integrated Model Construction

Table 8 summarizes the integrated ranking of all candidate features using a combined approach that incorporates both statistical significance and machine learning–based feature importance. Specifically, the ranking was derived by jointly evaluating:

- Statistical evidence from logistic regression (*p*-value and Odds Ratio), representing clinical relevance and strength of association; and
- Machine learning importance from the ReliefF and MRMR algorithms, capturing nonlinear interactions and redundancy-adjusted relevance.

Table 8. Statistical and machine learning–based feature importance analysis for early-stage diabetes prediction.

Rank	Feature	*p*-Value	OR	ReliefF	MRMR	Average ML	Significant	Combined Score	Importance
1	Polyuria	0.001	84.74	0.2412	0.2511	0.24615	Yes	0.622	Very High
2	Polydipsia	0.001	159.24	0.1717	0.0925	0.13210	Yes	0.566	Very High
3	Irritability	0.001	10.39	0.0570	0.1564	0.10670	Yes	0.554	Very High
4	Gender	0.001	0.013	0.0624	0.1854	0.12390	Yes	0.562	Very High
5	Genital thrush	0.001	6.45	0.1022	0.0261	0.06415	Yes	0.532	High
6	Partial paresis	0.027	3.19	0.0583	0.0563	0.05730	Yes	0.515	High
7	Sudden weight loss	0.728	1.21	0.0588	0.0655	0.06215	No	0.167	Low
8	Visual blurring	0.160	2.50	0.0589	0.0251	0.04200	No	0.441	Moderate
9	Polyphagia	0.025	3.30	0.0195	0.0506	0.03505	Yes	0.505	Moderate
10	Alopecia	0.808	1.16	0.0415	0.0224	0.03195	No	0.112	Low
11	Obesity	0.595	0.75	0.0445	0.0182	0.03135	No	0.218	Low
12	Muscle stiffness	0.209	0.48	0.0501	0.0092	0.02965	No	0.410	Moderate
13	Age	0.044	0.95	0.0192	0.0044	0.01180	Yes	0.484	Moderate
14	Delayed healing	0.476	0.68	0.0460	0.0000	0.02300	No	0.262	Low
15	Itching	0.001	0.061	0.0196	0.0000	0.00980	Yes	0.505	Moderate
16	Weakness	0.126	2.26	−0.0080	0.0133	0.00265	No	0.437	Moderate

A composite score was generated to reflect each variable's overall predictive value, thereby supporting a robust and interpretable Union Feature Selection framework for early-stage diabetes risk classification.

Using the combined scoring formula, defined as Combined Score = 0.5(1 − p) + 0.5 (Average ML Score), the integrated ranking identified polyuria, polydipsia, gender, and irritability as the most influential predictors. These variables consistently exhibited strong statistical significance alongside high machine learning importance. Collectively, they reflect classical early manifestations of hyperglycemia and key patient-level characteristics known to affect metabolic risk. A second tier of moderately important features—including genital thrush, partial paresis, and polyphagia—showed either statistical significance or meaningful machine learning scores, indicating their supportive role in improving model discrimination, particularly in nonlinear classifiers.

- Several features such as muscle stiffness, visual blurring, age, and itching exhibited moderate combined scores. Although not always statistically significant, their machine learning contributions suggest potential value in capturing subtle behavioral or physiological patterns.
- Conversely, variables such as obesity, alopecia, delayed healing, and weakness ranked lowest in combined importance and demonstrated limited discriminative power. These features may still be useful in specific clinical contexts but provide minimal incremental benefit for automated screening models.
- Overall, the combined statistical–machine learning ranking supports a parsimonious feature set centered on the most influential predictors while maintaining flexibility to incorporate moderate-level variables depending on model objectives and computational constraints. This integrated feature selection strategy enhances both model interpretability and predictive robustness for early-stage diabetes risk assessment.

To enable a structured comparison of predictive performance under varying levels of model complexity, three feature-selection scenarios were defined based on the integrated importance ranking:

1. Scenario 1: Core highly influential features.
   This scenario included only the most influential predictors, namely polyuria, polydipsia, gender, and irritability. These variables represent the strongest combined signals across both statistical and machine-learning evaluation and serve as the foundational feature set for a minimal yet high-impact screening model.
2. Scenario 2: Core features plus secondary-importance predictors.
   In this configuration, the core influential features were combined with secondary-importance variables, including genital thrush and partial paresis. This expanded set was designed to determine whether moderately strong predictors contribute additional discriminatory power beyond the core model.
3. Scenario 3: Comprehensive model including moderate-importance features.
   The final cenario incorporated the core, secondary, and moderate-importance features, adding polyphagia, muscle stiffness, visual blurring, age, and itching. This comprehensive feature set aimed to evaluate the upper bound of model performance when broader physiological and demographic indicators are included.

Table 9 presents the training time, validation accuracy, and test-set performance—including precision, recall, and F1-score—of five machine learning models (Decision Tree, Naïve Bayes, SVM, KNN, and Neural Network). Three scenarios were evaluated, each corresponding to a different subset of features ranked by machine learning–based feature importance. Scenario 3, which incorporated the most informative feature subset, yielded

the highest predictive performance across multiple classifiers, with SVM, KNN, and Neural Network achieving accuracy, precision, recall, and F1-scores exceeding 96%.

Table 9. Internal validation performance of machine learning classifiers across three scenarios based on ML-driven feature selection.

	Model Type	Training Time (Sec)	Accuracy %	Precision % (Test)	Recall % (Test)	F1 Score % (Test)
Scenario 1	Tree	30.9569	89.42	89.42	89.42	89.42
	Naive Bayes	23.4322	86.73	86.73	86.73	86.73
	SVM	20.3376	89.42	89.42	89.42	89.42
	KNN	18.4403	89.04	89.04	89.04	89.04
	Neural Network	443.7588	89.42	89.42	89.42	89.42
Scenario 2	Tree	18.0810	89.04	89.04	89.04	89.04
	Naive Bayes	25.1962	88.46	88.46	88.46	88.46
	SVM	17.3098	90.00	90.00	90.00	90.00
	KNN	18.0656	91.15	91.15	91.15	91.15
	Neural Network	219.9881	89.04	89.04	89.04	89.04
Scenario 3	Tree	15.0901	91.92	91.92	91.92	91.92
	Naive Bayes	31.5189	87.88	87.88	87.88	87.88
	SVM	49.0944	96.54	96.54	96.54	96.54
	KNN	17.9591	97.88	97.88	97.88	97.88
	Neural Network	140.4526	96.92	96.92	96.92	96.92

3.4. External Validation Using Retrospective Hospital Records

To further evaluate the generalizability of the developed machine learning models, an external validation was conducted using retrospective hospital records. In this section, feature subsets were constructed based solely on machine learning–based feature importance rankings derived from the internal training dataset. Five distinct feature-selection cases were generated to investigate how different combinations of top-ranked features influence predictive performance when applied to real-world clinical data. The selected models were then evaluated on an independent hospital dataset to assess their robustness against distributional shifts and their ability to retain discriminative capability outside the training environment.

Because all diabetic cases in the external validation cohort represented newly diagnosed and untreated individuals, the performance metrics reported in this section specifically reflect the ability of the proposed models to screen for early-stage diabetes, rather than to predict established or treated disease.

3.4.1. Machine Learning–Based Feature Importance

From Table 10, the table reports the performance of five classifiers—Decision Tree, Naïve Bayes, SVM, KNN, and Neural Network—evaluated using hospital retrospective records. Five ML-based feature selection cases were tested to assess the influence of different feature subsets on model generalization. Performance metrics include validation accuracy, test accuracy, precision, recall, and F1-score. Naïve Bayes consistently provided the strongest balance between validation and test performance, achieving the highest overall stability across all cases.

- Beyond conventional accuracy-based metrics, additional imbalance-aware evaluations further clarified the relative robustness of the examined models under external validation. In particular, Naïve Bayes consistently achieved the highest balanced accuracy across all five feature selection cases, with values ranging from approximately 81.7%

to 83.8%. This indicates stable sensitivity to both diabetic and non-diabetic classes when applied to unseen hospital data.

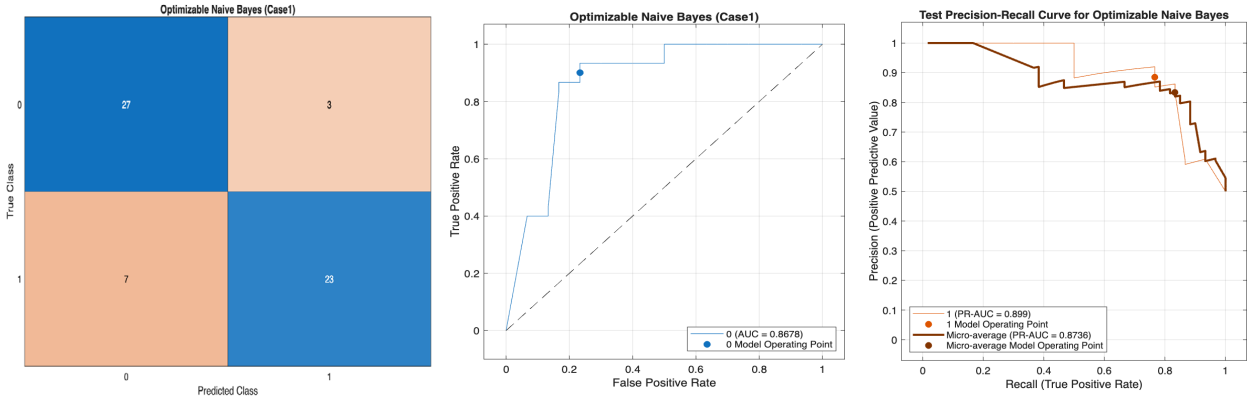
- Moreover, Naïve Bayes yielded the strongest Matthews correlation coefficient (MCC), with values between 0.636 and 0.673, substantially outperforming the other classifiers. These results reflect a high level of agreement between predicted and true class labels beyond random chance. In contrast, Decision Tree and Neural Network models exhibited markedly lower balanced accuracy (typically below 66%) and weak MCC values ($\leq$0.315), despite showing high validation accuracy. SVM and KNN demonstrated moderate and case-dependent improvements in balanced accuracy; however, their MCC values remained consistently inferior to those of Naïve Bayes, suggesting less reliable overall classification performance across different feature subsets.

Table 10. External validation results of machine learning models using feature subsets derived from ML-based feature importance across five cases.

	Model Type	Accuracy % (Validation)	Accuracy % (Test)	Accuracy % (Balance)	Precision % (Test)	Recall % (Test)	F1 Score % (Test)	MCC
Case 1	Tree	92.69	56.67	56.70	56.70	56.67	56.62	0.134
	Naive Bayes	89.81	83.33	83.30	83.94	83.33	83.26	0.673
	SVM	93.65	63.33	63.30	64.35	63.33	62.67	0.277
	KNN	93.65	63.33	63.30	64.35	63.33	62.67	0.277
	Neural Network	93.46	60.00	60.00	60.18	60.00	59.82	0.202
Case 2	Tree	93.27	65.00	65.00	66.48	65.00	64.19	0.315
	Naive Bayes	90.00	83.33	83.30	83.48	83.33	83.31	0.673
	SVM	95.77	66.67	66.70	68.75	66.67	65.71	0.354
	KNN	95.96	65.00	65.00	71.13	69.38	69.32	0.323
	Neural Network	94.81	58.33	58.30	58.57	58.33	58.04	0.169
Case 3	Tree	94.23	65.00	65.00	66.48	65.00	64.19	0.315
	Naive Bayes	90.00	81.67	81.70	81.99	81.67	81.62	0.636
	SVM	97.31	66.67	66.70	68.75	66.67	65.71	0.354
	KNN	94.81	61.67	61.70	62.34	61.67	61.14	0.240
	Neural Network	95.19	66.67	66.70	68.75	66.67	65.71	0.354
Case 4	Tree	93.65	58.33	58.30	59.16	58.33	57.37	0.175
	Naive Bayes	89.42	83.33	83.80	83.48	83.33	83.31	0.668
	SVM	97.31	65.00	65.00	67.33	65.00	63.78	0.323
	KNN	97.12	73.33	73.30	74.31	73.33	73.06	0.476
	Neural Network	96.73	66.67	66.70	68.75	66.67	65.71	0.354

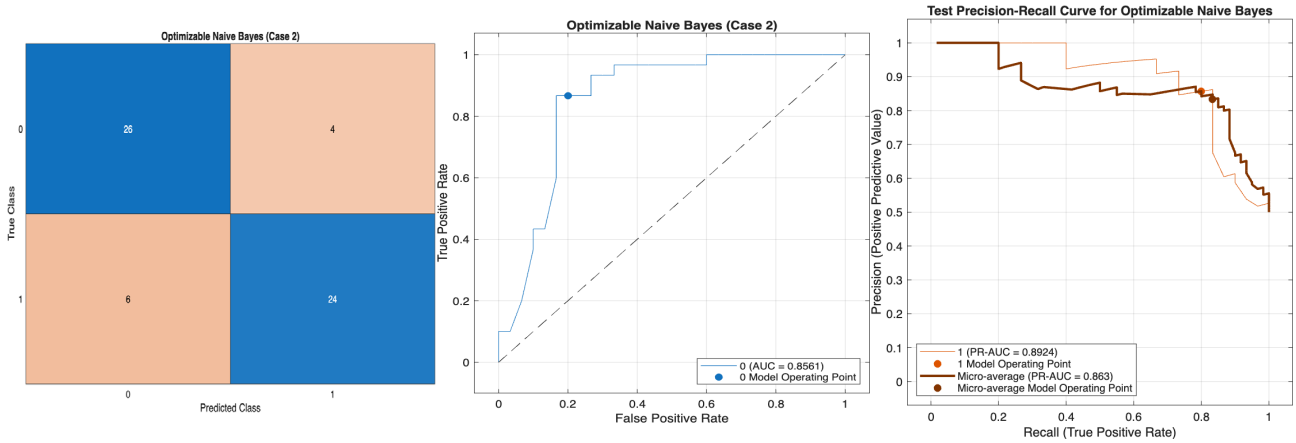
Across Cases 1–4, the Optimizable Naïve Bayes model consistently demonstrated strong and stable classification performance, correctly classifying approximately 50 out of 60 retrospective hospital records in each case. The confusion matrices indicate a balanced trade-off between sensitivity and specificity, with reliable detection of diabetic cases and effective exclusion of non-diabetic individuals across varying feature subsets. The ROC analyses further confirm robust discriminative ability, with AUC values ranging from 0.856 to 0.868 across all four cases, reflecting consistent separation between positive and negative

classes. In addition, the precision–recall curves exhibit favorable precision–recall trade-offs under class imbalance, indicating that high precision is maintained even at increased recall levels. Collectively, these results demonstrate the robustness and generalizability of the Optimizable Naïve Bayes model across different feature configurations, supporting its suitability as a reliable screening model for early-stage diabetes using real-world hospital data. Figures 3–6 present a comprehensive performance evaluation of the Optimizable Naïve Bayes model across the four experimental cases, including confusion matrices, ROC curves, and precision–recall (PR) curves.



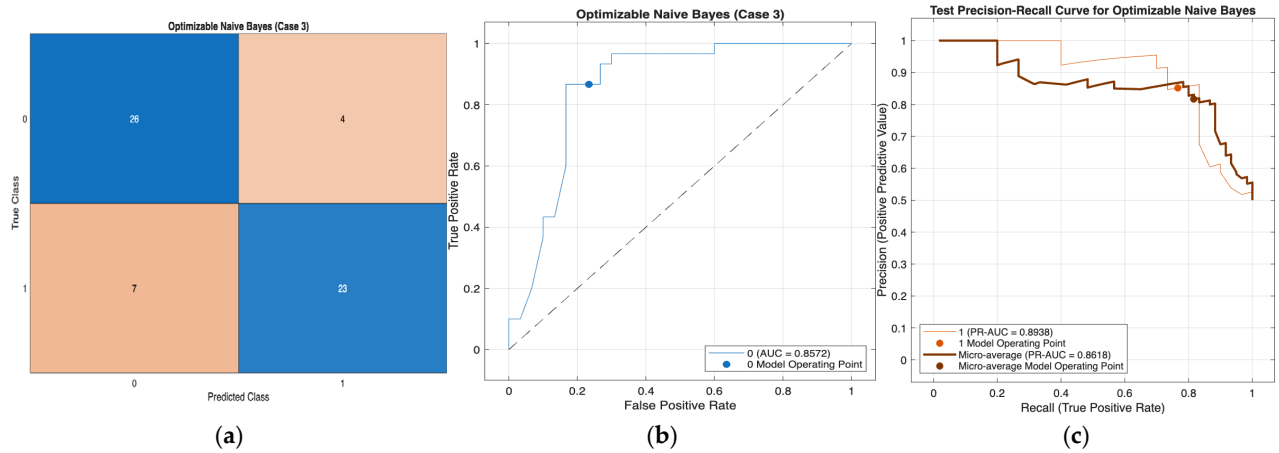
(**a**) (**b**) (**c**)

Figure 3. Performance evaluation of the Optimizable Naïve Bayes model for Case 1 in Table 10: (**a**) confusion matrix, (**b**) receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.8678, and (**c**) precision–recall (PR) curve illustrating the trade-off between precision and recall under class imbalance. In (**b**), the blue curve represents the ROC performance of the model, and the black dotted line indicates random classification. In (**c**), the colored curves represent class-wise and micro-averaged PR performance, and the circular markers denote the selected model operating point.



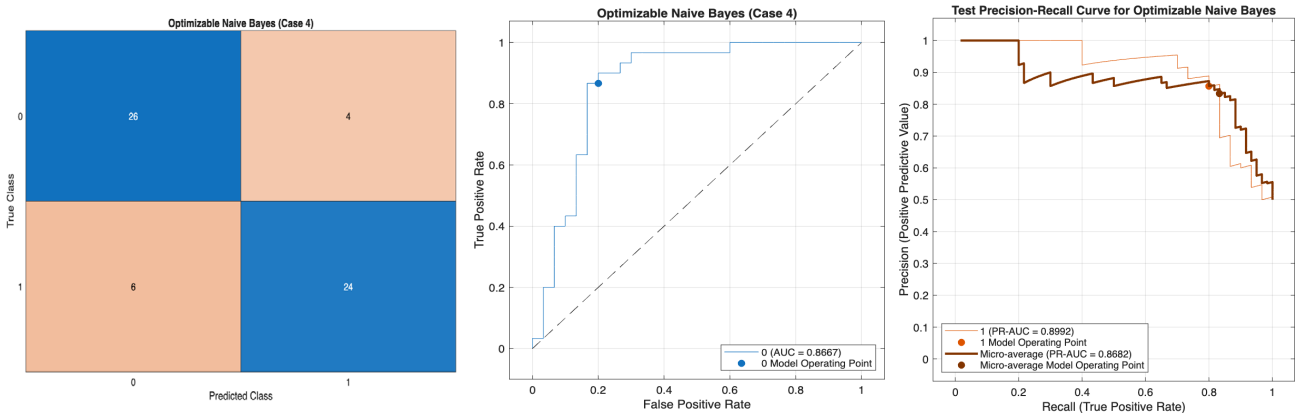
(**a**) (**b**) (**c**)

Figure 4. Performance evaluation of the Optimizable Naïve Bayes model for Case 2 in Table 10: (**a**) confusion matrix, (**b**) receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.8561, and (**c**) precision–recall (PR) curve illustrating the trade-off between precision and recall under class imbalance. In (**b**), the blue curve represents the ROC performance of the model, and the black dotted line indicates random classification. In (**c**), the colored curves represent class-wise and micro-averaged PR performance, and the circular markers denote the selected model operating point.



(**a**) (**b**) (**c**)

Figure 5. Performance evaluation of the Optimizable Naïve Bayes model for Case 3 in Table 10: (**a**) confusion matrix, (**b**) receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.8572, and (**c**) precision–recall (PR) curve illustrating the trade-off between precision and recall under class imbalance. In (**b**), the blue curve represents the ROC performance of the model, and the black dotted line indicates random classification. In (**c**), the colored curves represent class-wise and micro-averaged PR performance, and the circular markers denote the selected model operating point.



(**a**) (**b**) (**c**)

Figure 6. Performance evaluation of the Optimizable Naïve Bayes model for Case 4 in Table 10: (**a**) confusion matrix, (**b**) receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.8667, and (**c**) precision–recall (PR) curve illustrating the trade-off between precision and recall under class imbalance. In (**b**), the blue curve represents the ROC performance of the model, and the black dotted line indicates random classification. In (**c**), the colored curves represent class-wise and micro-averaged PR performance, and the circular markers denote the selected model operating point.

3.4.2. Union Feature Selection (Statistical and Machine Learning–Based Feature Importance)

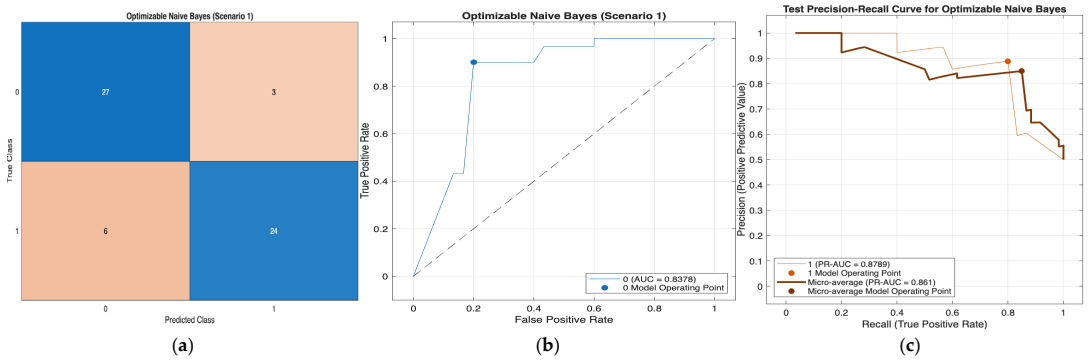
Table 11 presents the predictive performance of five machine learning classifiers—Decision Tree, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Neural Network—evaluated using retrospective hospital records. Three feature-selection scenarios were assessed: Scenario 1 (core highly influential features), Scenario 2 (core + secondary features), and Scenario 3 (core + secondary + moderate features). Performance metrics include validation accuracy, test accuracy, precision, recall, and F1 score. Naïve Bayes consistently achieved the highest test accuracy (85%) and demonstrated the most stable precision, recall, and F1 score across all scenarios.

- Beyond standard accuracy-based evaluation, the imbalance-aware metrics in Table 11 further emphasize the robustness of the screening models under external validation. Across all three feature-selection scenarios, Naïve Bayes consistently achieved the highest balanced accuracy (85.0%) and the strongest Matthews correlation coefficient (MCC = 0.703), indicating reliable discrimination between diabetic and non-diabetic classes.
- In contrast, Decision Tree, SVM, KNN, and Neural Network models exhibited noticeably lower balanced accuracy, ranging from approximately 60.0% to 70.0%, with corresponding MCC values remaining below 0.41. Although some models, particularly Decision Tree in Scenario 3, demonstrated moderate improvements in balanced accuracy, their overall agreement between predicted and true labels was substantially weaker than that of Naïve Bayes. These results confirm the superior and stable generalization performance of Naïve Bayes across all union-based feature selection scenarios.

Table 11. External validation results of screening models constructed using Union Feature Selection across three scenarios.

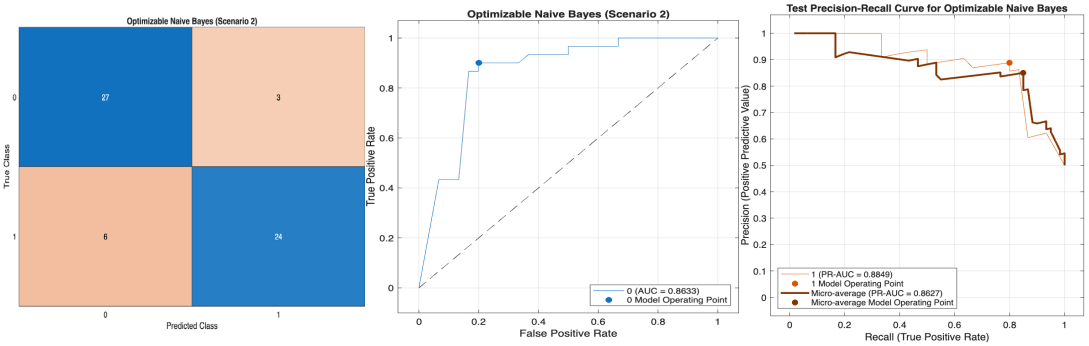
	Model Type	Accuracy % (Validation)	Accuracy % (Test)	Accuracy % (Balance)	Precision % (Test)	Recall % (Test)	F1 Score % (Test)	MCC
Scenario 1	Tree	89.42	65.00	65.00	66.48	65.00	64.19	0.315
	Naive Bayes	86.73	85.00	85.00	85.35	85.00	84.96	0.703
	SVM	89.42	63.33	63.30	65.87	63.33	61.81	0.291
	KNN	89.04	65.00	65.00	66.48	65.00	64.19	0.315
	Neural Network	89.42	63.33	63.30	65.87	63.33	61.81	0.291
Scenario 2	Tree	89.04	60.00	60.00	61.25	60.00	58.86	0.212
	Naive Bayes	88.46	85.00	85.00	85.35	85.00	84.96	0.703
	SVM	90.00	66.67	66.70	67.94	66.67	66.06	0.346
	KNN	91.15	63.33	63.30	64.35	63.33	62.67	0.277
	Neural Network	89.04	61.67	61.70	62.00	61.67	61.40	0.237
Scenario 3	Tree	91.92	70.00	70.00	70.83	70.00	69.70	0.408
	Naive Bayes	87.88	85.00	85.00	85.35	85.00	84.96	0.703
	SVM	96.54	66.67	66.70	68.75	66.67	65.71	0.346
	KNN	97.88	66.67	66.70	68.75	66.67	65.71	0.346
	Neural Network	96.92	66.67	66.70	71.31	66.67	64.75	0.377

Across all three scenarios, the Optimizable Naïve Bayes model demonstrated consistently strong and stable performance for early-stage diabetes classification using retrospective hospital records, as illustrated in Figures 7–9. The confusion matrices indicate balanced sensitivity and specificity, while ROC and precision–recall analyses confirm robust discriminative ability under class imbalance, with AUC values exceeding 0.83 in all scenarios. Model performance improved progressively from Scenario 1 to Scenario 3 as additional statistically and machine-learning–selected features were incorporated, supporting the effectiveness of the union feature-selection strategy. Nevertheless, Scenario 1 achieved competitive results with a reduced feature set, highlighting the strong discriminative power of core predictors. Overall, these findings demonstrate the strong generalizability and practical suitability of the Optimizable Naïve Bayes model for real-world early-stage diabetes screening and its potential integration into clinical decision-support workflows.



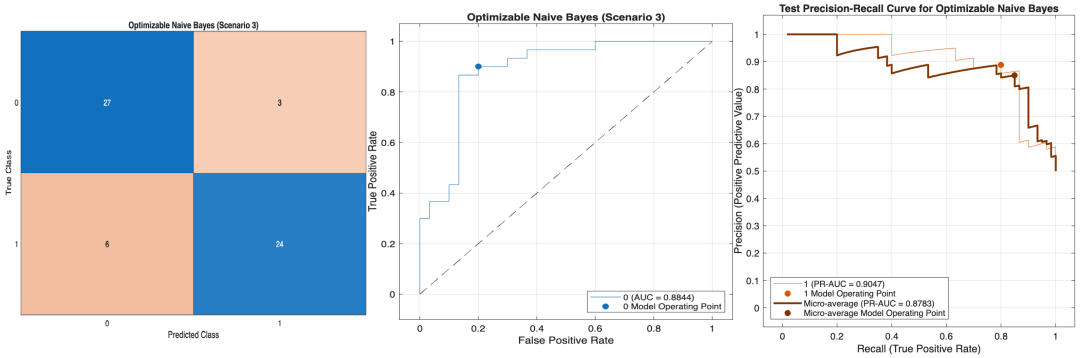
(**a**) (**b**) (**c**)

Figure 7. Performance evaluation of the Optimizable Naïve Bayes model for Scenario 1 in Table 11: (**a**) confusion matrix, (**b**)receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.8378, and (**c**) precision–recall (PR) curve illustrating the trade-off between precision and recall under class imbalance. In (**b**), the blue curve represents the ROC performance of the model, and the black dotted line indicates random classification. In (**c**), the colored curves represent class-wise and micro-averaged.



(**a**) (**b**) (**c**)

Figure 8. Performance evaluation of the Optimizable Naïve Bayes model for Scenario 2 in Table 11: (**a**) confusion matrix, (**b**)receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.8633, and (**c**) precision–recall (PR) curve illustrating the trade-off between precision and recall under class imbalance. In (**b**), the blue curve represents the ROC performance of the model, and the black dotted line indicates random classification. In (**c**), the colored curves represent class-wise and micro-averaged.



(**a**) (**b**) (**c**)

Figure 9. Performance evaluation of the Optimizable Naïve Bayes model for Scenario 3 in Table 11: (**a**) confusion matrix, (**b**)receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.8844, and (**c**) precision–recall (PR) curve illustrating the trade-off between precision and recall under class imbalance. In (**b**), the blue curve represents the ROC performance of the model, and the black dotted line indicates random classification. In (**c**), the colored curves represent class-wise and micro-averaged.

4. Discussion

The comparative evaluation between the full logistic regression model and multiple reduced variants offers important insights into the relative contribution of demographic and symptomatic factors to early-stage diabetes prediction. The full model, consisting of 17 predictors, demonstrated outstanding discriminative ability (AUC = 0.982), reinforcing that the combined effects of clinical, behavioral, and demographic factors effectively capture early pathophysiological changes associated with diabetes onset.

It is important to emphasize that the proposed framework is intended for early-stage diabetes screening rather than definitive diagnosis. While laboratory testing remains the diagnostic gold standard, symptom-based machine learning screening may support early identification of individuals in the initial diagnostic phase, thereby enabling timely confirmatory testing and early intervention before treatment initiation.

After systematic feature reduction, several reduced models retained strong predictive capability, indicating that some predictors contribute more substantially than others. Removing gender resulted in a modest decrease in AUC (0.956), suggesting that sex-related biological or behavioral differences play a meaningful role in risk stratification within this population. Conversely, the reduced model containing only statistically significant predictors achieved near-equivalent performance (AUC = 0.980) while improving accuracy (93.8%) and maintaining balanced sensitivity (0.947) and specificity (0.930). These results emphasize that a simplified model can preserve predictive strength while improving interpretability and mitigating overfitting.

Across analyses, polyuria, polydipsia, polyphagia, genital thrush, irritability, partial paresis, age, and gender consistently emerged as influential predictors. These observations align with existing evidence that excessive urination and thirst are early indicators of glycemic dysregulation, while irritability and generalized weakness reflect metabolic instability. The inverse trend observed for itching suggests potential symptom overlap with non-diabetic dermatologic conditions, meriting further clinical investigation.

Methodologically, the union feature-selection framework—integrating logistic regression significance with machine-learning–based importance (ReliefF and MRMR)—proved highly effective. This approach enhanced model robustness by removing redundant or low-impact variables while preserving features with strong statistical and predictive relevance. Such parsimony is especially valuable when transitioning to machine-learning pipelines, where excessive dimensionality can degrade generalization during external validation.

All reduced logistic models achieved AUC values above 0.94, demonstrating high stability even after substantial feature pruning. This highlights the feasibility of designing lightweight yet accurate screening tools suitable for primary care and community settings where comprehensive data collection may be impractical.

The machine-learning experiments further reinforced these findings. Across four feature-selection scenarios, all models achieved high internal accuracy ($\geq$90%) with balanced micro-averaged precision, recall, and F1-scores. SVM and KNN consistently yielded top performance, with SVM achieving the highest accuracy (97.69%) using all features, while KNN (Case 3) offered an efficient balance between accuracy (97.31%) and computational cost. Neural networks performed strongly but required substantial training time, whereas Decision Tree and Efficient Linear models provided interpretable alternatives with slightly lower accuracy.

Importantly, reducing features did not substantially degrade performance, confirming the effectiveness of the union selection framework in isolating core predictive variables. This finding suggests that simplified models can still achieve high diagnostic accuracy while improving feasibility for real-world deployment.

External validation with retrospective hospital records further underscored this point. Among all algorithms, Naïve Bayes consistently demonstrated the strongest generalization across all scenarios, maintaining test accuracy above 83% with balanced precision, recall, and F1-scores. This finding is consistent with previous diabetes screening studies reporting that Naïve Bayes performs competitively—or even outperforms more complex models—when applied to tabular clinical data, particularly under limited sample sizes and class imbalance conditions [15,37].

Notably, increasing the number of features did not yield additional performance gains, suggesting that a compact subset of clinically meaningful predictors is sufficient for practical early-stage diabetes screening. Similar observations have been reported in prior ML-based screening research, where simpler probabilistic models preserved robustness and interpretability while avoiding overfitting [32,37]. In contrast, models such as SVM, KNN, and neural networks exhibited decreased performance on external hospital data, indicating sensitivity to dataset shift and reduced transferability, despite strong internal validation results. This behavior aligns with recent findings that complex tabular models may overfit benchmark datasets and fail to generalize without rigorous external validation [32].

The superior performance of Naïve Bayes in terms of balanced accuracy and Matthews correlation coefficient (MCC) underscores its robustness under class imbalance and external validation conditions. Unlike accuracy alone, these imbalance-aware metrics provide a more clinically meaningful evaluation by accounting for both false-positive and false-negative predictions. The consistently high balanced accuracy and MCC values indicate strong predictive agreement beyond chance, even as feature subsets vary or expand across different scenarios.

In contrast, the discrepancy between high validation accuracy and poor imbalance-aware performance observed in Decision Tree and Neural Network models suggests potential overfitting and bias toward the majority class. Although SVM and KNN benefited from specific feature configurations, their performance stability remained inferior to that of Naïve Bayes, particularly when evaluated using MCC. Collectively, these findings reinforce Naïve Bayes as a reliable, stable, and interpretable baseline screening model for early-stage diabetes risk assessment using retrospective hospital datasets, especially in real-world clinical environments where feature heterogeneity and class imbalance are common.

Collectively, these findings highlight the importance of model simplicity, generalizability, and clinical interpretability when designing practical screening systems. The union feature-selection strategy, combined with external validation, provides a robust framework for identifying reliable predictors and ensuring model stability across populations. The results indicate that Naïve Bayes, KNN (Case 3), and SVM (Case 4) represent promising candidates for real-world deployment in hospital and community-based early diabetes risk assessment.

5. Conclusions

This study demonstrates that early-stage diabetes can be accurately predicted using a compact set of clinically meaningful features derived through the proposed union feature-selection framework. Logistic regression and multiple machine-learning models achieved high internal performance, and reducing the number of predictors did not compromise accuracy—highlighting the robustness of the selected core features. External validation using retrospective hospital records further confirmed that Naïve Bayes provided the most stable and generalizable performance across all scenarios, indicating strong resilience to real-world data variability. Overall, the findings support the development of lightweight, interpretable, and computationally efficient screening tools suitable for integration into clinical and community health workflows to facilitate early detection and timely intervention

for diabetes. Importantly, this model is intended as a risk-based screening tool rather than a diagnostic instrument, supporting early identification and prioritization for confirmatory laboratory testing rather than replacing standard diagnostic procedures.

Author Contributions: Conceptualization, P.S.; methodology and software, W.C. and N.E.; validation, P.S., W.C. and M.K.; formal analysis, W.C.; data curation, P.S.; writing—original draft preparation, W.C.; writing—review and editing, P.S. and M.K.; supervision, P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the revenue budget in 2025, School of Medicine, University of Phayao (Grant No. MD68-17) and Thailand Science Research and Innovation Fund (Fundamental Fund 2026, Grant No. 2253/2568).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, the Belmont report, CIOMS Guideline international conference on Harmonization in Good Cinical Pratice or ICH-GCP and with approval from the Human Research Ethics Committee of University of Phayao on Health Sciences and Technology (Institutional Review Board (IRB) approval, IRB Number: HREC-UP-HSST 1.1/002/69, approved on 17 October 2025).

Informed Consent Statement: This study has been reviewed and approved by The Human Research Ethics Committee of University of Phayao on Health Sciences and Technology and has been classified as Exempted, as it involves secondary analysis of anonymized data.

Data Availability Statement: The UCI Early Stage Diabetes Risk Prediction dataset is publicly available at https://doi.org/10.24432/C5VG8H. The retrospective hospital dataset from the University of Phayao Hospital is not publicly available due to ethical restrictions but can be obtained upon reasonable request to the corresponding author with appropriate ethics approval.

Acknowledgments: P. Suksorn and M. Khwanmuang would like to thank the revenue budget in 2025, School of Medicine, University of Phayao (Grant No. MD68-17). W. Cholamjiak and N. Eiamniran would like to thank University of Phayao and Thailand Science Research and Innovation Fund (Fundamental Fund 2026, Grant No. 2253/2568). During the preparation of this manuscript, the authors used ChatGPT (GPT-5, OpenAI, 2025) for language refinement, grammar correction, and improvement of academic phrasing in the Abstract, Methods, and Discussion sections. The tool was used solely to enhance clarity and readability; it did not generate scientific ideas, perform data analysis, or contribute to the interpretation of results. All scientific content, clinical interpretation, and conclusions were developed, verified, and approved by the authors, who take full responsibility for the integrity and accuracy of the work. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. International Diabetes Federation. *IDF Diabetes Atlas*, 10th ed.; IDF: Brussels, Belgium, 2021.
2. World Health Organization. *Global Report on Diabetes*; WHO: Geneva, Switzerland, 2016.
3. Zheng, Y.; Ley, S.H.; Hu, F.B. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat. Rev. Endocrinol.* **2018**, *14*, 88–98. [CrossRef]
4. Saeedi, P.; Petersohn, I.; Salpea, P.; Malanda, B.; Karuranga, S.; Unwin, N.; Colagiuri, S.; Guariguata, L.; Motala, A.A.; Ogurtsova, K.; et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res. Clin. Pract.* **2019**, *157*, 107843. [CrossRef] [PubMed]
5. American Diabetes Association. Introduction: Standards of Medical Care in Diabetes—2022. *Diabetes Care* **2022**, *45*, S1–S2. [CrossRef]
6. Hu, F.B. Globalization of diabetes: The role of diet, lifestyle, and genes. *Diabetes Care* **2011**, *34*, 1249–1257. [CrossRef]
7. Lindström, J.; Tuomilehto, J. The diabetes risk score: A practical tool to predict type 2 diabetes risk. *Diabetes Care* **2003**, *26*, 725–731. [CrossRef]
8. Wilson, P.W.F.; Meigs, J.B.; Sullivan, L.; Fox, C.S.; Nathan, D.M.; D'Agostino, R.B. Prediction of incident diabetes mellitus in middle-aged adults: The Framingham Offspring Study. *Arch. Intern. Med.* **2007**, *167*, 1068–1074. [CrossRef]

9. Sukumar, N.; Dallosso, H.; Saravanan, P.; Yates, T.; Telling, C.; Shorthose, K.; Khunti, K. Baby Steps—A Structured Group Education Programme with Accompanying Mobile Web Application Designed to Promote Physical Activity in Women with a History of Gestational Diabetes: Study Protocol for a Randomised Controlled Trial. *Trials* **2018**, *19*, 682. [CrossRef]
10. Chen, L.; Magliano, D.J.; Zimmet, P.Z. The worldwide epidemiology of type 2 diabetes mellitus—Present and future perspectives. *Nat. Rev. Endocrinol.* **2012**, *8*, 228–236. [CrossRef]
11. American Diabetes Association. Diagnosis and classification of diabetes: Standards of care in diabetes—2024. *Diabetes Care* **2024**, *47*, S20–S42. [CrossRef] [PubMed]
12. Prasad, R.B.; Groop, L. Genetics of Type 2 Diabetes—Pitfalls and Possibilities. *Genes* **2015**, *6*, 87–123. [CrossRef] [PubMed]
13. Chen, X.; Xie, N.; Feng, L.; Huang, Y.; Wu, Y.; Zhu, H.; Zhang, Y. Oxidative Stress in Diabetes Mellitus and Its Complications: From Pathophysiology to Therapeutic Strategies. *Chin. Med. J.* **2025**, *138*, 15–27. [CrossRef]
14. Donath, M.Y.; Shoelson, S.E. Type 2 diabetes as an inflammatory disease. *Nat. Rev. Immunol.* **2011**, *11*, 98–107. [CrossRef]
15. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [CrossRef] [PubMed]
16. Espíndola-Fernández, D.A.; Posada-Cano, A.M.; Aristizábal-Ocampo, D.; Gallo-Villegas, J. Measurement of Biological Age with Biomarkers: A Scoping Review. *Medicina* **2024**, *46*, 178–203. [CrossRef]
17. Choi, B.G.; Rha, S.W.; Kim, S.W.; Kang, J.H.; Park, J.Y.; Noh, Y.K. Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-Up in Non-Diabetic Patients with Cardiovascular Risks. *Yonsei Med. J.* **2019**, *60*, 191–199. [CrossRef]
18. Yahyaoui, A.; Jamil, A.; Rasheed, J.; Yesiltepe, M. A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. In Proceedings of the 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 6–7 November 2019; IEEE: New York City, NY, USA; pp. 1–4.
19. Tan, K.R.; Seng, J.J.B.; Kwan, Y.H.; Chen, Y.J.; Zainudin, S.B.; Loh, D.H.F.; Low, L.L. Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review. *J. Diabetes Sci. Technol.* **2023**, *17*, 474–489. [CrossRef] [PubMed]
20. Kiran, M.; Xie, Y.; Anjum, N.; Ball, G.; Pierscionek, B.; Russell, D. Machine Learning and Artificial Intelligence in Type 2 Diabetes Prediction: A Comprehensive 33-Year Bibliometric and Literature Analysis. *Front. Digit. Health* **2025**, *7*, 1557467. [CrossRef] [PubMed]
21. Alabdulhafith, M.; Saleh, H.; Elmannai, H.; Ali, Z.H.; El-Sappagh, S.; Hu, J.W.; El-Rashidy, N. A Clinical Decision Support System for Edge/Cloud ICU Readmission Model Based on Particle Swarm Optimization, Ensemble Machine Learning, and Explainable Artificial Intelligence. *IEEE Access* **2023**, *11*, 100604–100621. [CrossRef]
22. Oikonomou, E.K.; Khera, R. Machine Learning in Precision Diabetes Care and Cardiovascular Risk Prediction. *Cardiovasc. Diabetol.* **2023**, *22*, 259. [CrossRef]
23. Xu, W.; Zhou, Y.; Jiang, Q.; Fang, Y.; Yang, Q. Risk Prediction Models for Diabetic Nephropathy among Type 2 Diabetes Patients in China: A Systematic Review and Meta-Analysis. *Front. Endocrinol.* **2024**, *15*, 1407348. [CrossRef]
24. Gogoi, P.; Valan, J.A. Interpretable Machine Learning for Chronic Kidney Disease Prediction: A SHAP and Genetic Algorithm-Based Approach. *Biomed. Mater. Devices* **2025**, *3*, 1384–1402. [CrossRef]
25. Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **2010**, *21*, 128–138. [CrossRef] [PubMed]
26. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Feature selection and classification in multiple-class datasets: An application to KDD Cup 99 dataset. *Expert Syst. Appl.* **2011**, *38*, 5947–5957. [CrossRef]
27. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
28. Mwangi, B.; Tian, T.S.; Soares, J.C. A review of feature reduction techniques in neuroimaging. *Neuroinformatics* **2014**, *12*, 229–244. [CrossRef]
29. Moons, K.G.M.; Altman, D.G.; Reitsma, J.B.; Collins, G.S. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **2015**, *162*, W1–W73. [CrossRef]
30. Wynants, L.; Van Calster, B.; Collins, G.S.; Riley, R.D.; Heinze, G.; Schuit, E.; Bonten, M.M.J.; Dahly, D.L.; Damen, J.A.A.; Debray, T.P.A.; et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **2019**, *170*, 51–58.
31. Luo, W.; Phung, D.; Tran, T.; Gupta, S.; Rana, S.; Karmakar, C.; Shilton, A.; Yearwood, J.; Dimitrova, N.; Ho, T.B.; et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Nat. Med.* **2022**, *28*, 1773–1784. [CrossRef]
32. Somvanshi, S.; Das, S.; Javed, S.A.; Antariksa, G.; Hossain, A. A survey on deep tabular learning. *arXiv* **2024**, arXiv:2410.12034. [CrossRef]

33. The Jamovi Project. *Jamovi*; Version 2.7; The Jamovi Project: Sydney, Australia, 2025. Available online: https://www.jamovi.org (accessed on 10 November 2025).
34. R Core Team. *R: A Language and Environment for Statistical Computing*; Version 4.5; R Foundation for Statistical Computing: Vienna, Austria, 2025. Available online: https://cran.r-project.org (accessed on 10 November 2025).
35. Fox, J.; Weisberg, S. *Car: Companion to Applied Regression*; Version 3.1-2; R Foundation for Statistical Computing: Vienna, Austria. Available online: https://cran.r-project.org/package=car (accessed on 10 November 2025).
36. Early Stage Diabetes Risk Prediction. *UCI Machine Learning Repository [Dataset]*; University of California, Irvine: Irvine, CA, USA, 2020. Available online: https://doi.org/10.24432/C5VG8H (accessed on 10 November 2025). [CrossRef]
37. Perveen, S.; Shahbaz, M.; Guergachi, A.; Keshavjee, K. Performance analysis of data mining classification techniques to predict diabetes. *Procedia Comput. Sci.* **2016**, *82*, 115–121. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.