BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2665-2680
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2665-2680

GENE SELECTION FOR TYPE 2 DIABETES MELLITUS (T2DM) DISEASE USING MULTIPLE SUPPORT VECTOR MACHINE – RECURSIVE FEATURE ELIMINATION (MSVM-RFE) ALGORITHM

Andi Khalil Gibran Basir[1*], **Ahmad Husain**[2], **A. Fuad Ahsan Basir**[3]

[1]*Department of Product, Narasio Data*
Denver Apartment Unit 879, Made, Kec. Sambikerep, Surabaya, 60129, Indonesia

[2]*Department of Data Science, Institut Teknologi Bacharuddin Jusuf Habibie*
Jln. Balaikota No.1, Bumi Harapan, Kec. Bacukiki Bar, Parepare, 91122, Indonesia

[3]*Department of Engineering, BitHealth*
Jln. BSD Green Office Park, Green Office Park 9 3rd Floor, Sampora, Tangerang, 15345, Indonesia

*Corresponding author's e-mail: * **khalil@narasiodata.com***

Article Info

Article History:

Received: 18th October 2025
Revised: 20th January 2026
Accepted: 17th March 2026
Published: 8th April 2026

Keywords:

Type 2 diabetes mellitus;
Gene expression;
Multiple Support Vector Machine-Recursive Feature Elimination.

ABSTRACT

Gene selection is essential for improving classification performance and interpretability in high-dimensional microarray data. This study applies a Multiple Support Vector Machine–Recursive Feature Elimination (MSVM-RFE) framework for gene selection in Type 2 Diabetes Mellitus (T2DM). Experiments were conducted on a GEO microarray dataset comprising 118 samples (73 controls and 45 T2DM cases) with 25,770 genes. MSVM-RFE employs multiple linear SVM models within a 10-fold cross-validation scheme as feature selection to enhance accuracy and was evaluated under different train–test splits, with and without SMOTE resampling. The selected gene subsets were classified using SVM with linear, RBF, and polynomial kernels. The best configuration achieved 95.67% accuracy, with high sensitivity, specificity, and AUROC, using fewer than 100 genes. These results demonstrate that MSVM-RFE provides a robust and effective gene selection strategy for T2DM microarray analysis.

How to cite this article:

A. K. G. Basir, A. Husain and A. F. A. Basir., "GENE SELECTION FOR TYPE 2 DIABETES MELLITUS (T2DM) DISEASE USING MULTIPLE SUPPORT VECTOR MACHINE – RECURSIVE FEATURE ELIMINATION (MSVM-RFE) ALGORITHM", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2665-2680, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

The development of DNA microarray technology has enabled a wide range of applications in the medical field, including genetic research, disease screening, risk prediction, and the development of precision therapeutics tailored to individual genetic profiles. DNA microarrays provide high-throughput capability, high sensitivity and specificity, and are supported by well-established data analysis tools, thereby facilitating rapid advances in genomics research [1]. In genomic studies, microarray platforms such as cDNA microarrays, SNP microarrays, and oligonucleotide expression microarrays allow the simultaneous measurement of expression levels for thousands to millions of genes in a single experiment [1].

Despite these advantages, microarray-based gene expression data are inherently high-dimensional while typically comprising only a limited number of samples. This unfavorable dimensionality-to-sample-size ratio imposes significant analytical challenges, including increased computational burden, reduced statistical power, and heightened susceptibility to the curse of dimensionality and overfitting [2], [3],[4]. Without appropriate preprocessing, these issues can substantially degrade classification performance and limit the interpretability of downstream analyses. Consequently, effective feature reduction strategies are essential for reliable gene-expression–based disease modeling.

Gene selection is a critical preprocessing step that identifies a subset of informative genes strongly associated with specific phenotypic classes. In the context of disease classification, gene selection enhances predictive accuracy while improving model interpretability by highlighting potential diagnostic or prognostic biomarkers. Broadly, gene selection methods can be categorized into filter, embedded, and wrapper approaches. Filter methods rank genes based on statistical relevance independent of the classifier; embedded methods incorporate feature selection into the model training process; and wrapper methods evaluate gene subsets by repeatedly training and testing a predictive model to optimize a performance criterion, such as classification accuracy [2], [3].

Among wrapper-based methods, Support Vector Machine–Recursive Feature Elimination (SVM-RFE) has been widely adopted for gene selection in high-dimensional biomedical datasets. SVM-RFE iteratively removes genes with the smallest contribution to the SVM decision function, producing ranked gene subsets optimized for classification. Previous studies have demonstrated that SVM-RFE can substantially improve classification performance compared with SVM models without feature selection. For example, Rustam [5], reported superior performance of SVM combined with SVM-RFE in lung cancer microarray classification. Extensions and variants of SVM-RFE have also been explored in disease-related gene identification. Zhang [6], for instance, applied SVM-RFE in combination with the Boruta algorithm to type 2 diabetes mellitus (T2DM) datasets, identifying several candidate and hub genes relevant to disease mechanisms.

Despite its effectiveness, conventional SVM-RFE remains sensitive to sampling variability and is prone to overfitting, particularly in small-sample, high-dimensional settings. These limitations are further compounded in imbalanced datasets, which are common in biomedical research. Although several studies have proposed stabilization strategies or resampling techniques to mitigate these issues, only a limited number of works have systematically investigated the joint integration of multiple-model feature selection concepts and resampling methods to enhance accuracy and address class imbalance in gene-expression analysis, particularly for T2DM [7], [8], [9].

To address these challenges, this study applies a multiple-concept framework within the established SVM-RFE methodology, referred to as Multiple Support Vector Machine–Recursive Feature Elimination (MSVM-RFE). Rather than proposing a new algorithm, the novelty of this work lies in the systematic application of the MSVM-RFE framework combined with the Synthetic Minority Over-Sampling Technique (SMOTE) to enhance accuracy, mitigate overfitting, and address class imbalance in T2DM gene-expression data. In addition, this study performs a comprehensive comparative evaluation of different SVM kernels and train–test splitting schemes using gene subsets selected through the MSVM-RFE framework.

The main contributions of this work are summarized as follows:

1. This study applies an MSVM-RFE framework for gene selection to enhance accuracy in high-dimensional T2DM microarray data.
2. It integrates SMOTE resampling within the training process to address class imbalance and reduce overfitting effects.

3. It systematically compares multiple SVM kernels and data-splitting schemes to evaluate classification performance using MSVM-RFE–selected gene subsets.
4. It identifies biologically relevant candidate genes, including hub genes such as *GNAS*, that are potentially involved in T2DM pathophysiology.

Based on the above considerations, the following section describes the datasets, preprocessing procedures, and the MSVM-RFE–based analytical framework employed to perform gene selection and SVM classification for T2DM microarray data.

2. RESEARCH METHODS

2.1 Support Vector Machine-Recursive Feature Elimination (SVM-RFE)

Recursive Feature Elimination (RFE) is a prominent feature selection technique widely used in biomedical data analysis. It is conceptually designed to address the challenge of overfitting, a phenomenon that typically occurs when the number of features (e.g., genes) is substantially larger than the number of samples (e.g., patients). Therefore, RFE aims to discern the minimal subset of features that yields the highest classification efficacy [10].

In essence, RFE operates through an iterative, systematic procedure that ranks each feature, then removes the feature with the lowest criterion value (a method known as backward feature elimination). The operational workflow of RFE can be delineated as follows [11]:

1. Initialization: A classification model, for instance, a Support Vector Machine (SVM), is trained utilizing the complete set of available features.
2. Ranking Criterion Calculation: The importance of each feature to the model is quantified by computing a ranking criterion. In the context of a linear SVM, a frequently used criterion is the square of the feature's weight, denoted as w_i^2. A smaller w_i^2 value indicates diminished feature importance.
3. Weakest Feature Elimination: The feature (or features) possessing the lowest criterion value is excised from the feature set.
4. Iteration: Steps 1-3 are repeated with the reduced feature subset until all features have been eliminated. This process culminates in a comprehensive ranking of features, ordered from least to most significant.

RFE is typically employed in conjunction with a specific classification model. The most prevalent implementation observed in biomedical research is the integration of RFE with a Support Vector Machine (SVM), an approach subsequently termed SVM-RFE. [12]

SVM-RFE is a feature selection algorithm that generates a rank-ordered list of features according to their importance. Feature selection is then performed by selecting a specified number of top-ranked features. The scoring criterion in SVM-RFE is intrinsically linked to the SVM classifier, which is renowned for its high accuracy and robust generalization capabilities. [13]

In linear SVM-RFE, the selection process begins by training a linear SVM to compute the separating hyperplane's weight vector w. The square of each component wk2 serves as the feature ranking criterion. The feature associated with the smallest value is considered the least influential and is subsequently removed. This process is recursively iterated until all features have been eliminated. The resulting elimination order determines the final feature ranking [13].

In the linear form, the model assumes that the data can be separated by the following decision function:

$$f(x) = \mathbf{w}^{\boldsymbol{T}}\mathbf{x} + b, \tag{1}$$

where $\mathbf{w} = (w_1, w_2, \dots, w_d)^{\boldsymbol{T}}$ presents the weight vector that defines the orientation of the hyperplane, and b denotes the bias term that determines the hyperplane's position relative to the origin.

Each component w_k represents the relative contribution of the k^{th} feature. A higher absolute value of w_k indicates a greater significance of that feature in defining the decision boundary between classes. Consequently, Linear SVM-RFE employs the squared weight value w_k^2 as a quantitative measure of feature importance.

$$J(k) = w_k^2. \tag{2}$$

Using Eq. (2), the steps of the Linear SVM-RFE procedure can be described as follows:

1. Train a linear SVM model.
2. Compute the weight w_k for each feature.
3. The value $J(k) = w_k^2$ represents the importance of the k^{th} feature.
4. Remove the feature with the smallest $J(k)$ value.
5. Repeat the process recursively until all features are ranked according to their relative influence.

The linear SVM model is obtained by minimizing the following objective function:

$$\min_{\mathbf{w},\boldsymbol{b},\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i, \tag{3}$$

subject to the constraints:

$$y_i(\mathbf{w}^T x_i + b) \geq 1 - \xi i, \xi i \geq 0, i = 1, \ldots, n. \tag{4}$$

Here, $y_i \in \{+1, -1\}$ denotes the class label for the i^{th} sample, ξ_i represents the slack variable that allows margin violations, and C is the regularization parameter controlling the trade-off between maximizing the margin and minimizing classification errors.

The main advantage of this method lies in its simplicity and its ability to utilize the entire training dataset without relying on cross-validation accuracy. However, when the number of features is very large, removing one feature at a time becomes computationally expensive. Consequently, it is standard practice to remove a subset of features simultaneously in each iteration.

2.2 Multiple Support Vector Machine-Recursive Feature Elimination (MSVM-RFE)

Reducing the dimensionality of a dataset by eliminating redundant or irrelevant features can significantly improve classification accuracy. This process ensures that only the most informative features are retained, minimizing noise and improving the model's generalization capability. Among various feature selection methods, the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) has been widely recognized as an efficient wrapper approach, particularly in gene selection problems where datasets are often high-dimensional. By systematically removing less significant features, SVM-RFE identifies a subset of genes that contribute most effectively to class separation, thereby optimizing the performance of classification models [14].

Building on the foundation of SVM-RFE, the Multiple Support Vector Machine-Recursive Feature Elimination (MSVM-RFE) was developed as an enhanced, more robust variant. This advanced version demonstrates improved reliability in identifying informative genes when compared to the conventional SVM-RFE approach. The improvement stems primarily from the integration of a bootstrap-based stabilization mechanism that operates at every recursive elimination stage. This mechanism mitigates instability caused by sample variability and ensures more consistent feature selection across different data subsets, thereby enhancing the robustness of the overall feature selection process [15].

The distinctive contribution of MSVM-RFE lies in introducing the "multiple" concept: multiple SVM models are trained iteratively during the recursive elimination process. At each iteration, the method applies a backward elimination strategy, progressively discarding genes associated with the smallest weights. This iterative procedure allows the algorithm to assess feature importance more comprehensively by evaluating multiple models trained on varying subsets of the data rather than relying on a single model. As a result, the selected features are more representative and less sensitive to random fluctuations within the dataset.

In practice, as illustrated in Fig. 1, MSVM-RFE first trains several linear SVM models on different subsamples of the training dataset. These subsamples are typically generated using bootstrap resampling or alternative methods such as 10-fold cross-validation (CV). By training on these varied subsets, the algorithm captures different perspectives of the data distribution, enhancing the reliability of the resulting feature rankings. This step ensures that each model contributes to a more stable, generalizable estimate of feature importance, particularly valuable in biological datasets characterized by small sample sizes and high dimensionality.

Following the training stages, gene ranking scores are computed based on the squared weights $(w_i)^2$ obtained from the SVM-RFE computations. These scores reflect the relative importance of each gene in distinguishing between classes. Once all features have been evaluated and ranked, the top-ranked genes are extracted to form the final subset used in subsequent classification analyses. This refined subset not only improves classification accuracy but also enhances model interpretability by focusing on biologically meaningful genes. Consequently, MSVM-RFE serves as a powerful and stable approach for feature selection in complex, high-dimensional datasets, especially in the domain of bioinformatics and genomics [16].

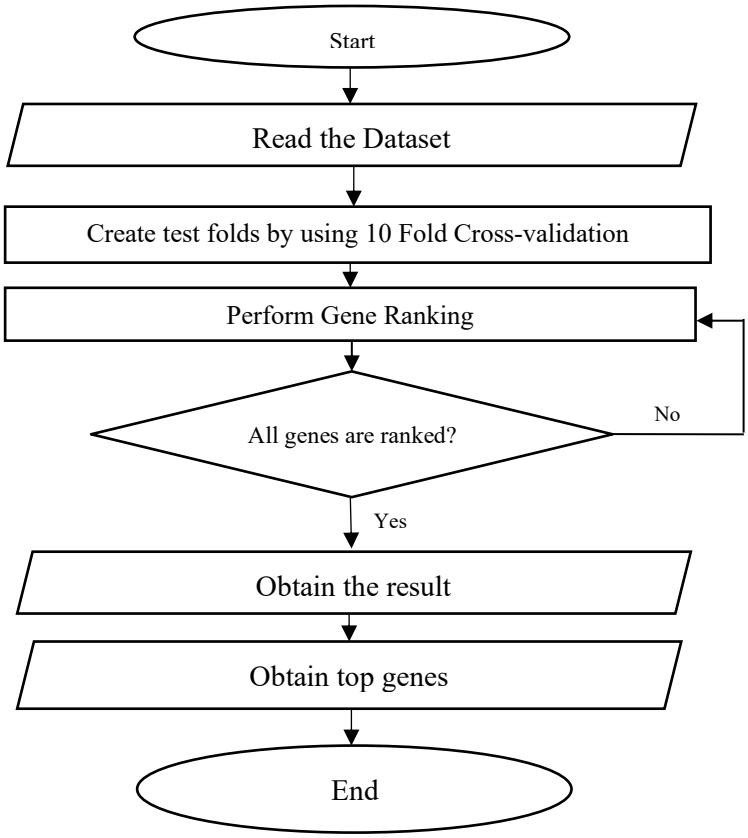

Figure 1. The Flow Chart of MSVM-RFE

Consider that t linear SVM models are trained on different subsamples of the original training dataset. Denote w_j as the weight vector of the j-th linear SVM, and w_{ji} as the component of this vector corresponding to the i-th feature. Define $v_{ji} = (w_{ji})^2$ as the squared weight for that feature. The ranking score for each feature can then be evaluated using the following criterion:

$$c_i = \frac{\bar{v}_l}{\sigma_{v_i}}, \tag{5}$$

$$\bar{v}_l = \frac{1}{t}\sum_{j=1}^{t} v_{ji}, \tag{6}$$

$$\bar{\sigma}_{v_i} = \sqrt{\frac{\sum_{j=1}^{t}\left(v_{ji} - \bar{v}_l\right)^2}{t-1}}. \tag{7}$$

Here, $\bar{v}_l$ represents the mean, and $\bar{\sigma}_{v_i}$ denotes the standard deviation of $\bar{v}_l$. It is crucial to normalize the weight vectors prior to calculating the ranking score for each gene.

$$w_i = \frac{w_j}{\|w_j\|}. \tag{8}$$

Fig. 2 below illustrates the recursive process of the MSVM-RFE algorithm. First, the MSVM-RFE procedure begins by initializing the complete set of genes $R = [\ ]$. From an initial selected subset $S = [1, \dots, d]$, the following steps are iteratively performed until all genes are ranked. Multiple linear SVMs are trained on subsamples of the original training data using the genes in subset S as input variables. Next, the weight vectors are computed and normalized. Using Eq. (5), the ranking score c is calculated for each gene in S. The gene with the lowest ranking score is then removed from the subset S. Finally, the ranked gene list in R is updated accordingly [15].

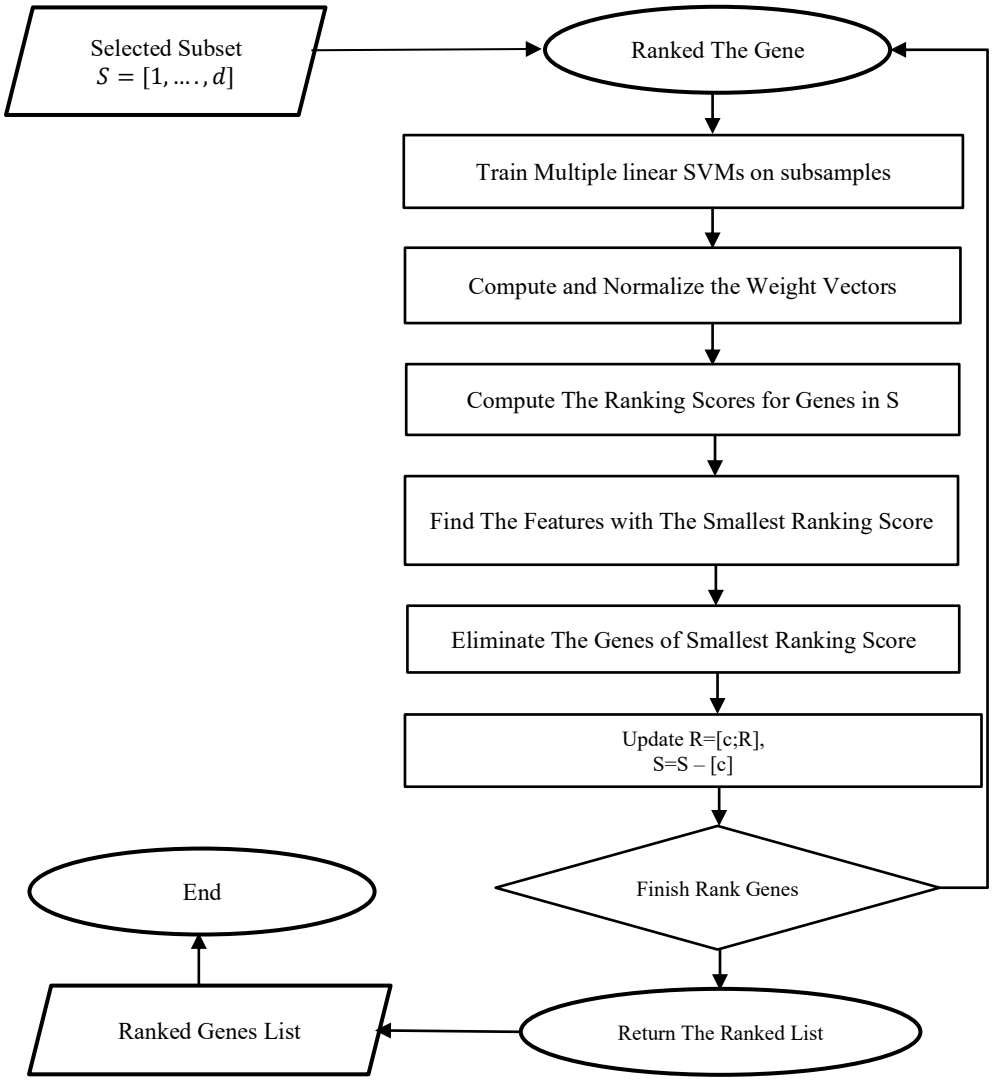

Figure 2. Recursive Procedure in MSVM-RFE

2.3 Gene Selection for Type 2 Diabetes Mellitus (T2DM) Microarray Datasets Using MSVM-RFE

Type 2 Diabetes Mellitus (T2DM) is a metabolic disorder characterized by chronic hyperglycemia and deficient β-cell function, resulting from dysregulated insulin secretion and/or insulin resistance in peripheral tissues [17], [18]. It is primarily triggered by some pathogenic factors, i.e., obesity, lifestyle, and genetic elements [17].

In recent years, with the rapid development of gene expression data analysis, earlier studies have shown that preparing gene expression data is a critical step in biological function analyses before moving on to classification. However, given the small number of samples relative to the large number of genes in T2DM

gene expression data, identifying the smallest possible subset of the most informative genes in T2DM remains challenging to avoid spurious results. Several gene selection methods for microarray data have been proposed in earlier studies, and the Multiple Support Vector Machine-Recursive Feature Elimination (MSVM-RFE) method has been shown to be one of the most stable for gene expression data.

The workflow of this research starts with acquiring gene expression phenotype data from the Gene Expression Omnibus (GEO) Dataset GSE18732, which can be downloaded from the NCBI GEO Datasets Database (http://ncbi.nlm.nih.gov). In total, the gene expression phenotype data in this dataset contains 118 observations and 25770 genes. The class consists of 73 patients with normal conditions and 45 patients with T2DM.

After the dataset is acquired, the next step is to preprocess it, such as normalizing and $\log_2$-transforming it, to align it with the standard input dataset. This dataset should be formatted as a dataframe, with rows representing observations and columns representing genes or features. The first column should contain the true class label, whereas in the common SVM classifier, the data is separated into classes and features. Afterward, the dataset will be transposed and sorted.

The MSVM-RFE procedure starts by constructing data partitions using a 10-fold cross-validation scheme applied to the training dataset. This process determines the allocation of samples into ten mutually exclusive folds, where each fold is alternately treated as the validation set while the remaining folds form the corresponding training subset. The resulting fold assignments are organized into an index-based list that specifies the test samples for each fold.

Based on this configuration, gene ranking is conducted independently within each of the ten training subsets. In this framework, the number of subsampled models is fixed at $k = 10$, which constitutes the "multiple" component of MSVM-RFE and distinguishes it from the conventional SVM-RFE approach, which employs only a single model ($k = 1$). Consequently, at each recursive elimination step, ten linear SVM models are trained, each using approximately 90% of the available training samples.

Feature elimination is performed iteratively using a halving strategy, whereby the number of retained genes is reduced by 50% at each iteration until the remaining feature set contains fewer than 100 genes. For each fold, this process produces an ordered list of genes ranked according to their importance scores derived from the SVM weight vectors.

To ensure reproducibility, gene rankings from the 10 cross-validation folds were aggregated by averaging their ranks. Genes with lower mean ranks were regarded as more informative and retained as higher-priority features. During the MSVM-RFE process, a linear SVM kernel was consistently used to preserve the interpretability of feature weights, while the regularization parameter C was held constant and optimized later during the classification phase via cross-validation. The MSVM-RFE algorithm was run to completion to generate a full ranking of all genes, without predefining a fixed number of selected features. The final gene subset was then selected by systematically evaluating multiple top-ranked subset sizes (e.g., 10–100 genes) and choosing the subset that yielded the highest cross-validated classification performance in the SVM stage.

At the SVM stage, the dataset is split into training and test sets, with approximately one-third of the samples reserved for testing. The test set is strictly excluded from all training and preprocessing steps to prevent data leakage and ensure unbiased performance evaluation.

All preprocessing procedures for classification, including normalization and $\log_2$ transformation, are conducted exclusively on the training data, and the corresponding parameters are subsequently applied to the test data. To mitigate class imbalance, SMOTE is applied only to the training set [19]. Table 1 and Table 2. summarize the data partitioning and class distributions used in this study.

Table 1. Splitting Method Using 70 30 Rules and 80 20 Rules, with SMOTE and Without SMOTE

Splitting Method	Resampling	Train Data	Test Data
Dataset 1 (70 30 Rules)	Without SMOTE	84	34
	With SMOTE	132	34
Dataset 2 (80 20 Rules)	Without SMOTE	95	23
	With SMOTE	148	23

Table 2. Train Data After SMOTE Resampling Applied

Splitting Method	Resampling	Class	
		Control	Patient with T2DM
Dataset 1 (70 30 Rules)	Without SMOTE	51	33
	With SMOTE	66	66
Dataset 2 (80 20 Rules)	Without SMOTE	58	37
	With SMOTE	74	74

After the data in the training subset is well-balanced, the next step is to produce a reasonable winning parameter by repeating the bootstrap step. The optimal parameters obtained during the bootstrap phase are subsequently used to train a new classifier on the entire training dataset, which is then evaluated on the test dataset. Since the true values of the target attribute in the test set are already known, the model's predictions can be directly compared with these ground truths, enabling straightforward assessment of prediction correctness and facilitating more accurate accuracy estimation.

In this research, the SVM classifier was first run without SMOTE resampling. After that, SVM was run with SMOTE resampling to address the dataset imbalance [19]. The top-ranked features obtained from the MSVM-RFE procedure, with varying subset sizes (e.g., 10, 20, 30, up to 100 genes), were subsequently used as input for the classifier package to train and evaluate the SVM model. The best subset of the gene is repeated the classification process 1000 times to obtain an average accuracy.

3. RESULTS AND DISCUSSION

3.1 HUGO Gene Nomenclature Committee (HGNC) Symbol

Through the MSVM-RFE–based gene selection procedure, the resulting output consists of vectors encompassing FeatureName, FeatureID, and an ordered representation of average ranking scores. Here, FeatureName and FeatureID serve as gene identifiers, which can subsequently be mapped to standardized symbols established by the HUGO Gene Nomenclature Committee (HGNC). The AvgRank variable captures the aggregated mean ranking score across ten cross-validation folds, with lower values signifying greater predictive significance in the selected features.

Table 3. The Output of Gene Selection Step Using MSVM-RFE

Feature Name	Feature ID	AvgRank
ENST00000371095_at	19118	46.8
ENST00000371095_at	19112	49.0
ENST00000371095_at	19109	51.7
ENST00000371095_at	13332	56.6
ENST00000371095_at	3906	58.3
⋮	⋮	⋮
ENST00000371095_at	16538	23930.7

The HUGO Gene Nomenclature Committee (HGNC) assigns a standardized and unique symbol to each known human gene, thereby ensuring unambiguous reference across studies and publications. These standardized gene symbols not only provide clarity in scientific communication but also facilitate efficient electronic retrieval of information from databases and research literature [20], [21].

In the Type 2 Diabetes Mellitus (T2DM) gene expression phenotype dataset, a total of 25,770 genes were ranked according to their average ranking scores. To associate each ranked gene with its standardized identifier, the corresponding *FeatureName* was mapped to the HGNC symbol. As presented in Table 4, the five top-ranked features were extracted based on their average ranking scores and subsequently converted from *FeatureName* into their respective HGNC symbols.

Table 4. HGNC Symbol of Sorted Genes in T2DM Microarray Dataset

Gene	Feature ID	AvgRank	hgnc_symbol
ENST00000371095_at	19118	46.8	GNAS
ENST00000371095_at	19112	49.0	GNAS
ENST00000371095_at	19109	51.7	GNAS
ENST00000371095_at	13332	56.6	GNAS
ENST00000371095_at	3906	58.3	ASPA
⋮	⋮	⋮	⋮
ENST00000371095_at	16538	23930.7	PCDH11X

As shown in Table 4, GNAS appeared as the top 4 features following by ASPA in the fifth position. The GNAS gene is an important regulator of insulin secretory capacity in pancreatic β-cells. A study conducted by (Taneera et.al,2019) [17] revealed that the GNAS gene is one of seven genes that capable of showing a higher expression of pancreatic β-cells compared to the expression of α-cells and exocrine cells. The GNAS gene encodes alpha-subunit heterotrimeric Gs protein (Gs α) which has a role as a molecular witch in conveying GPCR signals to control pancreatic β-cells growth, survival, and hormonal regulation. Pancreatic β-cells play a role in insulin synthesis and secretion. Insulin secretion dysfunction is a type of disorder that is considered the etiology of T2DM. GNAS has been proposed as a pivotal gene influencing insulin secretory capacity and represents a compelling candidate for further investigation into the pathogenesis of type 2 diabetes (T2D) [17].

3.2 Optimal Hyperparameter

A good classification result requires optimal parameters so that hyperparameter optimization or adjustment is needed. Hyperparameter is a parameter whose value is used to control the learning process. In addition to improving the quality of the resulting classification, parameter tuning is also needed to estimate the generalization errors on folds so that the optimal feature can be determined by looking at the fold which has the lowest error rate on the average error of multiple folds. This step started by initiating 10 folds using k-folds cross-validation. Each in the fold contains three lists, namely the sorted feature identity number (featureID), the train data fold and the test data fold which has been divided based on the 10-fold cross-validation rule. The samples used in this process are 132 samples (for the first dataset using 70 30 rule) and 148 samples (for the second dataset using 80 20 rule).

Each fold will be applied to the top 5 feature lists obtained from the MSVM-RFE gene selection so that as many as 50 folds are formed in one process to determine the best hyperparameter combination and generalization error. Each fold in this list will look for the best combination of hyperparameters and calculate the number of errors generated when using these hyperparameters. Because this process is carried out using three kernels (linear, RBF, and polynomial) [22], it means that 150 folds are formed at the end of this process. Table 5 shows the best parameter for each Kernels before moving onto classification step.

Table 5. Optimal Hyperparameter for Classification Step

Kernel	Dataset Scheme	Parameter	Value	Smallest GE
Linear	Dataset 1	Cost	8	0.0548590
	Dataset 2	Cost	1	0.0380889
RBF	Dataset 1	Cost	64	0.0147600
		Gamma	0.001953125	
	Dataset 2	Cost	32	0.0149270
		Gamma	0.015625	
Polynomial	Dataset 1	Cost	32	0.1571884
		Gamma	0.0004882812	
		Coef0	0.125	
	Dataset 2	Cost	16	0.1572367
		Gamma	0.015625	
		Coef0	0.125	

The Generalization Error (GE) analysis revealed that the SVM model with an RBF kernel achieved the lowest GE among all configurations. This is a promising outcome, as GE provides a quantitative measure for identifying overfitting. A higher GE suggests a wider gap between the model's performance on training versus test data, indicating a stronger tendency for the model to overfit [23].

3.3 Performance Measurement

The performance of the classifier is evaluated using several metrics derived from the confusion matrix. These metrics help determine the quality of the classifier by assessing its sensitivity, specificity, and overall accuracy.

Sensitivity, also referred to as the true positive rate, reflects the model's ability to correctly identify individuals who are actually classified as 'diseased'. The higher the sensitivity of a test, the more positive test results on people who have a disease and the lower number of false negatives. The sensitivity can be calculated as

$$Sensitivity = \frac{TP}{TP + FN} \times 100\%. \tag{9}$$

Specificity is the ability of a test to correctly classify an individual as *disease-free*. The higher the specificity of a test, the more negative results in people with *disease-free* or the fewer number of false positives. The sensitivity can be calculated as

$$Specificity = \frac{TN}{FP + TN} \times 100\%. \tag{10}$$

Sensitivity and specificity typically exhibit an inverse relationship, such that an increase in sensitivity is often accompanied by a decrease in specificity, and vice versa.

Accuracy is the proportion of true test results (true value) among all the experiment that has been checked and can be calculated as

$$Accuracy = \frac{TP + TN}{Total} \times 100\%. \tag{11}$$

3.3.1 Performance Measurement for SVM Classification Without MSVM-RFE

In this study, the classification performance measurement using Support Vector Machine (SVM) on diabetes mellitus gene expression data was conducted as an initial step (baseline) prior to the application of the MSVM-RFE feature selection method. The dataset employed exhibits high-dimensional characteristics, containing 25,770 features, which poses significant computational challenges for comprehensive processing. Consequently, an initial feature selection was performed by choosing the top 100 genes exhibiting the highest variance. This approach is grounded in the assumption that genes with the highest variance are more likely to carry biologically and statistically significant information, thereby serving as a preliminary proxy to reduce dimensionality without discarding critical signals. Although this method is straightforward, in the context of microarray data, which is often dominated by genes with constant expression levels (low-variance genes), the variance-based approach is a widely accepted baseline and has proven to be an effective initial dimensionality reduction step in several studies [24], [25]. Additionally, within computational environments constrained by limited resources, this approach facilitates model exploration without the necessity of handling extremely large data matrices.

Evaluation results across two data partitioning scenarios, Dataset 1 (70:30) and Dataset 2 (80:20), demonstrated considerable variability in performance depending on kernel type, the use of SMOTE, and the number of features utilized. Overall, no configuration indicated that SVM performance is highly sensitive to data representation or kernel characteristics.

Table 6. The SVM Classification Results on Dataset 1 Using the Top 100 Genes with Highest Variance

Kernel	Metrics	Number of Features									
		10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Linear + SMOTE	Accuracy	41.17%	41.17%	58.82%	44.11%	52.94%	50.00%	61.76%	58.82%	52.94%	52.94%
	Sensitivity	38.40%	46.15%	61.53%	38.46%	61.53%	69.23%	53.84%	61.53%	61.53%	53.84%
	Specificity	42.85%	38.09%	57.14%	47.61%	47.61%	38.09%	66.67%	57.14%	47.61%	52.38%
	AUROC	67.80%	63.40%	55.30%	54.60%	48.70%	64.80%	48.40%	43.20%	54.90%	57.10%

Kernel	Metrics	Number of Features									
		10	20	30	40	50	60	70	80	90	100
Linear	Accuracy	47.05%	32.35%	44.11%	47.05%	44.11%	64.70%	55.88%	50.00%	52.94%	58.82%
	Sensitivity	0.00%	15.38%	23.07%	30.76%	7.69%	46.15%	30.76%	30.76%	38.46%	53.84%
	Specificity	76.19%	42.85%	57.14%	57.14%	66.67%	76.19%	71.42%	61.90%	61.90%	61.90%
	AUROC	72.50%	67.40%	57.50%	57.90%	51.30%	62.30%	42.90%	44.70%	50.90%	54.20%
RBF + SMOTE	Accuracy	41.17%	50.00%	38.23%	38.23%	41.17%	58.82%	50.00%	50.00%	50.00%	50.00%
	Sensitivity	23.07%	61.53%	38.46%	38.46%	46.15%	69.23%	30.76%	38.40%	46.15%	53.84%
	Specificity	52.38%	42.85%	38.09%	38.09%	38.09%	52.38%	61.90%	57.14%	52.38%	47.61%
	AUROC	70.70%	57.10%	63.00%	68.50%	62.30%	62.30%	49.80%	61.50%	46.90%	49.10%
RBF	Accuracy	52.94%	47.05%	47.05%	41.17%	41.17%	64.70%	58.82%	41.17%	47.05%	50.00%
	Sensitivity	0.00%	46.15%	23.07%	23.07%	23.07%	46.15%	23.07%	15.38%	23.07%	30.76%
	Specificity	85.71%	47.61%	61.90%	52.38%	52.38%	76.19%	80.95%	57.14%	61.90%	61.90%
	AUROC	81.30%	61.50%	57.10%	63.70%	64.50%	61.20%	53.80%	65.20%	54.90%	53.10%
Polynomial + SMOTE	Accuracy	50.00%	55.88%	52.94%	47.05%	50.00%	55.88%	55.88%	50.00%	52.94%	50.00%
	Sensitivity	38.46%	46.15%	38.46%	46.15%	53.84%	69.23%	46.15%	53.84%	61.53%	53.84%
	Specificity	57.14%	61.90%	61.90%	47.61%	47.61%	47.61%	61.90%	47.61%	47.61%	47.61%
	AUROC	59.00%	57.90%	51.60%	50.90%	50.20%	39.90%	57.50%	52.40%	55.70%	49.50%
Polynomial	Accuracy	61.76%	61.76%	61.76%	61.76%	61.76%	61.76%	61.76%	61.76%	61.76%	61.76%
	Sensitivity	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.00%
	Specificity	100%	100%	100%	100%	100%	100%	100%	100%	100%	95.23%
	AUROC	79.90%	53.10%	48.70%	52.70%	56.00%	64.80%	57.50%	50.20%	57.50%	57.10%

Table 7. The SVM Classification Results on Dataset 2 Using the Top 100 Genes with Highest Variance

Kernel	Metrics	Number of Features									
		10	20	30	40	50	60	70	80	90	100
Linear + SMOTE	Accuracy	43.47%	43.47%	43.47%	47.82%	60.86%	52.17%	65.21%	47.82%	47.82%	56.52%
	Sensitivity	44.45%	33.34%	44.45%	44.45%	66.67%	66.67%	55.56%	55.56%	55.56%	55.56%
	Specificity	42.85%	50.00%	42.85%	50.00%	57.14%	42.85%	71.42%	42.85%	42.85%	57.14%
	AUROC	64.30%	58.70%	57.90%	49.20%	61.10%	58.70%	62.70%	54.80%	51.60%	54.00%
Linear	Accuracy	39.13%	52.17%	43.47%	65.21%	69.56%	73.91%	56.52%	65.21%	47.82%	47.82%
	Sensitivity	11.12%	33.34%	11.12%	44.45%	55.56%	66.67%	44.45%	44.45%	22.23%	44.45%
	Specificity	57.14%	64.28%	64.28%	78.57%	78.57%	78.57%	64.28%	78.57%	64.28%	50.00%
	AUROC	36.50%	61.10%	53.20%	61.90%	63.50%	69.80%	67.50%	50.80%	64.30%	58.70%
RBF + SMOTE	Accuracy	26.08%	34.78%	52.17%	56.52%	47.82%	47.82%	43.47%	52.17%	47.82%	47.82%
	Sensitivity	33.34%	44.45%	66.67%	77.78%	77.78%	77.78%	66.67%	55.54%	33.34%	44.45%
	Specificity	21.42%	28.57%	42.85%	42.85%	28.57%	28.57%	28.57%	50.00%	57.14%	50.00%
	AUROC	78.60%	65.10%	54.00%	61.10%	61.90%	63.50%	54.00%	42.90%	53.20%	50.80%
RBF	Accuracy	34.78%	30.43%	52.17%	60.86%	60.86%	65.12%	56.52%	52.17%	47.82%	47.82%
	Sensitivity	11.12%	22.23%	33.34%	55.56%	55.56%	66.67%	33.34%	11.12%	22.23%	11.12%
	Specificity	50.00%	35.71%	64.28%	64.28%	64.28%	64.28%	71.42%	78.57%	64.28%	71.42%
	AUROC	78.60%	77.00%	54.00%	65.90%	63.50%	62.70%	48.40%	53.20%	57.90%	56.30%
Polynomial + SMOTE	Accuracy	30.43%	43.47%	52.17%	52.17%	52.17%	56.52%	52.17%	47.82%	43.47%	47.82%
	Sensitivity	33.34%	55.54%	77.74%	77.74%	88.89%	88.89%	88.89%	66.67%	77.74%	77.74%
	Specificity	28.57%	35.71%	35.71%	35.71%	28.57%	35.71%	28.57%	35.71%	21.42%	28.57%
	AUROC	75.40%	59.50%	63.50%	61.90%	60.30%	66.70%	42.90%	56.30%	61.90%	54.00%
Polynomial	Accuracy	52.17%	43.47%	52.17%	52.17%	43.47%	52.17%	69.56%	65.21%	60.86%	52.17%
	Sensitivity	11.12%	22.23%	33.34%	55.56%	44.45%	55.56%	66.67%	55.56%	55.56%	44.45%
	Specificity	78.57%	57.14%	64.28%	50.00%	42.85%	50.00%	71.42%	71.42%	64.28%	57.14%
	AUROC	72.22%	70.60%	55.60%	65.90%	48.40%	60.30%	66.70%	58.70%	59.50%	41.30%

Table 6 presents the results for Dataset 1, where the Linear kernel combined with SMOTE demonstrated relatively stable performance, achieving a peak accuracy of 61.76% with 70 features. However, the low AUROC (48.40%) under the same configuration indicates the model's limited ability to discriminate between classes. This finding aligns with prior studies [26], [27] , which suggest that the linear kernel tends to perform well on nearly linearly separable data but lacks flexibility for more complex patterns.

In contrast, the RBF kernel exhibited higher sensitivity in certain configurations, such as at 60 features with SMOTE (Sensitivity: 69.23%), but this was accompanied by low specificity (52.38%). This pattern implies that RBF tends to be more aggressive in predicting the positive class, which could be advantageous for early detection but carries a risk of increased false positives. A similar trend was observed in Dataset 2, shown in Table 7, where RBF with SMOTE at 40 features attained a sensitivity of 77.78%, but specificity

dropped to 42.85%. These results corroborate findings from [28], [29], [30], highlighting the trade-off between sensitivity and specificity inherent to nonlinear kernels like RBF.

The Polynomial kernel without SMOTE on Dataset 1 showed a striking pattern: accuracy remained consistently at 61.76% across nearly all feature counts, while sensitivity was very low (even 0% in some cases), and specificity approached 100%. This indicates a tendency of the model to classify all samples as the negative class, a symptom of poor handling of class imbalance. Introducing SMOTE with the Polynomial kernel significantly improved sensitivity, albeit at the cost of reduced accuracy, reinforcing the conclusions from [31] that resampling is critical in imbalanced biomedical datasets.

These classification performance measurements also emphasize the absence of a consistent pattern in which increasing the number of features invariably improves performance. For instance, in Dataset 1 with the Linear kernel and SMOTE, the highest accuracy was achieved at 70 features (61.76%), rather than at 100 features (52.94%). This phenomenon is recognized as the curse of dimensionality, where adding features beyond an optimal point introduces noise and degrades model performance. Such results underscore that although variance-based feature selection is justifiable given computational constraints and high dimensionality, this method alone is insufficient for capturing complex patterns in gene expression data.

Overall, SVM performance is heavily influenced by kernel choice, class imbalance handling, and the number of features utilized. Consequently, this study focuses on applying various kernels, employing SMOTE for resampling, and implementing MSVM-RFE as a more sophisticated feature selection method to enhance model accuracy and identify biologically informative gene subsets.

3.3.2 Performance Measurement for SVM Classification with MSVM-RFE

In total, 120 classifiers were evaluated. As summarized in Table 8, the best-performing gene subsets for each dataset configuration were first identified based on accuracy, sensitivity, specificity, and AUROC.

Table 8. The Best Performance of SVM Classification using The Best Subset Genes for Each Type of Datasets before Getting Into *K-Fold Cross-Validation* Step

Kernel	Dataset Scheme	Resampling	Number of Features	Accuracy	Specificity	Sensitivity	AUROC
Linear	Dataset 1	SMOTE	60	91.18%	92.31%	90.48%	87%
			70	91.18%	92.31%	90.48%	87%
		Without SMOTE	60	97.06%	100%	95.24%	95.24%
			70	97.06%	100%	95.24%	95.24%
	Dataset 2	SMOTE	40	91.30%	100%	85.71%	85.71%
			60	91.30%	88.89%	92.86%	87.70%
			70	91.30%	88.89%	92.86%	87.70%
			80	91.30%	100%	85.71%	85.71%
		Without SMOTE	80	95.65%	88.89%	100%	94.44%
			90	95.65%	88.89%	100%	94.44%
			100	95.65%	88.89%	100%	94.44%
RBF	Dataset 1	SMOTE	60	94.12%	100%	90.48%	90.48%
			70	94.12%	100%	90.48%	90.48%
		Without SMOTE	40	97.06%	92.31%	100%	96.15%
			60	97.06%	100%	95.24%	95.24%
			70	97.06%	100%	95.24%	95.24%
			80	97.06%	100%	95.24%	95.24%
	Dataset 2	SMOTE	80	95.65%	100%	92.86%	92.86%
			90	95.65%	100%	92.86%	92.86%
			100	95.65%	100%	92.86%	92.86%
		Without SMOTE	70	91.30%	88.89%	92.86%	87.70%
			90	91.30%	88.89%	92.86%	87.70%
			100	91.30%	88.89%	92.86%	87.70%
Polynomial	Dataset 1	SMOTE	40	97.06%	92.31%	100%	96.15%
			70	97.06%	100%	95.24%	95.24%
		Without SMOTE	80	97.06%	100%	95.24%	95.24%
	Dataset 2	SMOTE	90	95.65%	100%	92.86%	92.86%
			100	95.65%	100%	92.86%	92.86%
		Without SMOTE	90	86.96%	77.78%	92.86%	82.54%

For each shortlisted configuration, the classification process was repeated 1,000 times to obtain a stable estimate of average accuracy. Based on the results in Table 9, the final model selection followed a criterion-driven evaluation procedure.

Table 9. Average Accuracy After Repeating the Classification Process for 1000 Times

Kernel	Dataset Scheme	Resampling	Number of Features	Average Accuracy
Linear	Dataset 1	SMOTE	60	91.11%
			70	90.97%
		No SMOTE	60	95.86%
			70	95.51%
	Dataset 2	SMOTE	40	87.95%
			60	87.78%
			70	90.30%
			80	89.47%
		No SMOTE	80	93.15%
			90	93.76%
			100	95.52%
RBF	Dataset 1	SMOTE	60	93.45%
			70	91.69%
		No SMOTE	40	93.52%
			60	96.76%
			70	95.24%
			80	95.26%
	Dataset 2	SMOTE	80	93.84%
			90	95.08%
			100	**95.67%**
		No SMOTE	70	91.84%
			90	90.04%
			100	92.70%
Polynomial	Dataset 1	SMOTE	40	**95.80%**
			70	94.94%
		No SMOTE	80	**97.06%**
	Dataset 2	SMOTE	90	93.87%
			100	94.43%
		No SMOTE	90	86.54%

Based on the results summarized in Table 9, the final model selection followed a criterion-driven evaluation procedure. Candidate configurations were first screened based on the average classification accuracy from 1,000 repeated runs. Three configurations consistently achieved average accuracies above 95%, namely the RBF kernel with SMOTE, the Polynomial kernel with SMOTE, and the Polynomial kernel without SMOTE.

To reach the final decision, a comparative analysis will be conducted among these three gene-subset configurations that yielded the highest accuracies. This comparative analysis evaluates the classification performance of each kernel and resampling method based on average accuracy, specificity, sensitivity, AUROC, and Generalization Error.

Table 10. Comparison of Classification Performance Between Kernels for Selected Gene Subsets

Kernel	Resampling	Number of Features	Average Accuracy	Specificity	Sensitivity	AUROC	Generalization Error
RBF	With SMOTE	100	95.67%	100%	92.86%	92.86%	0.0149270
Polynomial	With SMOTE	40	95.80%	92.31%	100%	96.15%	0.1571884
	Without SMOTE	80	97.06%	100%	95.24%	95.24%	0.1571884

The results presented in Table 10 underscore the critical influence of kernel selection and class imbalance mitigation strategies on the performance of classification models for Type 2 Diabetes Mellitus

(T2DM). Among the evaluated configurations, the Support Vector Machine (SVM) employing a Radial Basis Function (RBF) kernel combined with Synthetic Minority Over-sampling Technique (SMOTE) demonstrated the most balanced performance. This model achieved an exceptionally low Generalization Error (GE) of 0.0149, high sensitivity, and perfect specificity (100%), indicating robust generalization to unseen data and minimal risk of overfitting. The integration of SMOTE was instrumental in addressing class imbalance, a prevalent issue in medical datasets, thereby reducing potential bias in classification outcomes.

In contrast, the model employing a Polynomial kernel combined with SMOTE on the top 40 features achieved a slightly higher accuracy and attained the highest AUC score of 96.15%. This elevated AUC reflects strong discriminative capability between T2DM and non-T2DM individuals. The model also achieved 100% sensitivity, successfully identifying all T2DM cases. However, its specificity was lower (92.31%), and the generalization error was comparatively higher, indicating a greater susceptibility to overfitting. Although the model's performance metrics remain strong, its stability on external datasets may be constrained due to potential overadaptation to the training data.

The configuration utilizing a Polynomial kernel without the application of SMOTE achieved the highest classification accuracy (97.06%), alongside a robust AUC score (95.24%), perfect specificity (100%), and high sensitivity (95.24%). These results indicate excellent classification performance within the study dataset. Nevertheless, the absence of SMOTE raises concerns regarding class imbalance, particularly in scenarios where T2DM cases are underrepresented. The GE of 0.1572 further suggests that, despite high training accuracy, the model's generalization capability is constrained, indicating a heightened risk of overfitting to the training distribution.

Consequently, the RBF-kernel SVM integrated with MSVM-RFE and SMOTE was selected as the final configuration for performance reporting and subsequent discussion, as it provides the most favorable trade-off between predictive accuracy, robustness, and generalization capability.

3.4 Limitation

Despite the encouraging performance achieved by the proposed MSVM-RFE framework, several limitations should be acknowledged. First, the computational complexity of MSVM-RFE is considerably higher than that of standard SVM-RFE, as multiple linear SVM models are trained at each recursive elimination step and repeated across cross-validation folds. Although a halving strategy was employed to efficiently reduce the feature space, the overall procedure remains computationally demanding, particularly for large-scale transcriptomic datasets with tens of thousands of genes.

Second, the dataset used in this study is relatively small (118 samples), a common constraint in microarray-based biomedical research that nonetheless limits the statistical power and generalizability of the findings. While SMOTE and cross-validation were applied to mitigate class imbalance and sampling bias, synthetic oversampling cannot fully substitute for additional real-world samples.

Third, while MSVM-RFE incorporates multiple models and cross-validation to enhance accuracy, the feature selection outcome may still be sensitive to data partitioning and initial conditions. In this study, different train-test splits and resampling strategies yielded variations in the selected gene subsets, indicating that the stability of gene selection warrants further optimization. Future work should focus on refining the ensemble mechanism, exploring consensus strategies across multiple runs, or integrating stability selection techniques to achieve more reproducible, robust gene rankings.

Fourth, although the multiple-concept formulation of MSVM-RFE and the use of generalization error analysis substantially reduce the risk of overfitting, the possibility of residual overfitting cannot be entirely excluded, especially given the repeated evaluation of multiple feature subset sizes and kernel configurations. Future work should therefore focus on validating the selected gene subsets in independent external cohorts, optimizing computational efficiency and stability, and extending the framework to larger, more diverse datasets to further assess its robustness and clinical applicability.

4. CONCLUSION

Baseline SVM models without MSVM-RFE exhibited limited discriminative power and high-performance variability across kernels, resampling schemes, and feature dimensionalities, with variance-based gene filtering yielding moderate accuracies and suboptimal AUROC, underscoring its inadequacy for

high-dimensional, low-sample, and class-imbalanced T2DM transcriptomic data. In contrast, the MSVM-RFE + SMOTE framework with optimized hyperparameters consistently achieved superior, stable performance, with the RBF-SVM configuration attaining 95.67% mean accuracy, perfect specificity, 92.86% sensitivity, and a low generalization error (0.0149), indicating effective mitigation of the curse of dimensionality and overfitting. Importantly, the biological coherence of the selected gene subset further supports the framework's validity, as recurrent prioritization of GNAS, known to regulate insulin secretion and pancreatic β-cell signaling, and ASPA suggests biologically meaningful feature selection, warranting validation in independent cohorts and integrative pathway-level or multi-omics analyses.

Author Contributions

Andi Khalil Gibran Basir: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft, Writing – Review and Editing. Ahmad Husain: Methodology, Validation, Writing – Review and Editing. A. Fuad Ahsan Basir: Project Administration, Visualization, Writing – review and editing. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

This research was made possible in part through the dedicated support of Narasiodata's Chief Research Officer, Farida.

Declarations

The authors declare that there are no conflicts of interest related to the publication of this paper.

Declaration of Generative AI and AI-assisted technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text.

REFERENCES

[1] C. Shi, "DNA MICROARRAY TECHNOLOGY PRINCIPLES AND APPLICATIONS IN GENETIC RESEARCH," 2024, doi: https://doi.org/10.54097/a9b7d148

[2] Q. Chen, Z. Meng, and R. Su, "WERFE: A GENE SELECTION ALGORITHM BASED ON RECURSIVE FEATURE ELIMINATION AND ENSEMBLE STRATEGY," *Front. Bioeng. Biotechnol.*, vol. 8, May 2020, doi: https://doi.org/10.3389/fbioe.2020.00496

[3] X. Zhang, I. Jonassen, and A. Goksøyr, "MACHINE LEARNING APPROACHES FOR BIOMARKER DISCOVERY USING GENE EXPRESSION DATA," in *Bioinformatics*, Exon Publications, pp. 53–64, 2021, doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch4

[4] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A REVIEW OF FEATURE SELECTION METHODS FOR MACHINE LEARNING-BASED DISEASE RISK PREDICTION," *Frontiers Media SA*, 2022, doi: https://doi.org/10.3389/fbinf.2022.927312

[5] Z. Rustam and S. A. A. Kharis, "COMPARISON OF SUPPORT VECTOR MACHINE RECURSIVE FEATURE ELIMINATION AND KERNEL FUNCTION AS FEATURE SELECTION USING SUPPORT VECTOR MACHINE FOR LUNG CANCER CLASSIFICATION," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jan. 2020, doi: https://doi.org/10.1088/1742-6596/1442/1/012027

[6] Y. Zhang *et al.*, "EXPLORING POTENTIAL DIAGNOSTIC MARKERS AND THERAPEUTIC TARGETS FOR TYPE 2 DIABETES MELLITUS WITH MAJOR DEPRESSIVE DISORDER THROUGH BIOINFORMATICS AND IN VIVO EXPERIMENTS," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: https://doi.org/10.1038/s41598-025-01175-z

[7] S. Al-Azani, O. S. Alkhnbashi, E. Ramadan, and M. Alfarraj, "GENE EXPRESSION-BASED CANCER CLASSIFICATION FOR HANDLING THE CLASS IMBALANCE PROBLEM AND CURSE OF DIMENSIONALITY," *Int. J. Mol. Sci.*, vol. 25, no. 4, Feb. 2024, doi: https://doi.org/10.3390/ijms25042102

[8] J. Yang, J. Zhou, Z. Zhu, X. Ma, and Z. Ji, "ITERATIVE ENSEMBLE FEATURE SELECTION FOR MULTICLASS CLASSIFICATION OF IMBALANCED MICROARRAY DATA," *Journal of Biological Research (Greece)*, vol. 23, 2016, doi: https://doi.org/10.1186/s40709-016-0045-8

[9] R. F. W. Pratama, S. W. Purnami, and S. P. Rahayu, "BOOSTING SUPPORT VECTOR MACHINES FOR IMBALANCED MICROARRAY DATA," in *Procedia Computer Science*, Elsevier B.V., pp. 174–183, 2018, doi: https://doi.org/10.1016/j.procs.2018.10.517

[10] Y. Lee, M. Cappellato, and B. Di Camillo, "MACHINE LEARNING-BASED FEATURE SELECTION TO SEARCH STABLE MICROBIAL BIOMARKERS: APPLICATION TO INFLAMMATORY BOWEL DISEASE," *Gigascience*, vol. 12, 2023, doi: https://doi.org/10.1093/gigascience/giad083

[11] I. Guyon, J. Weston, and S. Barnhill, "GENE SELECTION FOR CANCER CLASSIFICATION USING SUPPORT VECTOR MACHINES," 2002, doi: https://doi.org/10.1023/A:1012487302797

[12] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, "SVM-RFE: SELECTION AND VISUALIZATION OF THE MOST RELEVANT FEATURES THROUGH NON-LINEAR KERNELS," *BMC Bioinformatics*, vol. 19, no. 1, Nov. 2018, doi: https://doi.org/10.1186/s12859-018-2451-4

[13] K. Yan and D. Zhang, "FEATURE SELECTION AND ANALYSIS ON CORRELATED GAS SENSOR DATA WITH RECURSIVE FEATURE ELIMINATION," *Sens. Actuators B Chem.*, vol. 212, pp. 353–363, 2015, doi: https://doi.org/10.1016/j.snb.2015.02.025

[14] D. Yang and X. Zhu, "GENE CORRELATION GUIDED GENE SELECTION FOR MICROARRAY DATA CLASSIFICATION," *Biomed Res. Int.*, vol. 2021, 2021, doi: https://doi.org/10.1155/2021/6490118

[15] N. N. M. Hasri, N. H. Wen, C. W. Howe, M. S. Mohamad, S. Deris, and S. Kasim, "IMPROVED SUPPORT VECTOR MACHINE USING MULTIPLE SVM-RFE FOR CANCER CLASSIFICATION," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 4–2 Special Issue, pp. 1589–1594, 2017, doi: https://doi.org/10.18517/ijaseit.7.4-2.3394

[16] Y. Zhang, Q. Deng, W. Liang, and X. Zou, "AN EFFICIENT FEATURE SELECTION STRATEGY BASED ON MULTIPLE SUPPORT VECTOR MACHINE TECHNOLOGY WITH GENE EXPRESSION DATA," *Biomed Res. Int.*, vol. 2018, 2018, doi: https://doi.org/10.1155/2018/7538204

[17] J. Taneera *et al.*, "GNAS GENE IS AN IMPORTANT REGULATOR OF INSULIN SECRETORY CAPACITY IN PANCREATIC B-CELLS," *Gene*, vol. 715, Oct. 2019, doi: https://doi.org/10.1016/j.gene.2019.144028

[18] L. Wang *et al.*, "ASSOCIATED FACTORS AND PRINCIPAL PATHOPHYSIOLOGICAL MECHANISMS OF TYPE 2 DIABETES MELLITUS," *Front. Endocrinol. (Lausanne).*, vol. 16, 2025, doi: https://doi.org/10.3389/fendo.2025.1499565

[19] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos, *DATA COMPLEXITY MEASURES FOR ANALYZING THE EFFECT OF SMOTE OVER MICROARRAYS*. [Online]. Available: http://www.i6doc.com/en/.

[20] R. L. Seal *et al.*, "GENENAMES.ORG: THE HGNC RESOURCES IN 2023," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1003–D1009, Jan. 2023, doi: https://doi.org/10.1093/nar/gkac888

[21] E. A. Bruford, B. Braschi, P. Denny, T. E. M. Jones, R. L. Seal, and S. Tweedie, "GUIDELINES FOR HUMAN GENE NOMENCLATURE,", *Nature Research*, Aug. 01, 2020, doi: https://doi.org/10.1038/s41588-020-0669-3

[22] M. Athoillah, E. Purnaningrum, and R. K. Putri, "MODIFIED MULTI-KERNEL SUPPORT VECTOR MACHINE FOR MASK DETECTION," [Online]. Available: https://github.com/prajnasb, 2022, doi: https://doi.org/10.21512/commit.v16i2.7873

[23] P. Jin, L. Lu, Y. Tang, and G. E. Karniadakis, "QUANTIFYING THE GENERALIZATION ERROR IN DEEP LEARNING IN TERMS OF DATA DISTRIBUTION AND NEURAL NETWORK SMOOTHNESS," *Neural Networks*, vol. 130, pp. 85–99, Oct. 2020, doi: https://doi.org/10.1016/j.neunet.2020.06.024

[24] S. Wolf, D. Melo, K. M. Garske, L. F. Pallares, A. J. Lea, and J. F. Ayroles, "CHARACTERIZING THE LANDSCAPE OF GENE EXPRESSION VARIANCE IN HUMANS," *PLoS Genet.*, vol. 19, no. 7 July, Jul. 2023, doi: https://doi.org/10.1371/journal.pgen.1010833

[25] Y. Xie, Z. Jing, H. Pan, X. Xu, and Q. Fang, "REDEFINING THE HIGH VARIABLE GENES BY OPTIMIZED LOESS REGRESSION WITH POSITIVE RATIO," *BMC Bioinformatics*, vol. 26, no. 1, Dec. 2025, doi: https://doi.org/10.1186/s12859-025-06112-5

[26] C. Savas and F. Dovis, "THE IMPACT OF DIFFERENT KERNEL FUNCTIONS ON THE PERFORMANCE OF SCINTILLATION DETECTION BASED ON SUPPORT VECTOR MACHINES," *Sensors (Switzerland)*, vol. 19, no. 23, Dec. 2019, doi: https://doi.org/10.3390/s19235219

[27] D. Aryo Anggoro and D. Permatasari, "PERFORMANCE COMPARISON OF THE KERNELS OF SUPPORT VECTOR MACHINE ALGORITHM FOR DIABETES MELLITUS CLASSIFICATION," [Online]. Available: www.ijacsa.thesai.org. 2023, doi: https://doi.org/10.14569/IJACSA.2023.0140163

[28] R. Trevethan, "SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUES: FOUNDATIONS, PLIABILITIES, AND PITFALLS IN RESEARCH AND PRACTICE," *Front. Public Health*, vol. 5, Nov. 2017, doi: https://doi.org/10.3389/fpubh.2017.00307

[29] A. M. Rahmani *et al.*, "MACHINE LEARNING (ML) IN MEDICINE: REVIEW, APPLICATIONS, AND CHALLENGES," *MDPI*, Nov. 01, 2021, doi: https://doi.org/10.3390/math9222970

[30] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A COMPREHENSIVE REVIEW ON MACHINE LEARNING IN HEALTHCARE INDUSTRY: CLASSIFICATION, RESTRICTIONS, OPPORTUNITIES AND CHALLENGES," May 01, *MDPI*. 2023, doi: https://doi.org/10.3390/s23094178

[31] J. Zhu *et al.*, "PROCESSING IMBALANCED MEDICAL DATA AT THE DATA LEVEL WITH ASSISTED-REPRODUCTION DATA AS AN EXAMPLE," *BioData Min.*, vol. 17, no. 1, Dec. 2024, doi: https://doi.org/10.1186/s13040-024-00384-y