Early Detection of Ovarian Cancer from Gene Expression Data Using an Optimized Multiple Indefinite Kernel-Based Twin Support Vector Machine with the Adaptive Lyrebird Optimization Algorithm

S. Swetha
Department of Information Science and Engineering, RV College of Engineering, Bengaluru-560059, affiliated to Visvesvaraya Technological University, Belagavi Karnataka, India
shwetha.ise@rvce.edu.in (corresponding author)

G. N. Srinivasan
Department of Information Science and Engineering, RV College of Engineering, Bengaluru-560059, affiliated to Visvesvaraya Technological University, Belagavi Karnataka, India
srinivasangn@rvce.edu.in

M. R. Anala
Department of Information Science and Engineering, RV College of Engineering, Bengaluru-560059, affiliated to Visvesvaraya Technological University, Belagavi Karnataka, India
analamr@rvce.edu.in

Received: 8 October 2025 | Revised: 5 December 2025 | Accepted: 19 February 2026

 DOI: https://doi.org/10.48084/etasr.15393

ABSTRACT

Ovarian cancer, a frequent gynecological tumor, is usually symptom-free in its early phases; therefore, early diagnosis is important. Gene expression microarrays have great potential for diagnosing and treating ovarian cancer by allowing high-throughput gene examination. However, processing such data presents several challenges, including noise, redundancy, errors, increased complexity, small sample sizes with high dimensionality, and difficulties in interpretation. Recently, microarray datasets have been analyzed using Machine Learning (ML) approaches to classify ovarian cancer. While ML methods are promising in ovarian cancer classification, most models still face issues such as class imbalance, excessive computational expense, weak scalability, and limited interpretability, which limit their clinical utility. To address these constraints, in the present research, a new ML-based model named Optimized Multiple Indefinite kernel-based Twin Support Vector Machine (OMI-Twin SVM) is proposed for the classification of ovarian cancer based on gene expression datasets. The collected gene expression dataset is pre-processed, including data cleaning and normalization to ensure data quality. To select the most critical gene expression features for precise classification of ovarian cancer, the minimum Redundancy–Maximum Relevance (mRMR) approach is applied. Our suggested classification model employs the OMI-Twin SVM structure to predict whether gene expression samples belong to normal or cancerous categories. The proposed classifier is based on a Twin Support Vector Machine (TWSVM) system that incorporates multiple indefinite kernels instead of positive semi-definite kernels alone. To enhance classification performance, we propose an Adaptive Lyrebird Optimization (ALO) algorithm that alternately optimizes the kernel combination and coefficients of the OMI-Twin SVM.

Keywords-gene expression data; ovarian cancer; multiple indefinite kernels; Twin Support Vector Machine (TWSVM); Adaptive Lyrebird Optimization (ALO) algorithm; Support Vector Machine (SVM)

I. INTRODUCTION

Ovarian cancer is the second most frequent cancer in women and the primary cause of cancer-related deaths in developing countries [1]. Early identification and appropriate treatment can reduce the chance of damaging other cells. However, early detection is quite challenging due to the disordered structure of cancer cells [2]. Furthermore, traditional treatment methods like chemotherapy and surgery have little effect on the disease's survival rate. Therefore, developing new methods for ovarian cancer in the early stages of the disease and creating individualized treatment plans are necessary to improve clinical efficacy and safety [3].

Over the past 20 years, research in health informatics has focused on a variety of topics, including cheminformatics, bioinformatics, and cancer prediction [4]. Scientists can now evaluate thousands of genes' activity at once because of recent developments in gene expression profiling technology, which provides crucial information for cancer diagnosis, classification, and outcome prediction [5-7]. However, the analysis of gene expression data is quite challenging because of their huge dimensions, complexity, and feature value duplications [8, 9].

Machine Learning (ML) has also proven to be an effective method for cancer prediction with its ability to learn deep, nonlinear patterns from large-scale biomedical data [10]. Authors in [11] employed artificial neural networks combined with nano-biosensors for non-invasive detection of ovarian cancer from blood samples, achieving promising sensitivity but lacking interpretability. Authors in [12] applied multi-omics integration and ensemble learning techniques such as Support Vector Machine–Recursive Feature Elimination (SVM-RFE), Extreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN), though their approach required complex preprocessing and validation. Generative Adversarial Network (GAN)-based data augmentation has also been explored to overcome limited sample availability, as seen in [13, 14], but these approaches suffer from instability and high computational cost. Other studies have utilized deep neural networks such as ResNet, transformers, and convolutional architectures for ovarian cancer diagnosis; however, most of these frameworks require large datasets and extensive computational resources, limiting their scalability and clinical feasibility [15-18].

II. PROPOSED METHODOLOGY

A. Overview

First, the collected gene expression dataset is pre-processed, including data cleaning and normalization to ensure data quality. For feature selection, the minimum Redundancy–Maximum Relevance (mRMR) algorithm is utilized to determine the most informative features for accurate ovarian cancer classification. The proposed classification model employs the Optimized Multiple Indefinite kernel-based Twin Support Vector Machine (OMI-Twin SVM) architecture for classifying the gene expression dataset into normal or malignant classes. The proposed classifier is based on a Twin Support Vector Machine (TWSVM) model incorporating multiple indefinite kernels instead of using only positive semi-definite kernels. An indefinite base kernel is assigned to each feature, which allows the model to capture complex nonlinear relationships in the gene expression data.

To improve classification performance, we propose an Adaptive Lyrebird Optimization (ALO) algorithm to optimize the OMI-Twin SVM's coefficients and kernel combination alternately, in order to enhance overall classification performance and accuracy. The overall framework of the proposed methodology is demonstrated in Figure 1.

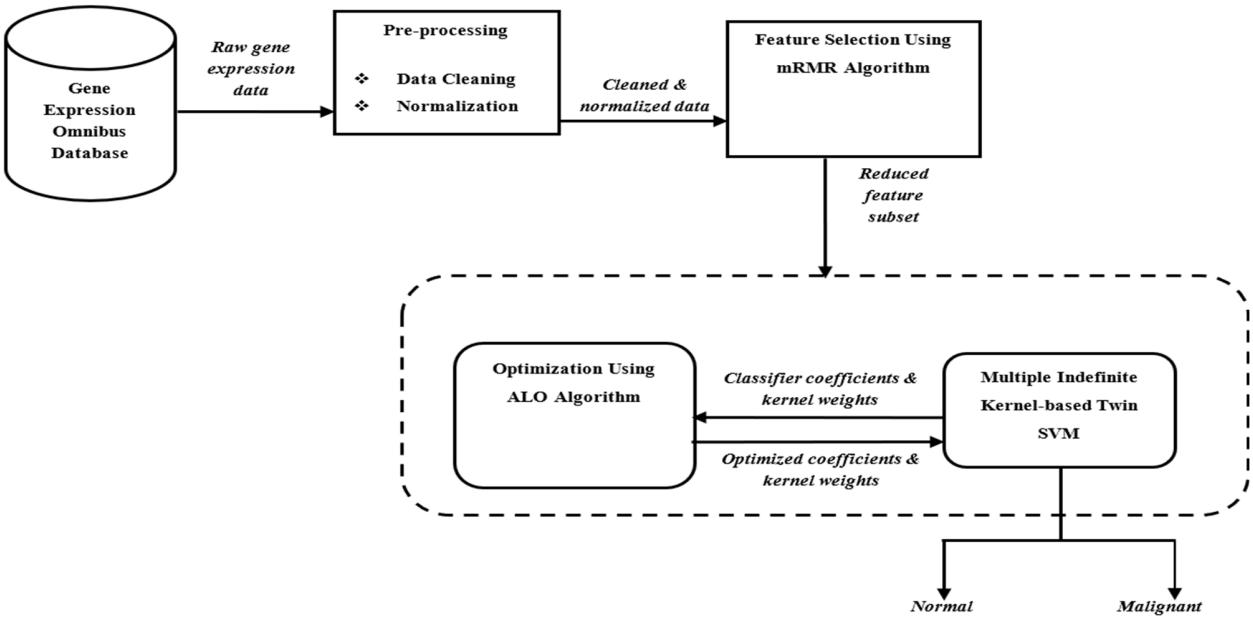

Fig. 1. Overall architecture of the proposed methodology.

B. Pre-Processing

Owing to experimental variations, gene expression datasets obtained from the Gene Expression Omnibus (GEO) database often contain noise, missing values, and errors. Hence, before applying feature selection and classification, pre-processing is a necessary step to make the data reliable and accurate. During the pre-processing phase, a data cleaning process was undertaken to verify the quality and reliability of the dataset. The process entailed identifying and resolving major data quality problems, including outliers, duplicates, missing values, and inconsistencies, all of which adversely affect model performance and interpretability.

Subsequently, normalization was applied to the dataset to ensure that every gene expression feature contributed equally to model training. Normalization seeks to scale the dataset's numeric values to a standard range, improving model performance and accuracy without altering the underlying data distribution or losing any information.

C. Feature Selection Using the Minimum Redundancy–Maximum Relevance Algorithm

The mRMR method is used to determine the most important gene expression features for accurate ovarian cancer classification. mRMR is a minimal-optimal multivariate filter-based feature selection method that selects the most relevant and non-redundant features. In this study, the GEO microarray gene expression data are continuous, whereas the response variable labels are binary. The F-statistic and Pearson correlation coefficient are employed to determine the relevance and redundancy of gene biomarkers with respect to class labels (0 and 1).

The following equations describe how the mRMR algorithm works. The F-statistic between the genes and the target variable H can be used to select the highest relevance score for continuous features (individual gene expression data). Equation (1) represents the F-test result of gene (feature) G_i in K classes, denoted as H:

$$f\left(G_i,H\right)=\left[\sum_k n_k\left(\bar{G}_k-\bar{G}\right)/\left(K-1\right)\right]/\sigma^2 \qquad (1)$$

The variance, along with the size of the k-th class, are denoted by σ_k and n_k, G represents the average value of a gene (G_i) for every tissue sample, $\bar{G}_k$ denotes the average value of G_i inside the k-th class, K represents the total number of classes, and σ^2 represents the overall variance. The F-test can be reduced to the t-test for the two-class classification problem via the relation $f=t_2$. As a result, the highest relevance ($\max v_f$) for the gene biomarker set can be expressed as:

$$v_f=\frac{1}{|s|}\sum_{i\in s} f\left(i,H\right) \qquad (2)$$

The present study uses the Pearson correlation coefficient $C(G_i,G_j)$ to specify the minimum redundancy criterion among the input variables. Both high negative and positive correlations are considered indicators of redundancy. Therefore, the absolute value of these correlations is used; hence, the particular circumstance (minimum relevance $(\min w_C)$ between the input variables) is expressed as:

$$w_C=\frac{1}{|s|^2}\sum_{i,j}\left|C(i,j)\right| \qquad (3)$$

The mRMR optimization criterion, combining the F-statistic with a correlation quotient, is then used to filter the selected features. Finally, we combine relevance and redundancy using a quotient-based formulation, expressed as follows:

$$\max_{i\in\Omega_s}\left\{f\left(i,H\right)\Big/\left[\frac{1}{|s|}\sum_{j\in s}\left|C\left(i,j\right)\right|\right]\right\} \qquad (4)$$

Following that, a subset of data is generated using selected genes (features) from the mRMR algorithm, which is then fed into the proposed OMI-Twin SVM classifier for ovarian cancer diagnosis.

D. OMI-Twin SVM for Ovarian Cancer Detection

After feature selection, the reduced feature subset is forwarded to the proposed OMI-Twin SVM classifier for ovarian cancer detection. Unlike conventional SVM classifiers that rely on a single positive semi-definite kernel, the OMI-Twin SVM incorporates a multiple indefinite kernels combination strategy that effectively captures the complex, non-linear structure of high-dimensional biomedical data.

The classifier is built upon the primal framework of the Indefinite Kernel Twin SVM (IKTSVM) [19], where each feature can be associated with an indefinite base kernel.

1) OMI-Twin SVM Formulation

Given a training set $\left(u_i,v_i\right), i=1,2,....,n$, where $u_i\in\mathbb{R}^m$ and $v_i\in\{-1,+1\}$, $i=1,2,....,n$, n represents the number of training samples, and m represents the dimension of the training samples. In the m-dimensional feature space, n_1 samples belong to class +1, whereas n_2 samples belong to class −1.

The purpose of OMI-Twin SVM for the linearly separable binary classification problem is to identify two non-parallel hyperplanes:

$$u_i^T W_1+b_1=0 \text{ and } v_i^T W_2+b_2=0 \qquad (5)$$

The OMI-Twin SVM model positions each hyperplane as close as possible to the samples of one class while keeping it as far as possible from those of the other class. In contrast to the original TWSVM, OMI-Twin SVM incorporates a regularization term to minimize structural risk and uses a smooth quadratic hinge loss function to make the model more robust and differentiable when indefinite kernels are utilized. The two optimization problems can be expressed as follows:

OMI-Twin SVM1:

$$\min_{W_1,b_1} \frac{1}{2}\|W_1\|^2 + \frac{1}{2}(PW_1 + d_1b_1)^T(PW_1 + d_1b_1) + a_1\chi^T\chi \quad (6)$$
$$\text{s.t.} \quad QW_1 + d_2b_1 + \chi \geq d_2,\ \chi \geq 0$$

OMI-Twin SVM2:

$$\min_{W_2,b_2} \frac{1}{2}\|W_2\|^2 + \frac{1}{2}(QW_2 + d_2b_2)^T(QW_2 + d_2b_2) + a_2\chi^T\chi \quad (7)$$
$$\text{s.t.} \quad PW_2 + d_1b_2 + \chi \geq d_1,\ \chi \geq 0$$

where a_1 and a_2 are penalty parameters, d_1 and d_2 are vectors of ones, χ and ∂ are both slack variables, and P in $\mathbb{R}^{\partial_1 \times m}$ and Q in $\mathbb{R}^{\partial_2 \times m}$ are the training sample matrices corresponding to the positive and negative classes, respectively.

From (6) and (7), the distance between the proximal hyperplanes $u^TW_i + b_i = 0$ and the bounding hyperplanes $u^TW_i + b_i = \pm 1$ is $1/\|W_i\|$, where $i = 1, 2$. As a result, the additional term in the objective function ensures that the bounding and proximal hyperplanes are separated as much as possible.

Lastly, the proposed method is more mathematically robust than the original TWSVM, as OMI-Twin SVM retains the advantages of standard SVM. To make the OMI-Twin SVM model continuously differentiable, a smooth quadratic hinge loss function is applied to the slack term.

Equations (6) and (7) are then reformulated as unconstrained optimization problems:

OMI-Twin SVM1:

$$\min_{W_1,b_1} \lambda\langle W_1, W_1\rangle + \frac{1}{2}\|PW_1 + d_1b_1\|^2 + a_1\|\max(0, d_2 + QW_1 + d_2b_1)\|^2 \quad (8)$$
$$= \lambda\langle W_1, W_1\rangle + \sum_{i=1}^{n} L(\langle W_1, u_i\rangle + b_1)$$

OMI-Twin SVM2:

$$\min_{W_2,b_2} \lambda\langle W_2, W_2\rangle + \frac{1}{2}\|QW_2 + d_2b_2\|^2 + a_2\|\max(0, d_1 + PW_2 + d_1b_2)\|^2 \quad (9)$$
$$= \lambda\langle W_2, W_2\rangle + \sum_{i=1}^{n} L(\langle W_2, u_i\rangle + b_2)$$

Equations (8) and (9) show that each OMI-Twin SVM consists of two components: the regularized term $\lambda\langle W, W\rangle$ and the loss function term $\sum_{i=1}^{n} L(\langle W, u_i\rangle + b)$.

2) Kernelization with Multiple Indefinite Kernels

The Representer Theorem states that we can use a kernel to expand (8) and (9) in Reproducing Kernel Hilbert Spaces (RKHS), which is expressed as:

OMI-Twin SVM1:

$$\min_{F_1,b_1} \lambda\langle F_1, F_1\rangle_k + \sum_{i=1}^{n} L_1(\langle W_1, u_i\rangle + b_1) \quad (10)$$

OMI-Twin SVM2:

$$\min_{F_2,b_2} \lambda\langle F_2, F_2\rangle_k + \sum_{i=1}^{n} L_2(\langle W_2, u_i\rangle + b_2) \quad (11)$$

Taking OMI-Twin SVM1 as an example, $\lambda\langle W_1, W_1\rangle$ can be represented as $\lambda\langle F_1, F_1\rangle_k$ and L_1 denotes a loss function. Expanding (10) and (11) in a wider Reproducing Kernel Krein Spaces (RKKS) is possible when the kernel is indefinite.

The Representer Theorem remains valid in RKKS, and the regularized risk function minimization task can be generalized as:

$$F^* = \sum_{i=1}^{n} \beta_i k(u_i) \quad (12)$$

Here, the coefficient $\beta_i \in \Re$, and k denotes the kernel function in RKKS. Furthermore, the OMI-Twin SVM1 model in RKKS can be expressed as:

OMI-Twin SVM1:

$$\min_{\beta_1,b_1} \lambda\beta^T k + \sum_{i=1}^{n} L_1(k^i\beta + b_1) \quad (13)$$

Here, $\beta = [\beta_1, \beta_2, ..., \beta_n]^T$, k denotes the indefinite kernel matrix derived from the respective kernel function $k_{i,j} = k(u_i, u_j)$, and k_i denotes the i-th row of k. This formulation, referred to as the kernelized Twin SVM1, can be directly applied even when k is indefinite, giving rise to the OMI-Twin SVM1. The flexibility of using indefinite kernels is particularly advantageous for biomedical datasets, where the similarity structure between samples may not conform to positive semi-definite kernels.

Nevertheless, (13) still relies on a single kernel function. To enhance flexibility, we extend the model using a multiple indefinite kernels strategy. Following the idea of multiple kernel learning, the kernel matrix is constructed as a weighted sum of base kernels, i.e.:

$$K(u_i, U) = \sum_{m=1}^{M} e_m K_m(u_{i,m}, U_m),\ e_m \geq 0 \quad (14)$$

where $u_{i,m}$ represents the m-th feature of u_i, and e_m denotes the coefficient of the kernel K_m.

By embedding this kernel combination into the ITWSVM framework, the two optimization problems for OMI-Twin SVM can be expressed as follows:

OMI-Twin SVM1:

$$\min_{\beta_1,b_1} \frac{\frac{1}{2}\lambda\beta_1^T K\beta_1 + \frac{1}{2}\sum_{i=1}^{n_1}(K^i\beta_1 + b_1)^2 + a_1\sum_{i=n_1+1}^{n_1+n_2}\left(\max(0, K^i\beta_1 + b_1 + 1)\right)^2}{\sum_{i=1}^{n} L_1(F_1(u_i) + b_1)} \quad (15)$$

OMI-Twin SVM2:

$$\min_{\beta_2,b_2} \frac{\frac{1}{2}\lambda\beta_2^T K\beta_2 + \frac{1}{2}\sum_{i=1}^{n_1}(K^i\beta_2 + b_2)^2 + a_2\sum_{i=n_1+1}^{n_1+n_2}\left(\max(0, K^i\beta_2 + b_2 + 1)\right)^2}{\sum_{i=1}^{n} L_2(F_2(u_i) + b_2)} \quad (16)$$

where n_1 denotes the number of samples belonging to class +1, n_2 denotes the number of samples belonging to class −1, and $n = n_1 + n_2$. To distinguish β in OMI-Twin SVM1 and OMI-Twin SVM2, we set β as β_1 in OMI-Twin SVM1 and β_2 in OMI-Twin SVM2, respectively.

E. Optimization of OMI-Twin SVM Using Adaptive Lyrebird Optimization

To further enhance classification performance, we propose a novel ALO algorithm to optimize the coefficients of OMI-Twin SVM (β, b) and the kernel combination (e) alternately, thereby enhancing overall classification performance and accuracy. The ALO metaheuristic algorithm is bio-inspired and relies on the behavior of lyrebirds in the wild.

The proposed ALO algorithm is adopted from the Lyrebird Optimization Algorithm (LOA) [20], which is based on the principles of exploration and exploitation. In general, in LOA, fixed parameter settings may lead to needless wandering or premature convergence.

To address this limitation, the proposed adaptive mechanism introduces dynamic weighting, which continuously adjusts the balance between exploration and exploitation according to iteration progress and solution quality. The procedures for tuning the OMI-Twin SVM classifier are as follows.

1) Initialization

The Lyrebird algorithm is responsible for initializing the population, which is sometimes referred to as the *P* matrix. As such, each lyrebird represents a vector-based solution corresponding to the hyperplane coefficients (β, b) and kernel weights (e) . The initial population matrix *P* is generated randomly within feasible ranges and is considered a set of viable starting solutions for optimizing the OMI-Twin SVM classifier:

$$P_i = \{p_1, p_2, ..., p_n\} \quad (17)$$

Here, *n* and P_i indicate the number of variables and the population size of the solution, respectively, and $P_1 = \{\beta, b, e\}$ represents the set of hyperplane coefficients and kernel weights.

2) Fitness Function

In every iteration, the fitness value of a solution is evaluated, and the overall fitness function is derived by selecting the highest accuracy obtained among the candidate solutions, which can be expressed as:

$$fitness = Max(accuracy) \quad (18)$$

3) Update Phase

In the architecture of ALO, (19) mimics the lyrebird's decision-making process when choosing between hiding and escaping under threat. As a result, in every iteration, each ALO member's position is updated based on either the first or the second phase.

$$\text{Update process for } P_i: \begin{cases} \text{based on Phase 1} & rand \le 0.5 \\ \text{based on Phase 2} & \text{else} \end{cases} \quad (19)$$

Here, *rand* represents a random number drawn from the range [0, 1].

a) Phase 1: Escaping Strategy

Every population member's location is continuously updated during the ALO stage to simulate the lyrebird's movement from risk to safe zones. These safe zones are determined by applying (20):

$$Safe\_area_i = \{P_k \mid f_k < f_i, k \in \{1, 2, \ldots, N\}\}, \quad i = 1, 2, \ldots, N \quad (20)$$

where $Safe_{area\,i}$ denotes the set of safe locations for the *i*-th lyrebird, and P_k represents the *k*-th row of the *P* matrix having a higher objective function value ($f_k < f_i$). The lyrebird selects one of these safe locations randomly, and each member's new position is determined using the lyrebird displacement model, as in (21).

If the objective function value improves, the corresponding member is replaced with this new position, as in (22).

$$p_{i,j}^{L1} = p_{i,j} + rand_{i,j} \cdot \left(SSafe_{area\,i,j} - R_{i,j} \cdot p_{i,j}\right) \quad (21)$$

$$P_i = \begin{cases} P_i^{L1}, & f_i^{L1} \le f_i \\ P_i, & \text{else} \end{cases} \quad (22)$$

Here, $SSafe_{area\,i,j}$ denotes the selected safe location for the *i*-th lyrebird along the *j*-th dimension, $p_{i,j}^{L1}$ represents the newly determined position of the *i*-th lyrebird using the LOA escape procedure along its *j*-th dimension, and f_i^{L1} represents the value of its objective function. Additionally, $rand_{i.j}$ are random values in [0, 1], and $R_{i,j}$ is randomly selected from {1,2}.

b) Phase 2: Hiding Strategy

Equation (23), which represents the lyrebird's movement toward a nearby safe location, is used to determine an alternative position for each member of the ALO framework. The member's previous position is replaced if the new position improves the objective function value, as expressed in (24).

$$p_{i,j}^{L2} = p_{i,j} + \left(1 - 2rand_{i,j}\right) \cdot \frac{Ub_j - Lb_j}{t} \tag{23}$$

$$P_i = \begin{cases} P_i^{L2} & f_i^{L2} \le f_i \\ P_i & \text{else} \end{cases} \tag{24}$$

Here, $p_{i,j}^{L2}$ is the newly estimated position generated by the *i*-th lyrebird using the LOA hiding strategy along its *j*-th dimension, f_i^{L2} is its objective function value, $rand_{i,j}$ denotes random values in [0, 1], and *t* represents the iteration counter.

4) Dynamic Weighting Strategy

The ALO refines Phase 2 of the original LOA. A dynamic weighting mechanism is introduced to adaptively adjust the weight value at each iteration. In Phase 2, the weight ω is iteratively updated to balance exploration and exploitation. The modified equation is given as:

$$p_{i,j}^{L2} = p_{i,j} + \omega\left(1 - 2rand_{i,j}\right)\frac{Ub_j - Lb_j}{t} \tag{25}$$

Within this framework, ω functions as a parameter that controls the search scales during the iteration. This process starts with a high value, around a quarter of the typical search domain scaling, and gradually drops to about 1% of the quarter of this length. The following monotonically decreasing function improves the stability of ALO:

$$\omega = \left(\left(\omega_o - \omega_\infty\right)/\left(1 - t_{\max}\right)\right) \times \left(t - t_{\max}\right) \| + \omega_\infty \tag{26}$$

In this case, the initial and final values are denoted by ω_o and ω_∞, respectively. In essence, ω takes charge of the iteration process. Typically, ω_o and ω_∞ can be configured as follows

$$\omega_o = (Ub - Lb)/4 \tag{27}$$

$$\omega_\infty = \omega_o/100 \tag{28}$$

Here, *Ub* and *Lb* represent the upper and lower bounds, respectively, and *t* denotes the current iteration, whereas $t_{\max}$ is the maximum number of iterations. This modified version of LOA provides improved solution quality and faster convergence compared to the original.

5) Termination

The population update process is repeated until the algorithm reaches the maximum number of iterations. At the end of the optimization process, the result corresponds to the optimal set of hyperplane parameters for the OMI-Twin SVM classifier.

F. Classification Rule for OMI-Twin SVM

After training, the OMI-Twin SVM produces two hyperplanes, one for each class, in the multiple indefinite kernels space. For a new sample u^*, its class is predicted as the hyperplane to which it is closest, measured in the kernel space:

$$Class(u^*) = \arg\min_{i=1,2} \frac{\left|K\left(u^*, U\right)^T \beta_i + b_i\right|}{\sqrt{\beta_i^T K \beta_i}} \tag{29}$$

where $K\left(u^*, U\right)$ is the combined multiple indefinite kernels vector between the new sample and all training samples, *K* is the kernel matrix of the training set, and β_i, b_i are the learned coefficients and bias for the *i*-th hyperplane.

III. RESULTS AND DISCUSSION

The proposed framework was implemented using Python. A CPU-based computer system with a 2 GHz Intel Core i7 processor, 8 GB of RAM, and 256 GB of storage was used for the implementation.

A. Dataset Description

The GEO database [21] was used as the benchmark dataset for training and testing the proposed models. GEO is a publicly accessible functional genomics data repository maintained by the National Center for Biotechnology Information (NCBI), designed to store and distribute high-throughput gene expression and molecular abundance data derived from microarray and next-generation sequencing experiments. It supports MIAME-compliant datasets and enables researchers to access curated gene expression profiles along with detailed experimental metadata.

In this study, ovarian cancer-related gene expression data were retrieved from GEO using disease-specific keywords to ensure the relevance and reliability of the dataset for classification tasks.

B. Comprehensive Performance Evaluation across Multiple Scenarios

1) Performance Evaluation of the Proposed Classifier with Other Baseline Methods

This section compares the effectiveness of the proposed OMI-Twin SVM method to IKTSVM, TWSVM, and SVM using the GEO dataset in terms of accuracy, recall, precision, F-score, and specificity. Table I depicts the comparative

analysis of the proposed and baseline methods using the GEO dataset.

The proposed OMI-Twin SVM method outperforms the IKTSVM, TWSVM, and SVM models in terms of all evaluation metrics. The experimental results demonstrate that the proposed OMI-Twin SVM method outperforms the IKTSVM model by 1.66% in accuracy, 4.76% in F-score, 0.095% in precision, 6.25% in recall, and 3.57% in specificity. Compared to the TWSVM model, the proposed method achieves notable improvements of 5% in accuracy, 4.76% in F-score, 3.17% in precision, 6.25% in recall, and 3.57% in specificity, respectively. When compared to the SVM model, the proposed method demonstrates significant enhancements of 6.66% in accuracy, 6.4% in F-score, 6.34% in precision, 6.41% in recall, and 6.77% in specificity, respectively.

TABLE I. PERFORMANCE COMPARISON OF THE PROPOSED OMI-TWIN SVM METHOD WITH BASELINE METHODS ON THE GEO DATASET

Evaluation metric	OMI-Twin SVM (proposed)	IKTSVM	TWSVM	SVM
Accuracy (%)	98.33	96.67	93.33	91.67
Precision (%)	96.97	96.875	93.8	90.625
Recall (%)	100	96.88	93.75	93.59
F-score (%)	98.46	96.87	93.7	92.06
Specificity (%)	96.42	96.427	92.85	89.65

C. Effect of mRMR Feature Selection on OMI-Twin SVM Performance

This section analyzes the impact of applying mRMR-based feature selection on the performance of the OMI-Twin SVM model in comparison to its performance without feature selection. Table II depicts the comparative analysis of the proposed method with and without the mRMR feature selection algorithm.

TABLE II. COMPARATIVE ANALYSIS OF THE PROPOSED METHOD WITH AND WITHOUT THE MRMR FEATURE SELECTION ALGORITHM

Evaluation metric	OMI-Twin SVM with mRMR	OMI-Twin SVM without mRMR
Accuracy (%)	98.33	95
Precision (%)	96.97	93.93
Recall (%)	100	96.87
F-score (%)	98.46	95.38
Specificity (%)	96.428	92.85

From Table II, the comparison shows that utilizing mRMR-based feature selection has a significant impact on improving the performance of the OMI-Twin SVM model. With mRMR, the model yielded higher values: 98.33% accuracy, 96.97% precision, 100% recall, 98.46% F-score, and 96.42% specificity. Without mRMR, the model obtained 95% accuracy, 93.93% precision, 96.87% recall, 95.38% F-score, and 92.85% specificity.

D. Comparison with State-of-the-Art Methods

Table III summarizes recent studies on ovarian cancer detection using gene expression and other biomarker-based approaches, along with the datasets employed, classification models used, and their reported performance.

TABLE III. COMPARISON OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS

Ref.	Key approach	Datasets used	Classification model	Result analysis
[17]	Transformer-based gene expression cancer classification	GEO and TCGA gene expression datasets	DistilBERT with self-attention	97.56% accuracy
[12]	DL-based ovarian cancer detection using metabolic and lipidomic biomarkers	HPLC-MS and NMR datasets	CNN, XGBoost	96% accuracy
[13]	Genomic data-driven ovarian cancer prediction with data augmentation	High-dimensional genetic dataset	GAN, XGBoost	98.01% accuracy
[14]	Explainable AI-based ensemble model for ovarian cancer detection	Multi-center ovarian cancer dataset	Ensemble classifiers (KNN, Logistic Regression, and SVM)	98.66% accuracy
[15]	Interpretable GAN-based ovarian cancer detection	High-dimensional genetic dataset	GAN	98.34% accuracy
[16]	Fuzzy deep learning for histopathology-based ovarian cancer detection	Ovarian Bevacizumab dataset	ResNet-50 + fuzzy classifier	97.99% accuracy, 99% sensitivity, 98.96% specificity
[11]	SERS-based neutrophil profiling for early ovarian cancer detection	Real-world blood samples (SERS-based signals)	ANN	90% sensitivity
Proposed	Gene expression-based ovarian cancer detection	GEO dataset	OMI-Twin SVM	98.64% accuracy, 98.45% precision, 98.78% recall, 98.34% specificity

IV. CONCLUSION

This paper introduced Optimized Multiple Indefinite kernel-based Twin Support Vector Machine (OMI-Twin SVM), an SVM-based Machine Learning (ML) approach for predicting ovarian cancer based on gene expression data. To ensure data integrity, the gathered gene expression dataset was initially pre-processed by applying data cleaning and normalization. The most significant gene expression characteristics for precise ovarian cancer classification were identified using the minimum Redundancy–Maximum Relevance (mRMR) approach for feature selection. The proposed classification model separated gene expression data into malignant and normal classes using the OMI-Twin SVM framework.

Rather than relying solely on positive semi-definite kernels, the proposed classifier was designed based on a Twin Support Vector Machine (TWSVM) framework that integrates multiple

indefinite kernels. In order to enhance classification performance, an Adaptive Lyrebird Optimization (ALO) algorithm is proposed and employed, which alternately optimizes the kernel combination and coefficients of the OMI-Twin SVM to improve overall classification accuracy and performance.

In the future, integrating clinical and multi-omics data together with gene expression data may reveal more insightful knowledge, making it possible to develop more robust and clinically relevant diagnostic tools.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGMENT

Not applicable to this work.

DATA AVAILABILITY

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] K. S. Prabhu, H. Q. Sadida, S. Kuttikrishnan, K. Junejo, A. A. Bhat, and S. Uddin, "Beyond genetics: Exploring the role of epigenetic alterations in breast cancer," *Pathology - Research and Practice*, vol. 254, Feb. 2024, Art. no. 155174, https://doi.org/10.1016/j.prp.2024.155174.

[2] A. Bajwa *et al.*, "Challenges and opportunities in ovarian cancer care: A qualitative study of clinician perspectives from 24 low- and middle-income countries," *Journal of Cancer Policy*, vol. 44, June 2025, Art. no. 100582, https://doi.org/10.1016/j.jcpo.2025.100582.

[3] M.-K. Hong and D.-C. Ding, "Early Diagnosis of Ovarian Cancer: A Comprehensive Review of the Advances, Challenges, and Future Directions," *Diagnostics*, vol. 15, no. 4, Feb. 2025, Art. no. 406, https://doi.org/10.3390/diagnostics15040406.

[4] M. H. Sadeghi, S. Sina, H. Omidi, A. H. Farshchitabrizi, and M. Alavi, "Deep learning in ovarian cancer diagnosis: a comprehensive review of various imaging modalities," *Polish Journal of Radiology*, vol. 89, pp. 30–48, Jan. 2024, https://doi.org/10.5114/pjr.2024.134817.

[5] J. A. Fadhil and A. M. Abdulazeez, "Classification of Cancer Microarray Data Based on Deep Learning: A Review," *The Indonesian Journal of Computer Science*, vol. 13, no. 1, pp. 55–77, Feb. 2024, https://doi.org/10.33022/ijcs.v13i1.3711.

[6] X. Wang, M. Yang, J. Zhu, Y. Zhou, and G. Li, "Role of exosomal non-coding RNAs in ovarian cancer (Review)," *International Journal of Molecular Medicine*, vol. 54, no. 4, Oct. 2024, Art. no. 87, https://doi.org/10.3892/ijmm.2024.5411.

[7] S. A. Alanazi, N. Alshammari, M. Alruwaili, K. Junaid, M. R. Abid, and F. Ahmad, "Integrative analysis of RNA expression data unveils distinct cancer types through machine learning techniques," *Saudi Journal of Biological Sciences*, vol. 31, no. 3, Mar. 2024, Art. no. 103918, https://doi.org/10.1016/j.sjbs.2023.103918.

[8] K. Abdelnaeem, R. M. Dawood, B. E. Fotouh, A. Ismail, M. S. Abdalla, and S. S. Ramadan, "Expression profiling of KRAS and NOXA genes as prospective biomarkers in ovarian carcinoma," *Scientific Reports*, vol. 15, no. 1, Sept. 2025, Art. no. 32370, https://doi.org/10.1038/s41598-025-17650-6.

[9] M. Sokouti and B. Sokouti, "Cancer genetics and deep learning applications for diagnosis, prognosis, and categorization," *Journal of Biological Methods*, vol. 11, no. 3, Aug. 2024, Art. no. e99010017, https://doi.org/10.14440/jbm.2024.0016.

[10] H. Ashayeri, N. Sobhi, P. Pławiak, S. Pedrammehr, R. Alizadehsani, and A. Jafarizadeh, "Transfer Learning in Cancer Genetics, Mutation Detection, Gene Expression Analysis, and Syndrome Recognition," *Cancers*, vol. 16, no. 11, June 2024, Art. no. 2138, https://doi.org/10.3390/cancers16112138.

[11] Y. Sekar, D. Ishwar, B. Tan, and K. Venkatakrishnan, "Nano biosensor unlocks tumor derived immune signals for the early detection of ovarian cancer," *Biosensors and Bioelectronics*, vol. 278, June 2025, Art. no. 117368, https://doi.org/10.1016/j.bios.2025.117368.

[12] A. Tokareva *et al.*, "Machine Learning Framework for Ovarian Cancer Diagnostics Using Plasma Lipidomics and Metabolomics," *International Journal of Molecular Sciences*, vol. 26, no. 14, Jan. 2025, Art. no. 6630, https://doi.org/10.3390/ijms26146630.

[13] J. Cai, Z.-J. Lee, Z. Lin, C.-H. Hsu, and Y. Lin, "An Integrated Algorithm with Feature Selection, Data Augmentation, and XGBoost for Ovarian Cancer," *Mathematics*, vol. 12, no. 24, Dec. 2024, Art. no. 4041, https://doi.org/10.3390/math12244041.

[14] A. Kodipalli, V. S. Devi, S. Guruvare, and T. Ismail, "Explainable AI-based feature importance analysis for ovarian cancer classification with ensemble methods," *Frontiers in Public Health*, vol. 13, Mar. 2025, Art. no. 1479095, https://doi.org/10.3389/fpubh.2025.1479095.

[15] J. Cai, Z.-J. Lee, Z. Lin, and M.-R. Yang, "A Novel SHAP-GAN Network for Interpretable Ovarian Cancer Diagnosis," *Mathematics*, vol. 13, no. 5, Jan. 2025, Art. no. 882, https://doi.org/10.3390/math13050882.

[16] E. I. A. El-Latif, M. El-dosuky, A. Darwish, and A. E. Hassanien, "A deep learning approach for ovarian cancer detection and classification based on fuzzy deep learning," *Scientific Reports*, vol. 14, no. 1, Nov. 2024, Art. no. 26463, https://doi.org/10.1038/s41598-024-75830-2.

[17] M. M. H. Aziz and S. A. Mahmood, "Utilizing Machine Learning Techniques for Cancer Prediction and Classification based on Gene Expression Data," *UHD Journal of Science and Technology*, vol. 9, no. 1, pp. 135–148, June 2025, https://doi.org/10.21928/uhdjst.v9n1y2025.pp135-148.

[18] L. K. Hema *et al.*, "Region-Based Segmentation and Classification for Ovarian Cancer Detection Using Convolution Neural Network," *Contrast Media & Molecular Imaging*, vol. 2022, no. 1, Nov. 2022, Art. no. 5968939, https://doi.org/10.1155/2022/5968939.

[19] Y. An and H. Xue, "Indefinite twin support vector machine with DC functions programming," *Pattern Recognition*, vol. 121, Jan. 2022, Art. no. 108195, https://doi.org/10.1016/j.patcog.2021.108195.

[20] M. Dehghani, G. Bektemyssova, Z. Montazeri, G. Shaikemelev, O. P. Malik, and G. Dhiman, "Lyrebird Optimization Algorithm: A New Bio-Inspired Metaheuristic Algorithm for Solving Optimization Problems," *Biomimetics*, vol. 8, no. 6, Oct. 2023, Art. no. 507, https://doi.org/10.3390/biomimetics8060507.

[21] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, Jan. 2002, https://doi.org/10.1093/nar/30.1.207.