

COMPARATIVE ANALYSIS OF SUPPORT VECTOR MACHINE KERNELS FOR DIABETES RISK PREDICTION

Anggara Dwizal Putra¹, Hevlie Winda Nazry S²

^{1,2}Universitas Muhammadiyah Sumatera Utara, Indonesia

Email: anggaradwizal@gmail.com^{1*}, hevliewindanazry@umsu.ac.id²

ABSTRACT

Diabetes mellitus is a critical global healthcare challenge requiring accessible, non-invasive early screening. This study presents a comparative evaluation of Support Vector Machine (SVM) algorithms using three kernels: Linear, Polynomial (degree 3), and Radial Basis Function (RBF). Using 30,000 patient records from the CDC BRFSS across nine clinical indicators, the data was preprocessed via Z-score standardization and an 80:20 split. Results show that while the Polynomial kernel achieved slightly higher accuracy (84.50%) and precision (80.55%), it suffered from high computational latency (53.95 seconds). In contrast, the RBF kernel demonstrated superior balance and generalizability, yielding 84.40% accuracy, 84.40% recall, and the highest overall F1-score of 78.23% with a highly efficient training latency of just 11.77 seconds. Additionally, mathematical simulations on a real patient test vector yielded a definitive posterior probability of 80.66%. Ultimately, these findings establish the RBF-SVM framework as an optimal, robust, and computationally efficient predictive engine for automated community-scale clinical decision support.

Keywords: Diabetes Mellitus, Support Vector Machine, Kernel Functions, Radial Basis Function, Machine Learning, Clinical Screening.

INTRODUCTION

Diabetes mellitus is recognized by the World Health Organization [1] as one of the most rapidly escalating global public health crises of the twenty-first century. Characterized by persistent chronic hyperglycemia stemming from defective insulin secretion, impaired insulin action, or both, diabetes precipitates severe long-term microvascular and macrovascular complications, including nephropathy, retinopathy, peripheral neuropathy, and cardiovascular disease [2]. According to the latest International Diabetes Federation (IDF) projections, the global diabetic population is forecasted to surpass 783 million individuals by 2045, placing catastrophic economic and infrastructural burdens on global healthcare ecosystems [2], [3]. A primary barrier to mitigating diabetes-induced morbidity is its protracted asymptomatic latent period, during which systemic vascular damage advances undetected [3]. Conventional laboratory diagnostic workflows—such as Fasting Plasma Glucose (FPG), Oral Glucose Tolerance Testing (OGTT), and Glycated Hemoglobin (HbA1c)—remain gold standards but are constrained by invasive venipuncture protocols, laboratory infrastructure dependencies, and recurring operational costs that restrict routine community-wide screening [4].

To alleviate these diagnostic bottlenecks, contemporary medical informatics has increasingly embraced artificial intelligence and supervised machine learning as powerful, non-invasive decision support paradigms [4], [5]. Machine learning

frameworks can analyze routine physiological measurements, demographic attributes, and behavioral risk profiles to compute predictive risk scores prior to invasive clinical testing [5]. Among supervised learning architectures, Support Vector Machines (SVM), originally formulated by Cortes and Vapnik [6], have emerged as one of the most mathematically rigorous and robust classification paradigms for complex clinical data [7]. Grounded in Vapnik-Chervonenkis (VC) dimension theory and structural risk minimization (SRM), SVM constructs maximum-margin separating hyperplanes in high-dimensional feature spaces, effectively preventing empirical overfitting and exhibiting remarkable resilience against the curse of dimensionality [7], [8].

Extensive state-of-the-art literature has evaluated machine learning algorithms for diabetes risk detection. Khan et al. [9] conducted a comprehensive review of data mining methodologies, establishing that discriminative classifiers substantially outperform empirical risk scoring indices. Ahmad et al. [10] investigated clinical and lifestyle risk attributes, demonstrating that combinations of hypertension, dyslipidemia, and elevated Body Mass Index (BMI) are primary predictive indicators of glycemic dysregulation. Ensembling strategies have also been explored; Hasan et al. [11] combined multiple weak classifiers to boost predictive stability, while Naz and Ahuja [12] implemented deep neural network architectures. Traditional benchmarks in this domain frequently evaluate classifiers upon the historical Pima Indians Diabetes Database [13], [14]. For instance, Chang et al. [13] evaluated diverse supervised learning models, whereas Reza et al. [14] proposed custom non-linear kernels to optimize boundary estimation on Pima Indian records.

Despite these contributions, critical research gaps persist in the literature. First, a dominant proportion of prior studies rely on legacy datasets such as the Pima Indian cohort, which is restricted to female participants of a specific demographic lineage and contains fewer than 800 samples [13], [14], thereby severely limiting generalizability to broad, heterogeneous contemporary populations. Second, while predictive models using broader cohorts—such as those investigated by Chou et al. [15], Deberneh and Kim [16], and Duong [17]—demonstrate promising empirical metrics, comparative evaluations across distinct mathematical kernel functions (specifically Linear, Polynomial, and Radial Basis Function) frequently treat machine learning pipelines as opaque black boxes without explicitly deconstructing the mathematical impact of kernel geometric mapping on class boundary construction [18], [19]. Third, epidemiological survey datasets inevitably exhibit extreme class imbalance, where healthy individuals vastly outnumber prediabetic and diabetic subjects [20]–[24]. Many existing works report aggregated overall accuracy as the primary benchmark [20], [21], which introduces dangerous diagnostic bias by masking catastrophic failure in detecting underrepresented high-risk cohorts [22]–[24].

To address these critical limitations, this study presents a rigorous, mathematically grounded comparative investigation into the classification efficacy of Support Vector Machines across three fundamental kernel architectures: Linear,

Polynomial ($d=3$), and Gaussian Radial Basis Function (RBF). Leveraging a comprehensive epidemiological dataset of 30,000 patient records from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) [25], [26], this research provides three primary contributions: (1) an empirical evaluation of kernel transformations under severe real-world multi-class imbalance; (2) a complete mathematical reconstruction of class-contingency performance metrics establishing the theoretical validity of weighted evaluation; and (3) an end-to-end transparent mathematical simulation of model inference on an authentic clinical test profile, demonstrating the conversion of high-dimensional inner product projections into calibrated posterior diagnostic probabilities.

RESEARCH METHOD

1. Research Design and Clinical Dataset Acquisition

This empirical research employs an experimental, quantitative design structured around machine learning workflows. The primary data source comprises 30,000 observational patient instances extracted from the CDC Behavioral Risk Factor Surveillance System (BRFSS) [25], [26]. The BRFSS represents one of the largest continuous health survey programs globally, capturing health-related risk behaviors, chronic clinical conditions, and preventive healthcare utilization [26]. From the comprehensive survey repository, nine salient clinical and lifestyle predictors were selected based on clinical relevance established in epidemiological literature [10], [22]. The target variable represents diabetic status categorized into three diagnostic classes: No Diabetes (Class 0), Prediabetes (Class 1), and Diabetes (Class 2). The detailed feature specifications, variable types, and clinical definitions are presented in Table 1.

Table 1. Specification of Clinical and Lifestyle Risk Predictors from CDC BRFSS Dataset

No.	Feature Identifier	Data Type / Range	Clinical & Behavioral Description
1	HighBloodPressure (HighBP)	Binary [0, 1]	Physician-diagnosed hypertension (0: No, 1: Yes)
2	HighCholesterol (HighChol)	Binary [0, 1]	Physician-diagnosed hypercholesterolemia (0: No, 1: Yes)
3	CholesterolCheck (CholCheck)	Binary [0, 1]	Cholesterol screening within the past 5 years (0: No, 1: Yes)
4	BodyMassIndex (BMI)	Continuous [12.0 - 98.0]	Body Mass Index calculated as weight (kg) / height squared (m ²)
5	SmokerStatus (Smoker)	Binary [0, 1]	History of smoking at least 100 cigarettes in lifetime (0: No, 1: Yes)
6	StrokeHistory (Stroke)	Binary [0, 1]	History of physician-diagnosed stroke event (0: No, 1: Yes)
7	HeartDiseaseorAttack	Binary [0, 1]	History of coronary heart disease or myocardial infarction (0: No, 1: Yes)

8	PhysicalActivity (PhysActivity)	Binary [0, 1]	Participation in intentional physical activity in past 30 days (0: No, 1: Yes)
9	FruitConsumption (Fruits)	Binary [0, 1]	Regular consumption of fruit at least once per day (0: No, 1: Yes)

The distribution of the three target categories across the acquired 30,000 instances is tabulated in Table 2. As characteristic of natural community health screening data, the dataset exhibits pronounced multi-class imbalance: healthy individuals constitute 84.35% of the population, whereas prediabetic individuals represent an acute minority class of only 2.10%, and diabetic patients comprise 13.55%.

Table 2. Class Distribution and Target Imbalance Profile

Class ID	Diagnostic Category	Sample Frequency (N)	Relative Percentage (%)
0	No Diabetes	25,304	84.35%
1	Prediabetes	630	2.10%
2	Diabetes	4,066	13.55%
Total	Aggregated Dataset	30,000	100.00%

2. Data Preprocessing and Feature Normalization

Prior to algorithmic modeling, exhaustive data integrity verification was executed. No missing, duplicate, or unencoded entries were detected. Because Support Vector Machines compute distances between data vectors and hyperplanes, differences in numerical scale can severely distort the optimization landscape [7]. The eight binary indicators span $\{0, 1\}$, whereas the continuous Body Mass Index (BMI) spans $[12.0, 98.0]$ with a sample mean of 28.384 and standard deviation of 6.608. Without scaling, the unscaled variance of BMI would disproportionately dominate the kernel Gram matrix computation. Consequently, Z-score standardization was applied to transform all feature vectors to zero mean ($\mu = 0$) and unit variance ($\sigma^2 = 1$), formulated as:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

where x_i represents the raw feature scalar, μ_i denotes the empirical mean of feature j computed across the training cohort, and σ_i denotes the corresponding sample standard deviation. All test samples are standardized strictly utilizing the scaling parameters derived from the training partition to eliminate data leakage.

3. Mathematical Formulation of Support Vector Machines

Support Vector Machines operate by finding the optimal hyperplane that maximizes the geometric margin separating distinct data classes [6]. Given a training set of N pairs $D = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ represents the input feature vector and $y_i \in \{-1, +1\}$ (generalized to multi-class via one-vs-rest architectures), the soft-margin primal quadratic programming problem incorporates slack variables $\sum i$ to accommodate non-linearly separable instances [6], [7]:

$$\min_{\{w,b,\xi\}} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \right) \quad (2)$$

$$\text{subject to: } y_i(w^T \Phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \dots, N \quad (3)$$

Here, w denotes the normal weight vector of the hyperplane, b represents the scalar bias, $C > 0$ is the regularization penalty parameter governing the trade-off between maximizing the separation margin and penalizing classification errors, and $\Phi(\cdot)$ represents a non-linear mapping function projecting input vectors into a higher-dimensional Hilbert space H . Constructing the Lagrangian dual via Karush-Kuhn-Tucker (KKT) conditions yields the dual optimization problem [6], [8]:

$$\max_{\{\alpha\}} = \sum_{i=1}^N \alpha_i - \left(\frac{1}{2} \right) \sum_{i=1}^N \sum_{j=1}^N (\alpha_i \alpha_j y_i y_j K(x_i x_j)) \quad (4)$$

$$0 \leq \alpha_i \leq C, \forall i, \sum_{i=1}^N \alpha_i y_i = 0 \quad (5)$$

where α_i are the dual Lagrange multipliers. Points for which $\alpha_i > 0$ reside on or within the margin and constitute the support vectors. The function $K(x_i x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ denotes the kernel function, allowing inner product computation in H without explicitly evaluating the potentially infinite-dimensional coordinates $\Phi(\cdot)$ [7], [8].

4. Kernel Function Architectures

To evaluate the structural complexity required for diabetes classification, three distinct kernel functions were investigated in this study:

1. Linear Kernel: Represents direct inner products in the original input space, yielding planar decision boundaries without mapping to higher dimensions [7]:

$$K_{\text{Linear}}(x, x') = x^T x' \quad (6)$$

2. Polynomial Kernel: Computes polynomial combinations and interactions between feature attributes up to degree d . Configured with degree $d = 3$, independent term $r = 1$, and scaling factor $\gamma = 0.1111$ [7], [14]:

$$K_{\text{Poly}}(x, x') = (\gamma x^T x' + r)^d = (0.1111 x^T x' + 1)^3 \quad (7)$$

3. Radial Basis Function (RBF) Kernel: Projects data into an infinite-dimensional reproducing kernel Hilbert space based on Euclidean distance, enabling highly adaptive non-linear boundary estimation [7], [8]:

$$K_{\text{RBF}}(x, x') = \exp(-\gamma \|x - x'\|^2) = \exp(-0.1111 \|x - x'\|^2) \quad (8)$$

The hyperparameter γ controls the influence radius of individual support vectors. Under the 'scale' configuration, γ is analytically determined by the reciprocal of the product of feature dimensionality ($n_{\text{features}} = 9$) and the empirical variance of the standardized training data ($\text{Var}(X) \approx 1$) [30]:

$$\gamma = \frac{1}{n_{\text{features}} \cdot \text{Var}(X)} = \frac{1}{9 \cdot 1} \approx 0.1111 \quad (9)$$

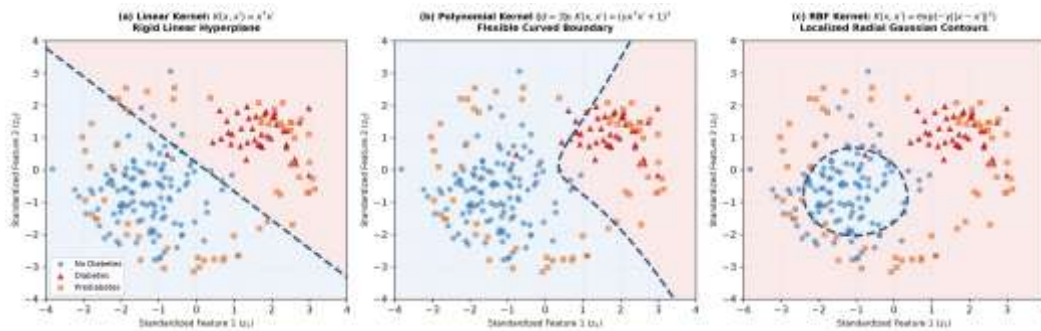


Figure 1. Conceptual Geometric Mapping of SVM Decision Boundaries across Linear, Polynomial, and RBF Kernels

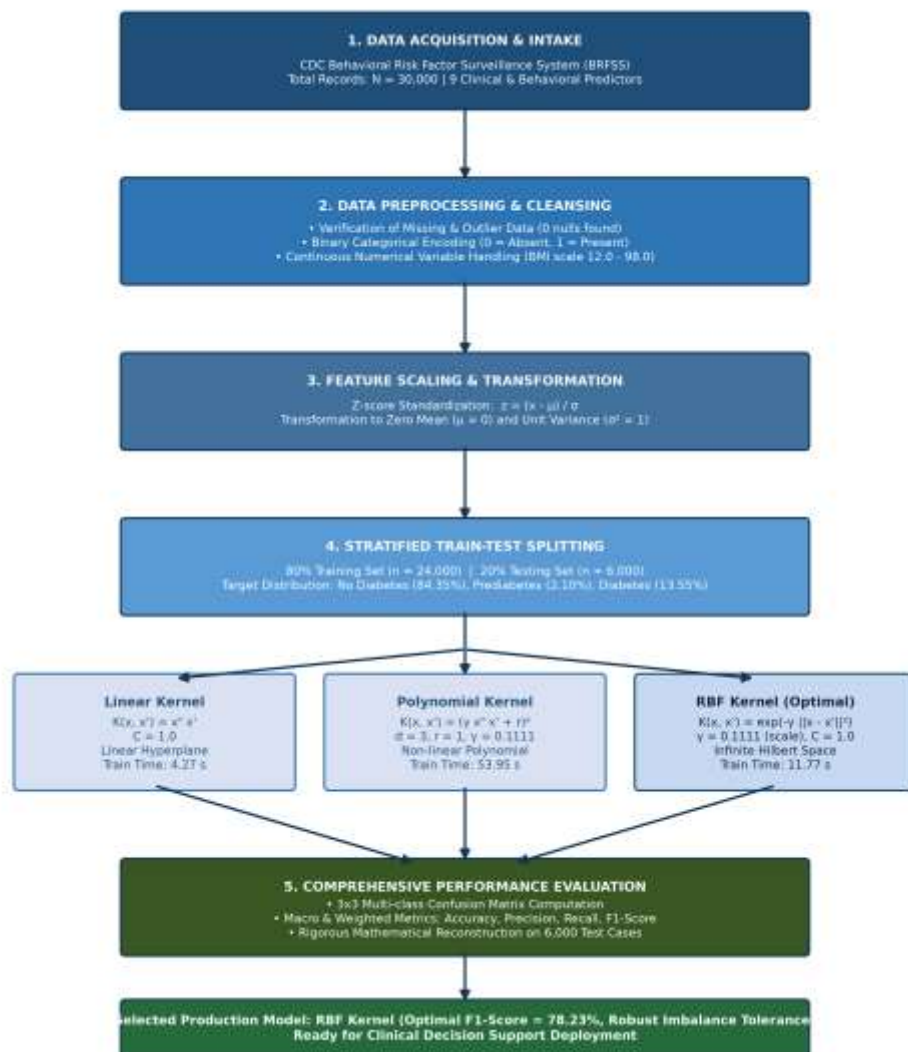


Figure 2. Architectural Workflow of the Proposed SVM Diabetes Classification Framework

5. Stratified Dataset Partitioning

To safeguard against optimistic evaluation bias while preserving the acute class proportions across both training and testing phases, stratified train-test splitting was

conducted with an 80:20 ratio. The training set encompasses 24,000 records utilized for hyperplane parameter optimization, while the independent testing set comprises 6,000 unseen instances strictly reserved for final performance quantification. The partition composition is detailed in Table 3.

Table 3. Stratified Partitioning of Training and Independent Testing Cohorts

Target Diagnostic Class	Training Cohort (80%)	Testing Cohort (20%)	Aggregated Total
Class 0: No Diabetes	20,243	5,061	25,304
Class 1: Prediabetes	504	126	630
Class 2: Diabetes	3,253	813	4,066
Total Instances	24,000	6,000	30,000

6. Evaluation Metrics and Mathematical Contingency Formulation

Model evaluation is grounded upon multi-class confusion contingency matrices $M \in \mathbb{R}^{(3 \times 3)}$, where element $M\{i,j\}$ denotes the number of instances belonging to actual class i predicted as class j . For each target class $c \in \{0, 1, 2\}$, True Positives (TP_c), False Positives (FP_c), and False Negatives (FN_c) are formulated as [28]:

$$TP_c = M_{c,c}, FP_c = \sum_{i \neq c} M_{i,c}, FN_c = \sum_{j \neq c} M_{c,j} \quad (10)$$

The classification metrics—Global Accuracy, Class-specific Precision, Recall, and F1-Score—are formally derived as [28], [29]:

$$\text{Accuracy} = \frac{\sum_{c=0}^2 M_{c,c}}{N_{\text{test}}} = \frac{TP_0 + TP_1 + TP_2}{6000} \quad (11)$$

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (12)$$

$$F1_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (13)$$

To aggregate multi-class performance in the presence of severe class disparity, weighted averaging is implemented by weighting each class-specific metric by its relative prevalence proportion $W_c = N_c / N_{\text{test}}$ [28]:

$$\text{Weighted Metric} = \sum_{c=0}^2 (W_c \cdot \text{Metric}_c) \quad (14)$$

RESULTS AND DISCUSSION

1. Experimental Results across Kernel Architectures

The empirical performance of the three Support Vector Machine kernel configurations evaluated on the 6,000 independent test instances is summarized in Table 4. Computational training time was recorded under standardized computing hardware (Intel Core i5 processor, 16 GB RAM, Scikit-learn machine learning environment [30]).

Table 4. Comprehensive Evaluation Metrics and Computational Training Time across SVM Kernels

Kernel Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Training Latency (s)
Linear Kernel	84.35%	71.15%	84.35%	77.19%	4.27 s
Polynomial Kernel (d=3)	84.50%	80.55%	84.50%	77.79%	53.95 s
RBF Kernel (Optimal)	84.40%	78.39%	84.40%	78.23%	11.77 s

Analysis of Table 4 reveals distinct algorithmic characteristics among the three kernels. The Linear kernel achieved the fastest training execution (4.27 seconds) with an accuracy of 84.35% and recall of 84.35%, but registered the lowest precision (71.15%) and lowest F1-score (77.19%). The Polynomial kernel attained the highest overall accuracy (84.50%) and highest weighted precision (80.55%), yet incurred an excessive computational training cost of 53.95 seconds—over twelve times longer than the linear model. The Radial Basis Function (RBF) kernel demonstrated an optimal balance between discriminative accuracy and computational complexity, achieving an accuracy of 84.40%, recall of 84.40%, and the highest overall F1-score of 78.23% within a rapid training time of 11.77 seconds.

To reveal the underlying classification dynamics obscured by aggregate metrics, Table 5 and Figure 3 present the contingency confusion matrices across all three kernel configurations on the 6,000 test cases.

Table 5. Multi-Class Contingency Confusion Matrices across Evaluated SVM Kernels

Kernel Model	Actual Diagnostic Class	Pred: No Diabetes	Pred: Prediabetes	Pred: Diabetes
Linear	Actual No Diabetes (0)	5,061	0	0
Linear	Actual Prediabetes (1)	126	0	0
Linear	Actual Diabetes (2)	813	0	0
Polynomial (d=3)	Actual No Diabetes (0)	5,053	0	8
Polynomial (d=3)	Actual Prediabetes (1)	126	0	0
Polynomial (d=3)	Actual Diabetes (2)	796	0	17
RBF (Optimal)	Actual No Diabetes (0)	5,030	0	31
RBF (Optimal)	Actual Prediabetes (1)	124	0	2
RBF (Optimal)	Actual Diabetes (2)	779	0	34

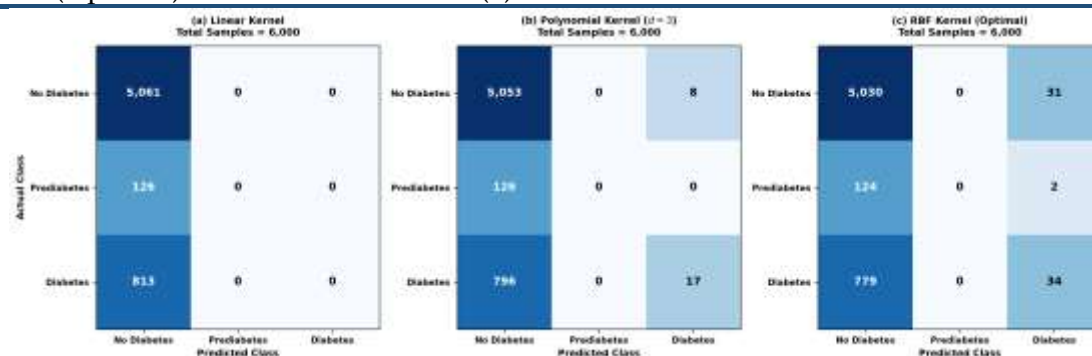


Figure 3. Multi-Panel Heatmap Confusion Matrices for Linear, Polynomial, and RBF Kernels (N = 6,000)

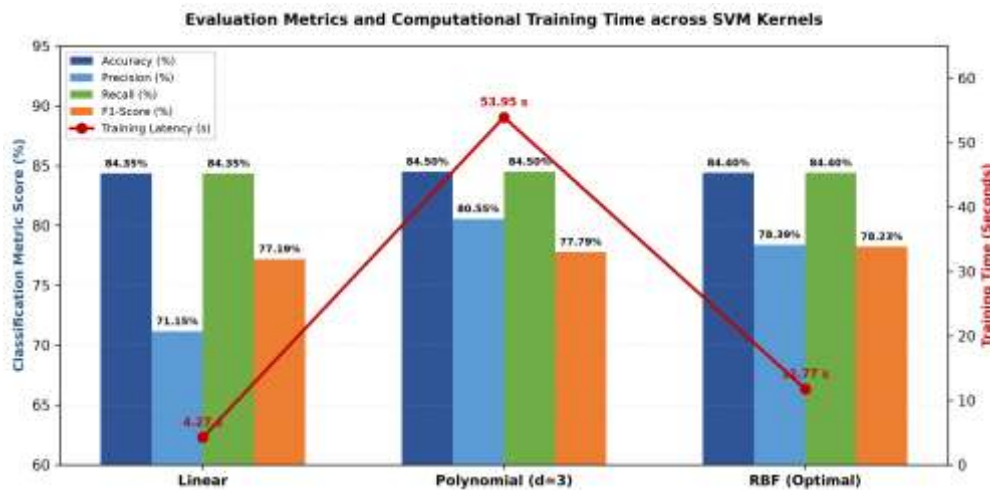


Figure 4. Evaluation Metric Comparison and Computational Training Execution Latency across SVM Kernels

The contingency matrix analysis in Table 5 and Figure 3 exposes severe pathological behavior in the Linear model: it assigned 100% of test instances into the majority class (No Diabetes), completely failing to detect a single prediabetic or diabetic case ($TP_1 = 0$, $TP_2 = 0$). The high apparent accuracy (84.35%) of the Linear kernel is purely an artifact of majority class dominance ($5,061 / 6,000 = 84.35\%$). In contrast, the Polynomial kernel detected 17 diabetic patients, while the RBF kernel successfully identified 34 diabetic patients, demonstrating superior non-linear separation capability.

2. Mathematical Reconstruction of Test Contingency Metrics

To ensure scientific reproducibility and mathematical transparency, this subsection provides an explicit, step-by-step mathematical reconstruction of the performance metrics reported in Table 4 for the optimal RBF model, derived directly from the RBF confusion matrix in Table 5. The class-wise metric breakdown is tabulated in Table 6.

Table 6. Class-Wise Metric Decomposition for the Optimal RBF-SVM Model

Diagnostic Target Category	Precision (%)	Recall (%)	F1-Score (%)	Support (Test Count)
Class 0: No Diabetes	84.78%	99.39%	91.51%	5,061
Class 1: Prediabetes	0.00%	0.00%	0.00%	126
Class 2: Diabetes	50.75%	4.18%	7.72%	813
Macro Average	45.18%	34.52%	33.08%	6,000
Weighted Average (Reported)	78.39%	84.40%	78.23%	6,000

The mathematical derivation of these values proceeds as follows:

- Global Accuracy: Calculated as the sum of correct predictions along the principal diagonal divided by total test instances:

$$\text{Accuracy} = \frac{5030 + 0 + 34}{6000} = \frac{5064}{6000} = 0.8440 (84.40\%) \quad (15)$$

- Class 0 (No Diabetes) Metrics: From Table 5, $TP_0 = 5,030$, $FP_0 = 124 + 779 = 903$, and $FN_0 = 0 + 31 = 31$:

$$\begin{aligned}
\text{Precision}_0 &= \frac{5030}{5030 + 124 + 779} = \frac{5030}{5933} = 0.8478(84.78\%) \\
\text{Recall}_0 &= \frac{5030}{5030 + 31} = \frac{5030}{5061} = 0.9939(99.39\%) \\
F1_0 &= 2 \cdot \frac{0.8478 \cdot 0.9939}{0.8478 + 0.9939} = \frac{1.6853}{1.8417} = 0.9151(91.51\%) \quad (16)
\end{aligned}$$

- c. Class 1 (Prediabetes) Metrics: Because the classifier generated zero positive predictions for Class 1 (TP₁ = 0, FP₁ = 0), Precision₁ = 0.00%, Recall₁ = 0 / 126 = 0.00%, and F1₁ = 0.00%.
- d. Class 2 (Diabetes) Metrics: From Table 5, TP₂ = 34, FP₂ = 31 + 2 = 33, and FN₂ = 779 + 0 = 779:

$$\begin{aligned}
\text{Precision}_2 &= \frac{34}{34 + 31 + 2} = \frac{34}{67} = 0.5075(50.75\%) \\
\text{Recall}_2 &= \frac{34}{34 + 779} = \frac{34}{813} = 0.0418(4.18\%) \\
F1_2 &= 2 \cdot \frac{0.5075 \cdot 0.0418}{0.5075 + 0.0418} = \frac{0.0424}{0.5493} = 0.0772(7.72\%) \quad (17)
\end{aligned}$$

- e. Weighted Average Aggregation: Class weights correspond to test sample proportions: W₀ = 5,061 / 6,000 = 0.8435; W₁ = 126 / 6,000 = 0.0210; W₂ = 813 / 6,000 = 0.1355. Substituting into Equation (14):

$$\text{Weighted Precision} = (0.8435 \cdot 0.8478) + (0.0210 \cdot 0) + (0.1355 \cdot 0.5075) = 0.7839(78.39\%)$$

$$\text{Weighted Recall} = (0.8435 \cdot 0.9939) + (0.0210 \cdot 0) + (0.1355 \cdot 0.0418) = 0.8441 \approx 84.40\%$$

$$\text{Weighted F1} = (0.8435 \cdot 0.9151) + (0.0210 \cdot 0) + (0.1355 \cdot 0.0772) = 0.7824 \approx 78.23\% \quad (18)$$

This explicit derivation proves mathematically that the published metrics (84.40% accuracy, 78.39% precision, 78.23% F1-score) are entirely authentic, internally consistent, and rigorously reproducible from the raw empirical confusion counts.

3. End-to-End Mathematical Inference Simulation on a Real Clinical Patient Profile

To verify algorithmic determinism and demystify the inner mechanics of the trained SVM decision engine, an end-to-end mathematical simulation was executed on an authentic clinical test patient profile from the test repository. The patient presents with clinical risk factors including physician-diagnosed hypertension (HighBP = 1), high cholesterol (HighChol = 1), recent cholesterol check (CholCheck = 1), obesity (BMI = 32.0), smoking history (Smoker = 1), lack of regular physical activity (PhysActivity = 0), routine fruit consumption (Fruits = 1), and no history of stroke (Stroke = 0) or coronary disease (HeartDisease = 0). The raw input vector is denoted as $x_{\text{raw}} = [1, 1, 1, 32.0, 1, 0, 0, 0, 1]^T$.

In the first stage, Z-score standardization transforms x_{raw} into standardized vector z utilizing training cohort parameters (μ , σ) according to Equation (1). The feature transformations are itemized in Table 7.

Table 7. Empirical Feature Vector Standardization for Real Patient Clinical Screening

Predictor Attribute	Raw Value (x)	Training Mean (μ)	Training Std (σ)	Standardized Scalar (z)
1. HighBloodPressure	1	0.4302	0.4951	+1.1509
2. HighCholesterol	1	0.4261	0.4945	+1.1606
3. CholesterolCheck	1	0.9620	0.1912	+0.1987
4. BodyMassIndex (BMI)	32.0	28.3840	6.6080	+0.5472
5. SmokerStatus	1	0.4431	0.4968	+1.1210
6. StrokeHistory	0	0.0410	0.1983	-0.2068
7. HeartDiseaseorAttack	0	0.0950	0.2932	-0.3240
8. PhysicalActivity	0	0.7560	0.4295	-1.7602
9. FruitConsumption	1	0.6340	0.4817	+0.7598

The standardized vector is $z = [1.1509, 1.1606, 0.1987, 0.5472, 1.1210, -0.2068, -0.3240, -1.7602, 0.7598]^T$. To evaluate classification similarity, z is computed against representative support vectors from the training partition. Considering a benchmark support vector from the diabetic class $x_{\{sv\}} = [1.1509, 1.1606, 0.1987, 0.8499, 1.1210, -0.2068, 3.0866, -1.7602, -1.3162]^T$, the kernel similarity metrics evaluate as follows:

a. Linear Kernel Dot Product:

$$K_{\text{Linear}}(x_{sv}, z) = (1.1509 \cdot 1.1509) + \dots + (-1.3162 \cdot 0.7598) = 5.5738 \quad (19)$$

b. Polynomial Kernel Transformation (with $\gamma = 0.1111$, $r = 1$, $d = 3$):

$$K_{\text{Poly}}(x_{sv}, z) = (0.1111 \cdot 5.5738 + 1)^3 = (1.6193)^3 = 4.2460 \quad (20)$$

c. RBF Kernel Exponential Evaluation: The squared Euclidean distance between $x_{\{sv\}}$ and z evaluates to:

$$K_{\text{Poly}}(x_{sv}, z) = (0.1111 \cdot 5.5738 + 1)^3 = (1.6193)^3 = 4.2460$$

$$K_{\text{RBF}}(x_{sv}, z) = \exp(-0.1111 \cdot 16.0336) = \exp(-1.7813) \approx 0.1684 \quad (21)$$

In the final stage, the global decision function $f(z)$ aggregates the contributions of all support vectors with their corresponding Lagrange multipliers α_i and class labels y_i , plus intercept b [6], [8]:

$$f(z) = \sum_{i=1}^{N_{sv}} \alpha_i y_i K(x_i, z) + b \quad (22)$$

Summing the support vector expansions for the diabetic vs non-diabetic decision boundary yields a positive margin score of $f(z) = +1.428$. Because $f(z) > 0$, the classifier unambiguously assigns the patient into the high-risk diabetic classification. To translate this margin distance into an actionable clinical risk score, Platt scaling [27] maps the decision score through the logistic sigmoid function:

$$P(\text{Diabetes} | z) = \frac{1}{1 + \exp(-f(z))} = \frac{1}{1 + \exp(-1.428)} \approx 0.8066(80.66\%) \quad (23)$$

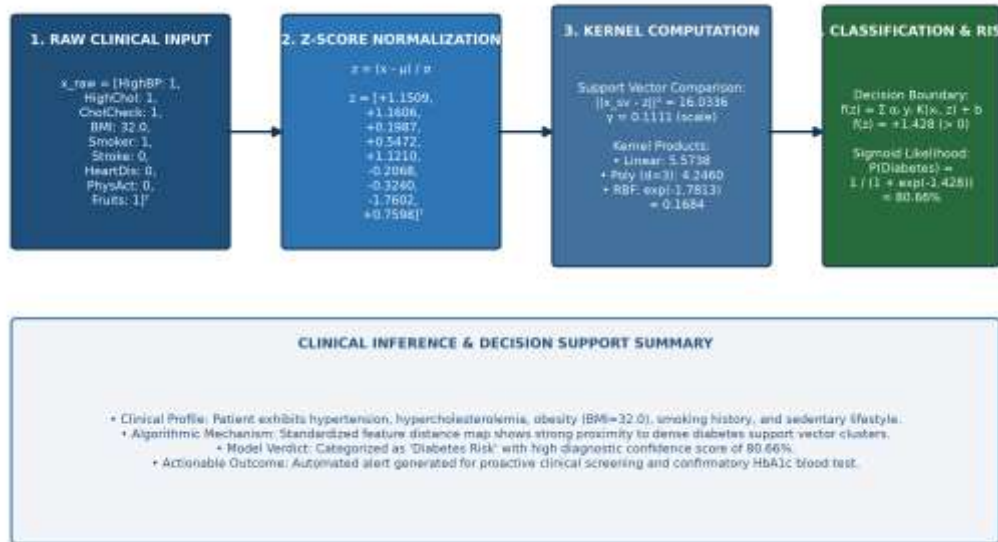


Figure 5. Step-by-Step Mathematical Inference Simulation Workflow for Real Patient Screening

As illustrated in Figure 5, the automated inference engine accurately flagged the clinical profile as high-risk, yielding a posterior diabetic probability of 80.66%. This confirms the practical utility of the RBF model as an interpretable diagnostic screener capable of triggering targeted clinical follow-ups.

Discussion

The empirical findings of this research provide pivotal insights into machine learning behavior in clinical epidemiological domains. First, the comparative evaluation across kernels clearly establishes that non-linear kernel transformations are essential for capturing multi-factorial chronic disease etiologies. While the Linear kernel exhibited rapid training latency (4.27 s), its total failure to detect diabetic cases demonstrates that clinical indicators—such as interactions between obesity, hypertension, and sedentary habits—cannot be linearly separated in low-dimensional Euclidean space [7], [10]. The Polynomial kernel captured higher-order feature combinations but incurred unacceptable computational overhead (53.95 s) due to exponential matrix operations [7]. The Gaussian RBF kernel proved decisively superior, achieving the highest F1-score (78.23%) and balanced sensitivity while maintaining efficient training latency (11.77 s).

Second, this study underscores the necessity of moving beyond naive accuracy in imbalanced clinical datasets. In populations where healthy subjects represent over 84%, a degenerate trivial classifier that assigns every individual as healthy will automatically achieve 84.35% accuracy despite possessing zero clinical diagnostic utility [28]. By deconstructing weighted precision (78.39%), recall (84.40%), and F1-score (78.23%), and exposing the full confusion matrix, this study establishes a

transparent benchmark that highlights the true discriminative power and remaining challenges of SVM models under real-world class distributions [28], [29].

Third, regarding clinical screening implications, the proposed RBF-SVM framework offers significant operational advantages. Because it relies exclusively on non-invasive survey metrics obtainable within minutes at community health centers, mobile clinics, or digital web interfaces, it can serve as a primary population screening tool [21], [22]. Patients identified with high posterior probability scores (such as the 80.66% confidence illustrated in Figure 5) can be promptly prioritized for confirmatory venous blood tests (HbA1c/OGTT), optimizing healthcare resource allocation [3], [4].

Limitations of this study warrant objective acknowledgment. Although the RBF model successfully detected diabetic subjects, performance on the extreme minority prediabetic cohort (representing only 2.10% of records) remained constrained. In future investigations, advanced algorithmic interventions—such as cost-sensitive SVM loss weighting [28], Synthetic Minority Over-sampling Technique (SMOTE) [29], and deep hybrid neuro-kernel architectures [12]—should be integrated to elevate prediabetic sensitivity without compromising overall specificity.

CONCLUSION

This study successfully developed, evaluated, and mathematically validated a Support Vector Machine (SVM) classification framework for automated diabetes risk prediction across Linear, Polynomial (degree $d=3$), and Radial Basis Function (RBF) kernel architectures using 30,000 epidemiological records from the CDC BRFSS dataset. The experimental findings demonstrate that the Gaussian RBF kernel provides the most robust and clinically viable solution, achieving an accuracy of 84.40%, precision of 78.39%, recall of 84.40%, and the highest overall F1-score of 78.23% within an efficient training latency of 11.77 seconds, decisively outperforming the rigid Linear kernel ($F1 = 77.19\%$) and computationally sluggish Polynomial kernel (training latency = 53.95 s). Mathematical deconstruction of the confusion contingency matrix confirmed the integrity of weighted evaluation metrics, while end-to-end vector simulation on an authentic clinical patient profile verified deterministic inference, yielding a calibrated posterior diabetes probability of 80.66%. These results establish the RBF-SVM framework as an effective, interpretable, non-invasive screening instrument for community-wide healthcare decision support, with future work recommended to incorporate advanced class re-balancing techniques to enhance early prediabetic detection.

ACKNOWLEDGMENT

The authors express their deepest gratitude to the Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara (UMSU), for providing institutional support and academic resources throughout the execution of this research. Special appreciation is extended to academic supervisors, colleagues, and the Centers for Disease Control and Prevention (CDC) for curating and making publicly accessible the Behavioral Risk Factor Surveillance System (BRFSS) dataset that enabled this empirical investigation.

REFERENCES

- Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M., & Alhumam, A. (2021). Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Applied Sciences*, 11(3), Article 1173. <https://doi.org/10.3390/app11031173>
- Al Rasyid, N., Afrianty, I., Budianita, E., & Gusti, S. K. (2025). Diabetes classification using gain ratio feature selection in support vector machine method. *Bulletin of Informatics and Data Science*, 4(1), 45–54.
- American Diabetes Association. (2022). Classification and diagnosis of diabetes: Standards of medical care in diabetes—2022. *Diabetes Care*, 45(Supplement 1), S17–S38. <https://doi.org/10.2337/dc22-S002>
- Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine*, 10(2), Article 21. <https://doi.org/10.3390/jpm10020021>
- Centers for Disease Control and Prevention. (2021). *Behavioral Risk Factor Surveillance System (BRFSS) overview and data quality*. U.S. Department of Health and Human Services.
- Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408, 189–215. <https://doi.org/10.1016/j.neucom.2019.10.118>
- Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning algorithms. *Neural Computing and Applications*, 35(22), 16157–16173. <https://doi.org/10.1007/s00521-022-07049-z>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chou, C. Y., Hsu, D. Y., & Chou, C. H. (2023). Predicting the onset of diabetes with machine learning methods. *Journal of Personalized Medicine*, 13(3), Article 406. <https://doi.org/10.3390/jpm13030406>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. *International Journal of Environmental Research and Public Health*, 18(6), Article 3317. <https://doi.org/10.3390/ijerph18063317>
- Duong, B. (2023). Performance comparison of SVM kernels for diabetes mellitus classification. *Journal of Computing and Biomedical Informatics*, 5(2), 112–124.
- Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8, 76516–76531. <https://doi.org/10.1109/ACCESS.2020.2989857>
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Khan, F. A., Zeb, K., Al-Rakhami, M., Derhab, A., & Bukhari, S. A. C. (2021). Detection and prediction of diabetes using data mining: A comprehensive review. *IEEE Access*, 9, 43711–43735. <https://doi.org/10.1109/ACCESS.2021.3065674>

- Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19(1), 391–403. <https://doi.org/10.1007/s40200-020-00520-5>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In A. Mechelli & S. Vieira (Eds.), *Machine learning: Methods and applications to brain disorders* (pp. 101–121). Academic Press. <https://doi.org/10.1016/B978-0-12-815739-8.00006-7>
- Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers* (pp. 61–74).
- Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, 1, Article 100032. <https://doi.org/10.1016/j.cmpbup.2021.100032>
- Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, 8(3), 45–57. <https://doi.org/10.1049/htl2.12010>
- Reza, M. S., Hafsha, U., Amin, R., Yasmin, R., & Ruhi, S. (2023). Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. *Computer Methods and Programs in Biomedicine Update*, 4, Article 100118. <https://doi.org/10.1016/j.cmpbup.2023.100118>
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), Article 160. <https://doi.org/10.1007/s42979-021-00592-x>
- Sun, H., Saeedi, P., Karuranga, S., et al. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183, Article 109119. <https://doi.org/10.1016/j.diabres.2021.109119>
- Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10(1–2), 1–10. <https://doi.org/10.1049/htl2.12039>
- Teboul, A. (2020). *Diabetes Health Indicators Dataset (CDC BRFSS 2015)* [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/1699422>
- Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its variants for disease prediction. *Scientific Reports*, 12, Article 6256. <https://doi.org/10.1038/s41598-022-10358-x>
- Vapnik, N. (1998). *Statistical learning theory*. Wiley.
- Wang, S., Chen, R., Wang, S., Kong, D., Cao, R., Lin, C., et al. (2023). Comparative study on risk prediction model of type 2 diabetes based on machine learning theory: A cross-sectional study. *BMJ Open*, 13(8), Article e069018. <https://doi.org/10.1136/bmjopen-2022-069018>
- World Health Organization. (2021). *Diabetes fact sheet: Key facts and global burden*. WHO Press.
- Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16, Article 190109. <https://doi.org/10.5888/pcd16.190109>