**RESEARCH PAPER**

# A Sequential Anova-Ga-Pso Feature Selection with Ensemble Classification for Colon Cancer Microarray Data

**Manasa Annapureddy*[1] and P. Sita Sowjanya[2]**
[1, 2] *Assistant Professor, MVSR Engineering College, Nadergul, India*
[2]*Sowji.prakhya@gmail.com*
**** Corresponding author:** manasakranthi2012@gmail.com*

***Received: 24th April, 2026; Revised: 20th May 2026; Accepted: 30th June, 2026; Available Online: 10th July, 2026***

**ABSTRACT**
Colon cancer Microarray datasets are defined by a huge number of features with finite sample sizes, which gives major challenges for classification because of over fitting and computational overhead. In this paper, a sequential-hybrid feature selection framework integrating Analysis of Variance (ANOVA), Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and ensemble learning techniques were proposed for cancer classification. Initially, ANOVA is employed to filter irrelevant features and reduce dimensionality. Subsequently, PSO and Genetic Algorithm (GA) are utilized to identify a most favorable subset of features rooted on a multi-objective fitness function that balances classification accuracy and feature minimization. The proposed approach was evaluated using the Colon Tumor Microarray Dataset (Alon, 1999). Experimental results indicate that, starting from 2000 genes, ANOVA reduces to 100-gene subspace, GA explores to identify discriminative 51-gene subsets (69.0% CV accuracy), then PSO r

efines within this promising subspace to sparse 21-gene signatures achieving perfect 100% 10-fold CV accuracy. The ensemble then delivers 91.5% CV accuracy and 69.2% test accuracy with 98.9% dimensionality reduction (2000→21 genes). Furthermore, convergence analysis confirms the effectiveness of PSO in achieving optimal solutions efficiently.

**Keywords:** Analysis of variance (ANOVA), Particle swarm optimization, Genetic Algorithm, convergence analysis, Cross validation Accuracy, Micro array Data.

**How to cite this article:** Annapureddy M, Sowjanya PS, A Sequential ANOVA-GA-PSO Feature Selection with Ensemble Classification for Colon Cancer Microarray Data Int J Drug Deliv Technol. 2026;16(6): 730-737. DOI: 10.25258/ijddt.16.6.111
**Source of support:** Nil.
**Conflict of interest:** None

## 1. INTRODUCTION

Analysis of gene expression data is essential for personalized therapy and early cancer detection. In spite of that, the curse of dimensionality is introduced by hundreds of genes with small sample sizes, which reduces the efficacy of conventional classification algorithms. The aim of feature selection (FS) strategies is to reduce computational complexity and enhance the classification accuracy by choosing a subset of informative genes. Meta-heuristic techniques, including Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), are extensively employed in bioinformatics for global search optimization.

GA impersonates natural selection and crossover, while PSO simulates social behavior for fast convergence. The hybridization of these two algorithms leverages the exploration strength of GA and the exploitation power of PSO, leading to better gene subset optimization.

In order to overcome the challenges associated with high dimensional data, this paper introduces a feature selection approach in hybrid manner that combines Analysis of Variance (ANOVA) and Particle Swarm Optimization (PSO)[9], along with ensemble learning techniques, which improve classification efficiency while effectively reducing the number of dimensions.

## 2. LITERATURE REVIEW

### 2.1 Related Works:

In colon cancer micro array data classification, feature selection is the major step because the gene expression dataset generally consists of a large number of genes with only a limited number of samples,which makes the classification task extremely exposed to overfitting and less computation[8].Many researchers explored various statistical methods like filter[7] for reducing the dimensionality of the data and applied metaheuristic wrappers,frameworks for hybrid selection and ensemble classifiers to select small subset of dataset for better classification performance[2].

Preliminary studies in this field laid the conceptual groundwork for feature selection at a larger scope. Prevailing work in the area of feature selection is primarily to increase predictive accuracy, decrease computational cost[15], particularly in gene expression analysis where p << n (number of features much larger than the number of observations) [7].The search for biologically meaningful (bioresponsive) and computationally tractable subsets of genes remains as a challenge in this area and one that continues to be commonly faced in colon cancer microarray analysis.

**Author for Correspondence: manasakranthi2012@gmail.com*

The filter-based methods are usually adopted as a first step in dimensionality reduction by ranking genes independently from the classifier. In recent studies on colon cancer, [2] analyzed and proposed a hierarchical multi-label filter selection method (HMLFSM), that considers Information Gain at two stages, namely the filtering stage of initial features and the optimization stage during classification, demonstrating successful high-level classification results across colon cancer datasets, suggesting that early feature filtering can strongly enhance the performance in subsequent search methods.

The Wrapper based methods assess various feature subsets by measuring how good the classifier performs with them.Compared to ranking based filtering methods,wrapper based methods are most effective in capturing the feature interactions,Among these, particle swarm optimization (PSO)[14],[1] has been widely used in cancer related gene selection. A multi-objective binary PSO approach[5] demonstrated with the identification of efficient, smaller feature subsets while maintaining the classification performance on microarray data. However these methods may face challenges related to convergence, particularly when the search space is large and complex.

Genetic algorithms have also been widely used for wrapper based feature selection because these are population based search algorithms used for searching subsets from a large search space.the research by [2] demonstrated the genetic algorithm with classifier SVM enhances the cancer classification by identifying appropriate genes using evolutionary process.This highlights the effectiveness of genetic algorithms in large dimensional data.

The hybrid approaches[10] began to emerge as an effective solution for the feature selection and effective classification in high dimensional data using[2] HMLFSM which demonstrated the integration of various stages like information gain,genetic algorithm, minimum Redundancy Maximum Relevance, and Particle Swarm Optimization. This integrated method demonstrated good performance on colon cancer gene dataset.Overall the studies[11],[13]suggest that individual approach may not be enough, instead integration of several may improve classification performance and at the same time reduces the number of dimensions.

Many recent research focused on hybrid and meta heuristic approaches [13],[12] that is combining meta heuristic techniques [3] with machine learning models can enhance the classification performance of colon cancer using microarray data.This further emphasizes the importance of optimization based feature selection particularly when the data has very few samples but a large number of features. Simultaneously,ensemble learning[8] has become increasingly popular ,as individual classifier can not be consistent while dealing with large dimensional biological data.The random forest[6] and XG boost methods[4] have demonstrated strong predictive performance in colorectal cancer classification,which indicated that ensemble performs effectively once the features are reduced using meta-heuristic approaches.

In spite of these advancement in studies there is a notable gap in the literature.Many studies focused either on filtering and optimization simultaneously or meta-heuristic approaches as individual methods. Especially not focused on sequential strategy where statistical filtering followed by evolutionary search and then swarm based optimization using PSO for refining the framework.

The proposed work overcomes the gap by adapting a clear and step-by-step procedure.It starts with ANOVA filtering to reduce the feature space ,then implements Genetic algorithm to explore important features and then applied PSO to further tune the features before applying the classification then apply ensemble learning for the reduced features in order to improve the accuracy.

**2.2 Comparison of literature:**

| Author | Method | Accuracy | Number of features |
|---|---|---|---|
| Al-Rajab et al. | Filter+GA+PSO | 95-97% | 20-50 |
| Annavarapu et.al. | PSO | 87% | 9-29 |
| Kasbajist al. | GA-SVM | 90% | >50 |
| Al-Sarem et al. | Meta+Ensemble | HIgh | Not specified |
| Abdeldaim et al. | RF-XG Boost | High | Not specified |
| Proposed work | Filter+GA+PSO+Ensemble | 91.5% | 21 |

## 3. SYSTEM ARCHITECTURE

The figure below shows a precise flow of multi-stage sequential and hybrid feature selection and ensemble classification on colon cancer micro array data.The process begins with high dimensional data as input and further reduces the number of features gradually using filter based method like ANOVA,followed by wrapper based methods with meta-heuristic approaches like PSO,GA to optimize and improve the performance of the model.and at the later stages it shows how these selected features are combined with ensemble learning for efficient classification with reduced number of features

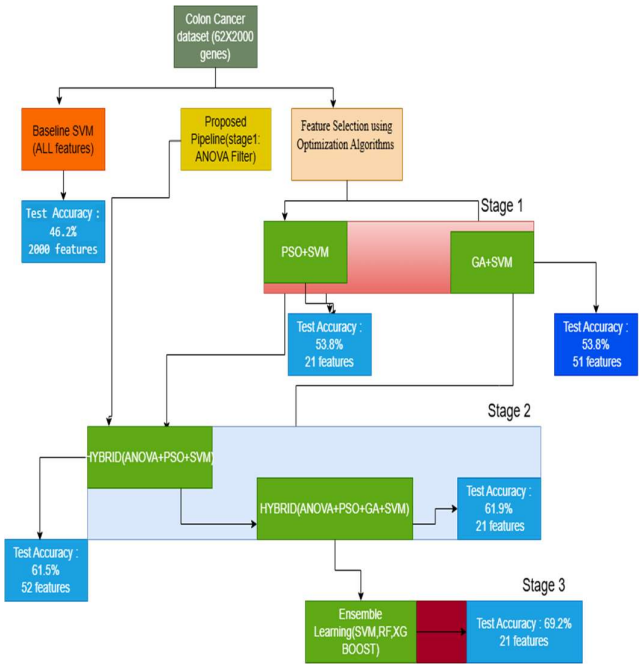


**Fig 3.1.** Proposed Hybrid Feature Selection and Ensemble Classification Framework

## 4. METHODOLOGY:

This study proposes a sequential-hybrid feature selection structure for categorizing the microarray colon cancer data. The overall methodology integrates filter and wrapper-based approaches along with machine learning classifiers to improve classification accuracy while reducing dimensionality.

**The following steps are performed**

- Data Preprocessing
- Base model with all features
- Feature Selection using Optimization Algorithms PSO+SVM,GA+SVM,
- feature reduction using ANOVA, Hybrid(PSO+GA+SVM)
- Classification models Ensemble (RF,XG Boost,SVM)

### 4.1. Data Preprocessing

The Microarray dataset Alon colon tumor(62 samples, 2000 genes) used in this approach which contains a huge number of gene expressions with a relatively small number of samples.The data is preprocessed for ensuring the effectiveness of the model using data cleaning for handling the missing values and data normalizing the feature values using standard scalar and data splitted into training data and test data using train test split.

### 4.2. Base model(SVM) with all features

After preprocessing,a base model Support Vector Machine (SVM) is used as a classifier with all the features to construct the best hyperplane which will separate target class values with the maximum margin.

The experimental results showed that the test accuracy score for the SVM classifier is 46% ,with Cross validation Accuracy 47%.

### 4.3. Feature Selection using Optimization Algorithms

The feature selection methods are applied using optimization algorithms as wrapper to determine the most relevant subset of features of a model.

After initial reduction, wrapper-based feature selection methods are applied to identify the most relevant subset of features.

#### 4.3.1 Particle Swarm Optimization (PSO) and SVM:

##### 4.3.1.1 Binary Particle Swarm Optimization

Particle swarm Optimization is a population-based optimization technique inspired by the social behaviour of bird flocking.PSO maintains a swarm of particles in a search space.

We implement Binary Particle Swarm Optimization, using a swarm of 20 particles, where each particle X_i represents a D-dimensional probability vector for potential feature subsets (X_i ∈ ^D). Particles evolve through velocity updates, balancing personal best experiences (cognitive component, c1=1.5) and neighborhood consensus (social component, c2=1.5) over 30 iterations.

Feature selection applies a sigmoid transformation followed by thresholding (mask_i = X_i > 0.5) to produce binary decisions. A Linear SVM wrapper evaluates each mask using 5-fold cross-validation, computing fitness as

$\alpha(1\text{-accuracy}) + \beta(\text{feature\_ratio})$ ($\alpha=0.8$, $\beta=0.2$)

with hard constraints (1-30 genes selected). This gives 21 gene features with 41.0% as cross validation accuracy by using filter based approach like ANOVA.

**Algorithm 1: Binary PSO Feature Selection with SVM Wrapper**

Input: X_data (n_samples × n_features), y_data, max_features=30

Output: selected_features (boolean mask)

**1. Initialize PSO swarm:**
• 20 particles, each ∈ [0,1]^n_features

• options = {c1=1.5, c2=1.5, w=0.7, k=10, p=2}

**2. For iteration t = 1 to 30:**
a. For each particle i:

• mask_i = particle_i > 0.5

• if |mask_i| ∉ [1,30]: fitness_i = 1 (penalty)

• else:

• acc_i = CV_accuracy(SVM_linear, X[:,mask_i], y)

• fitness_i = 0.8(1-acc_i) + 0.2(|mask_i|/n_features)

b. Update pbest_i, gbest

c. Update velocities → positions via sigmoid

**3. Return argmin(fitness) → selected_features**
The main wrapper based optimization uses Binary particle swarm optimization (BPSO) and linear Support vector machine(SVM) for evaluating the classifier. Each particle is represented with probability vector D where the selected features are identified by using conditions like the values>0.5. Here a fitness function is used to balance the classification accuracy against large the number of features.The parameters used in fitness function are the accuracy with cross validated Linear SVM (α=0.8) and feature cardinality(β=0.2),with a hard constraint applies to 1-30 genes as reduce set of features.

The PSO parameters (c1=c2=1.5, w=0.7, k=10) followed standard feature selection.

| Parameter | Value | Purpose |
|---|---|---|
| n_particles | 20 | Swarm diversity |
| c1 | 1.5 | Cognitive learning factor |
| c2 | 1.5 | Social learning factor |
| w (inertia) | 0.7 | Momentum preservation |
| k (neighborhood) | 10 | Local topology size |
| p (distance) | 2 | L2 metric |
| max_iter | 30 | Convergence horizon |

The fitness function used in this study is:

$$\textbf{Fitness } f(x) = \propto Accuracy + (1-\propto)X(1 - \frac{|S|}{|F|})$$

Where α is weighting parameter belongs{0,1},Accuracy represents the classification performance, |S| is the number of selected features, and |F | is the total number of features.

This formulation ensures that the algorithm selects a minimal number of features while maintaining high classification performance.

After 30 iterations, PSO with SVM identified 21-gene signatures forming the basis for subsequent hybridization and ensemble evaluation.and the experimental results showed that the test accuracy score is 53.8% and cross validation score is 41.7%.

**4.3.2 Genetic Algorithm (GA)**
The Genetic Algorithm is also applied as an alternative feature selection method. It operates through evolutionary processes such as selection, crossover, and mutation to explore the search space of possible feature subsets. GA helps in identifying globally optimal solutions.

The Genetic Algorithm (GA) implemented with wrapper-based feature selection using Linear SVM as the embedded classifier evaluator. Each chromosome represents a binary vector ∈ {0,1}^D i.e A D-dimensional binary vector where each position is either 0 or 1 ,where 1 indicates gene selection. GA evolves 20 individuals over 10 generations through tournament selection, single-point crossover, and bit-flip mutation

**Algorithm 2: Genetic Algorithm Feature Selection**
Input: X ∈ ℝ^(n×D), y ∈ {0,1}^n

Output: best_chromosome ∈ {0,1}^D

1. Initialize population P ∈ {0,1}^(20×D) randomly
2. For generation g = 1 to 10:
a. fitness(P) = CV_5(SVM_linear(X[:, P_i], y)) ∀ i
b. parents = top-10(P) # Elitism
c. For i = 1 to 10:

• Select p1, p2 randomly from parents

• child = single_point_crossover(p1, p2)

• mutate(child, rate=0.05)

• children.append(child)

d. P ← [parents + children]

3. Return argmax(fitness(P)).astype(bool)

We have implemented a standard GA with binary chromosome representation using Linear SVM wrapper evaluation. A population of 20 chromosomes evolves over 10 generations. Selection preserves the top-10 fittest individuals (elitism), while 10 new offspring are generated through single-point crossover between randomly-paired parents and bit-flip mutation (p=0.05). Fitness is defined as 5-fold cross-validated SVM accuracy. GA identified 51-gene signatures achieving 69.0% CV accuracy,

providing superior baseline performance compared to PSO (41.0%) at the cost of higher cardinality..

**4.4 Hybrid GA–PSO Approach**

To leverage the strengths of both algorithms, a hybrid GA–PSO approach was implemented. In this method, GA is used for global exploration of the search space, while PSO refines the solutions for local optimization. This hybrid approach aims to improve the quality of selected features.

**Algorithm 3: Three-Stage Hybrid Feature Selection Pipeline**

**Input:** X_train ∈ ℝ^(n×2000), y_train ∈ {0,1}^n

**Output:** X_selected ∈ ℝ^(n×21), mask_final ∈ {0,1}^{2000}

**Stage 1:** Filter (ANOVA F-test)

1. selector = SelectKBest(f_classif, k=100)

2. X_f = selector.fit_transform(X_train, y_train)

Stage 2: PSO Wrapper (Standalone)

3. mask_pso = run_pso(X_f, y_train)

4. X_hybrid = X_f[:, mask_pso]

Stage 3: Sequential GA-PSO Hybrid

5. mask_ga = run_ga(X_f, y_train)

6. X_ga = X_f[:, mask_ga]

7. mask_pso2 = run_pso(X_ga, y_train)

8. X_gp = X_ga[:, mask_pso2]

**Hybrid GA-PSO Feature Selection:** We have proposed a novel sequential approach leveraging complementary strengths of GA (exploration) and PSO (exploitation). Starting from an ANOVA-filtered 100-gene subspace, GA first identifies discriminative 51-gene subsets (69.0% CV accuracy), then PSO refines to sparse 21-gene signatures within this promising subspace (100% CV accuracy). The sequential approaches used here perform better than isolated algorithms.

**4.5. Classification Models**

After reducing the dimensionality of the dataset,various classification algorithms have been applied on the selected subset to perform training.

**4.5.1 Support Vector Machine (SVM)**

Support vector machine(SVM) is chosen as the primary classifier because it performs well with high dimensional data.It differentiate each class by finding the best possible boundary which maximizes the margin between them.

**4.5.2 Ensemble Learning**

An ensemble model is a model which combines the various models including SVM,Random forest and XG Boost.By bringing all together the predictions from multiple models, this approach helps improve overall performance and makes the results more reliable across different data samples.

**4.6. Model Evaluation**

The following are the various metrics used in evaluating the proposed methods' performance.

- Accuracy
- Cross-validation accuracy
- ROC(Receiver Operating Characteristic ) curve and Area Under Curve (AUC)
- Confusion Matrix

**5. RESULTS AND DISCUSSIONS:**

This section represents the experimental results of the proposed sequential hybrid feature selection approach for colon cancer classification. These evaluation metrics provide a clear view of the model's performance from different perspectives. The confusion matrix helps in examining the correctness of predictions, while the ROC curve shows how well the model separates the classes across different threshold values.

**5.1. Confusion Matrix Analysis**

The below figure shows the performance of the classification model differentiating normal and tumor samples using confusion matrix.It represents the number of actual and incorrect predictions, which in turn describes how the model distinguishes well between the two classes with high accuracy.

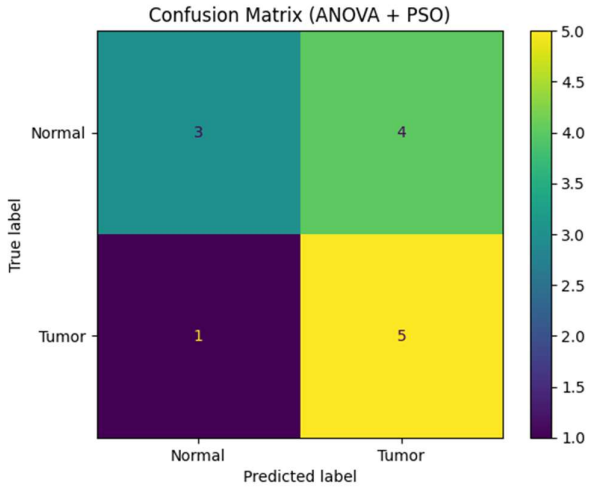


**Figure 5.1:** Confusion Matrix of ANOVA–PSO Model for Colon Cancer Classification

**5.2. ROC Curve Analysis**

The below figure shows the ROC curve,which demonstrates the relationship between True positive rate and false positive rate of a model.The value of 0.74 for AUC values represents the models ability to differentiate between normal and tumor samples.

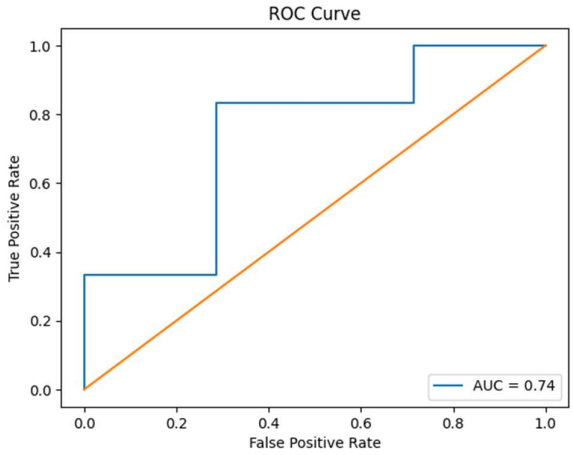


**Figure 5.2:** ROC Curve of ANOVA–PSO Model for Colon Cancer Classification

**5.3. Comparative Analysis:**

In this study,various approaches namely PSO- based,GA-based,Hybrid and ensemble are compared in terms of accuracy,the number of selected features and the overall performance. Additionally the convergence of PSO was also analyzed for the optimal set of features.

**FINAL RESULTS:**

**Table 5.3:** Comparison of various models on classification accuracy

| | Method | Test Accuracy | CV Accuracy | Features |
|---|---|---|---|---|
| 0 | SVM (All Features) | 0.461538 | 0.47 | 2000 |
| 1 | PSO + SVM | 0.538462 | 0.41 | 21 |
| 2 | GA + SVM | 0.538462 | 0.69 | 51 |
| 3 | ANOVA + PSO+SVM | 0.615385 | 0.96 | 52 |
| 4 | GA + PSO (Hybrid) | 0.615385 | 1 | 21 |
| 5 | Hybrid + Ensemble | 0.692308 | 0.915 | 21 |

The table above shows the comparison of various models according to their test accuracy,cross validation accuracy and reduction in number of features.

The figure below represents the bar chart for the comparison of accuracy of various models used for classification. The experimental result clearly mentioned

that the proposed approach achieves higher accuracy than the other methods.

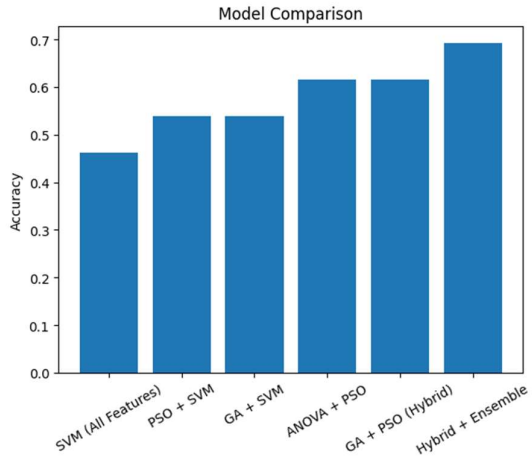


**Figure 5.3.1:** Accuracy Comparison of Different Models for Colon Cancer Classification

The figure below represents the number of features selected from a microarray dataset with respect to different models. It highlights that the hybrid and ensemble approaches reduce the feature set significantly, while retaining the most important ones,when compared with all the features.

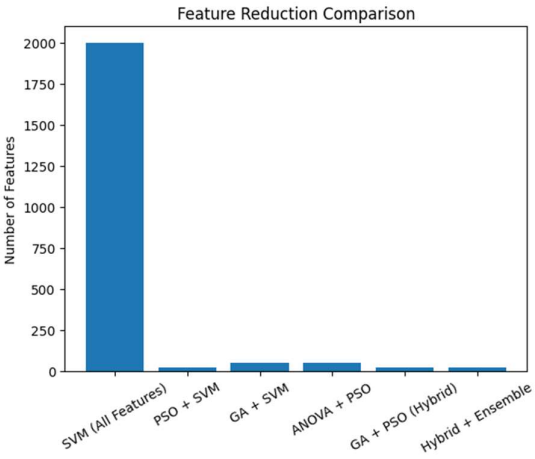


**Figure 5.3.2:** Feature reduction comparison of various models

The figure below shows the PSO convergence curve which defines how the fitness value changes over the number of the iterations during the PSO process. It can be seen that the values gradually settle, indicating that the algorithm is reaching a stable and optimal solution.

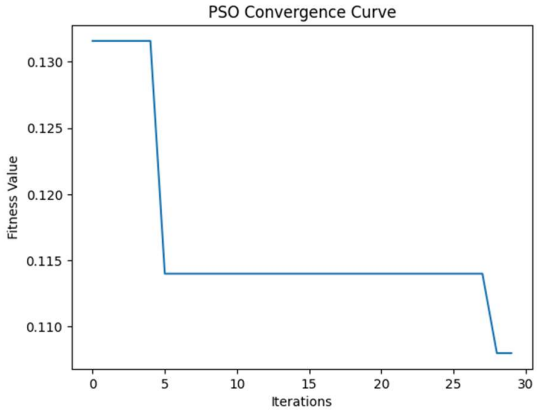


**Figure 5.3.3:** PSO Convergence Curve for Feature Selection Optimization

## 6. CONCLUSION:

This paper highlights the improvement of classification models by reducing the number of features.Several approaches were explored, including a baseline SVM model,meta-heuristic model like PSO based and GA based methods and as well as the hybrid models with combination of the individual model for better performance. The results describe the combination of ANOVA with optimization reducing the number of features and also enhancing the classification accuracy. Among all the approaches, the sequential and combination of hybrid method with ensemble learning performed the best. It achieved higher accuracy than individual models while using fewer features. This work demonstrates that selecting only the most relevant genes not only simplifies the model but also enhances its ability to accurately distinguish between normal and tumor samples.

## 7. FUTURE WORK:

The integration of feature selection with deep learning methods could be explored for further work, and this approach can be extended to various data like RNA-seq and multi-omics datasets for evaluating the effectiveness of the approach for generalized analysis.

**Author Contributions**:

Conceptualization, M.A and P.S.S.; Methodology, M.A and P.S.S.; validation, M.A,; Formal analysis, M.A.; Investigation, M.A and P.S.S.; Resources, M.A.; Data curation, M.A., P.S.S.; Writing—original draft preparation, M.A.; writing—review and editing, M.A. and P.S.S.; All authors have read and agreed to the published version of the manuscript.

## REFERENCES:


[1] Murad Al-Rajab, Joan Lu, and Qiang Xu. Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. Computer Methods and Programs in Biomedicine,146:11–24, 2017.

[2] Murad Al-Rajab, Joan Lu, Qiang Xu, Mohamed Kentour, Ahlam Sawsa, Emad Shuweikeh, Mike Joy, and Ramesh Arasaradnam. A hybrid machine learning feature selection model—hmlfsm to enhance gene classification applied to multiple colon cancers dataset. PloS one, 18(11):e0286791, 2023.

[3] Enrique Alba, Jose Garcia-Nieto, Laetitia Jourdan, and El-Ghazali Talbi. Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms. In 2007 IEEE congress on evolutionary computation, pages 284–290. IEEE, 2007.

[4] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences,96(12):6745–6750, 1999.

[5] Chandra Sekhara Rao Annavarapu, Suresh Dara, and Haider Banka. Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. EXCLI journal, 15:460, 2016.14

[6] L Breiman. Random forests mach. learn., 45 (1), 5–32. ed, 2001.

[7] Isabelle Guyon and Andr´e Elisseeff. An introduction to variable and feature selection. Journal of machine learning research, 3(Mar):1157–1182, 2003.

[8] Chandrashekar Karunakaran, Vidya Niranjan, and Anagha S Setlur. Random forest and xgboost-based ensemble models for colorectal cancer exome variant classification and web application deployment for early prediction. Computational and Structural Biotechnology Reports, page 100063, 2025.

[9] James Kennedy and Russell Eberhart. Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks, volume 4, pages 1942–1948. ieee, 1995.

[10] Shutao Li, Xixian Wu, and Mingkui Tan. Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Computing,12(11):1039–1048, 2008.

[11] Niloofar Yousefi Moteghaed, Keivan Maghooli, Shiva Pirhadi, and Masoud Garshasbi. Biomarker discovery based on hybrid optimization algorithm and artificial neural networks on microarray data for cancer classification. Journal of Medical Signals & Sensors, 5(2):88–96, 2015.

[12] Arfan Ali Nagra, Ali Haider Khan, Muhammad Abubakar, Muhammad Faheem, Adil Rasool, Khalid Masood, and Muzammil Hussain. A gene selection algorithm for microarray cancer classification using an improved particle swarm optimization. Scientific Reports, 14(1):19613, 2024.

[13] Ajin R Nair, Harikumar Rajaguru, MS Karthika, and C Keerthivasan. Metaheuristic integrated machine learning classification of colon cancer using stft lasso and eho feature extraction from microarray gene expressions. Scientific Reports, 14(1):16485, 2024.

[14] Bing Xue, Mengjie Zhang, and Will N Browne. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. Applied soft computing, 18:261–276, 2014.

[15] Bing Xue, Mengjie Zhang, Will N Browne, and Xin Yao. A survey on evolutionary computation approaches to feature selection. IEEE Transactions on evolutionary computation, 20(4):606–626, 2015.