Title: Enhancing Classification Soft Computing using GA and PSO

To: Dr. Nashwa Elbendary

Members:

- **Mahmoud AbdEl-Baset** 211008597
- **Ahmed Abdulrazak** 211013224
- **Ahmed Saad** 211013060
- **Khaled Abdelazeem** 211013156

Table of Contents

1. Abstract

This Documentation presents the development of models for the efficiency of soft computing techniques, specifically **Genetic Algorithm (GA)** and **Particle Swarm Optimization (PSO)**, in improving the performance of machine learning models on high-dimensional chemical datasets. Two classification models were selected: **Random Forest (RF)** and **Support Vector Machine (SVM)**. GA and PSO algorithms were applied for optimal feature selection among 168 descriptors.

The "Musk Version 2" dataset from the UCI repository, which contains numerical representations of molecules (classified as either Musk or Non-Musk), was utilized. The study demonstrated the significant impact of optimized feature selection on improving model accuracy and efficiency, showcasing the role of soft computing as a powerful tool in complex classification problems.

2. Introduction and Objective

In machine learning systems, high-dimensional data poses a major challenge, especially when the dataset contains redundant or irrelevant features. This leads to overfitting, increased training time, and reduced model accuracy. Feature selection techniques aim to reduce the number of features while retaining meaningful information.

This study aims to:

- Analyze the performance of RF and SVM models on the Musk dataset.
- Apply GA and PSO to select the most influential features.
- Compare the performance before and after feature optimization using multiple evaluation metrics.
- Identify the optimal model in terms of accuracy and efficiency.

3. Dataset

Dataset Name: Musk Version 2

Source: UCI Machine Learning Repository

Number of Instances: 6598

Number of Features: 168 molecular descriptors

Classification Type: Binary (1 = Musk, 0 = Non-Musk)

Preprocessing Steps:

- Data loading using pandas
- Feature scaling using MinMaxScaler
- Splitting into training (75%) and testing (25%) sets using train_test_split

This dataset reflects complex molecular properties, making it ideal for testing feature selection algorithms.

4. Methodology

4.1 Base Models

- **Random Forest (RF):** An ensemble method based on multiple decision trees.
- **Support Vector Machine (SVM):** A linear classifier using maximum-margin separation.

4.2 Optimization Algorithms

Genetic Algorithm (GA):

- Each feature subset is encoded as a binary chromosome.
- Evolutionary operations include selection, crossover, and mutation.
- Model performance is used as the fitness function.

Particle Swarm Optimization (PSO):

- Each particle represents a feature subset.
- Position updates are guided by personal and global best solutions.
- Goal: maximize model accuracy via intelligent updates.

4.3 Evaluation Metrics

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **Confusion Matrix**
- **Classification Report**

5. Models Architecture & Flow Diagram

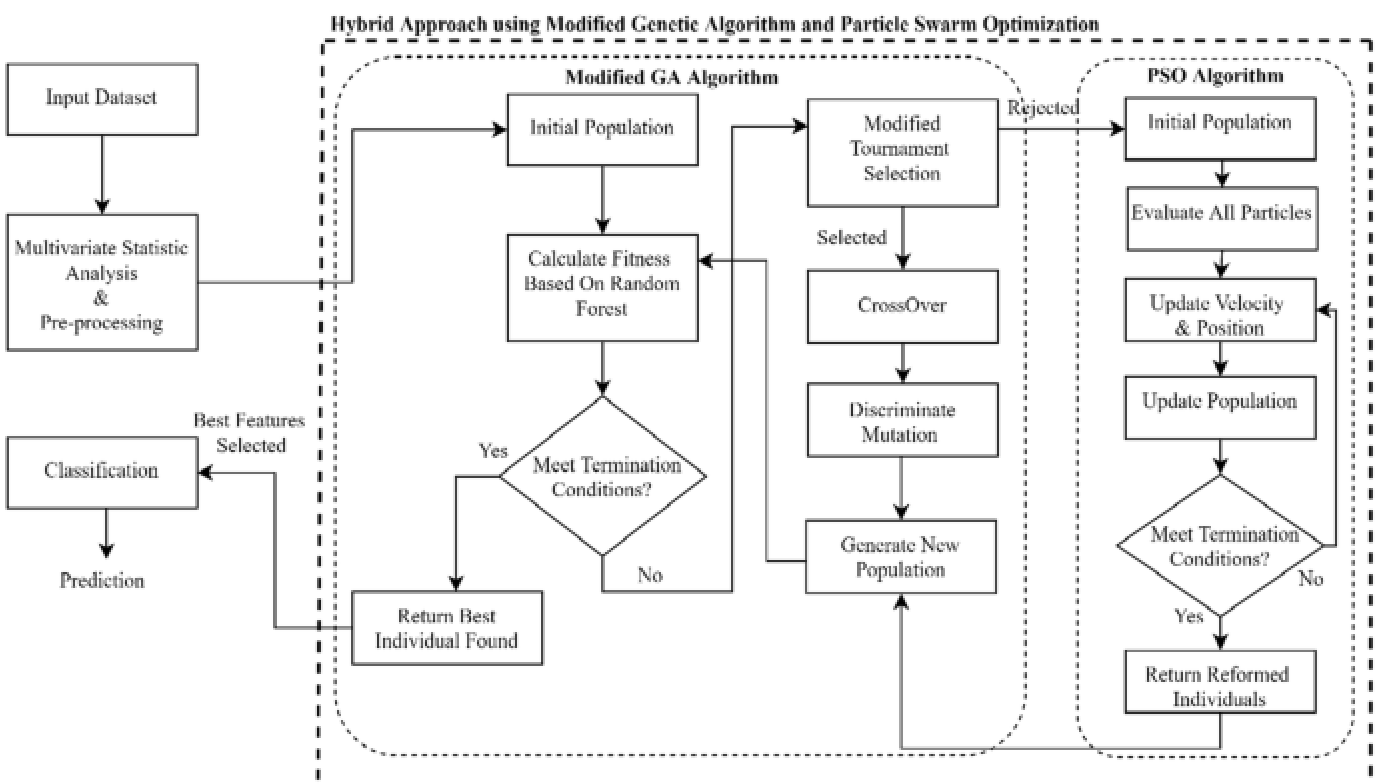


Figure 1 : Flow Chart for modified GA & PSO [2]

6. Result

6.1 Confusion Matrix Analysis

Confusion matrices were generated for each model to visualize correct and incorrect predictions. The optimized models reduced false positives and false negatives, indicating improved distinction between Musk and Non-Musk classes.

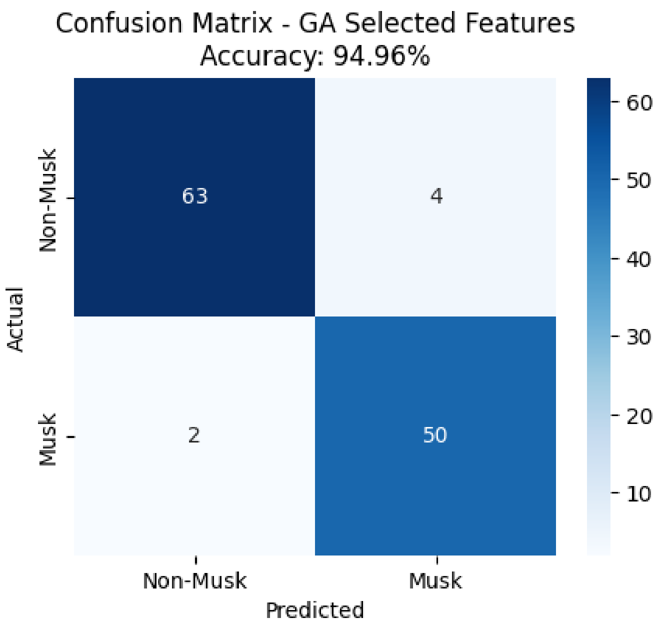


Figure 2 : Confusion matrix for GA Selected Features

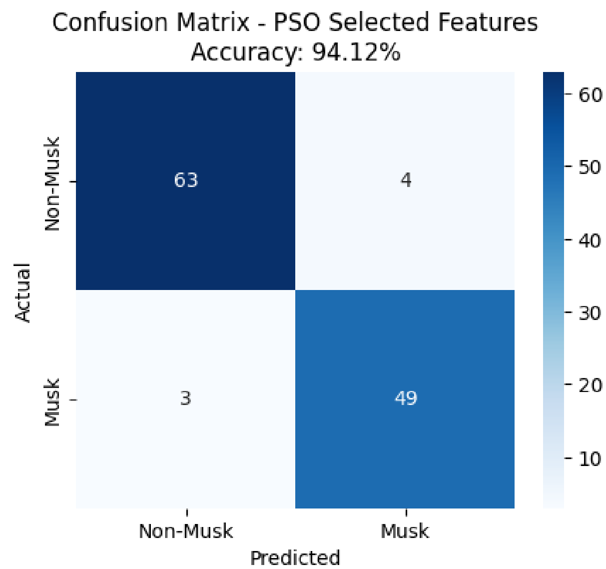
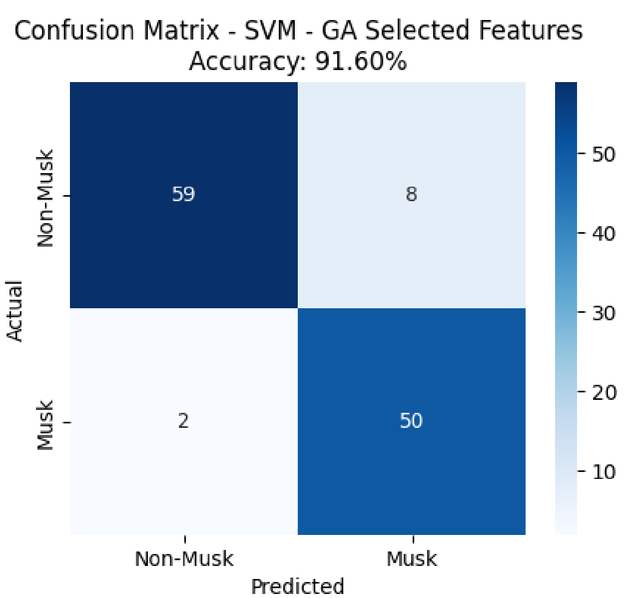


Figure 3 : Confusion matrix for PSO Selected Features

- **SVM with GA &PSO Optimization :**



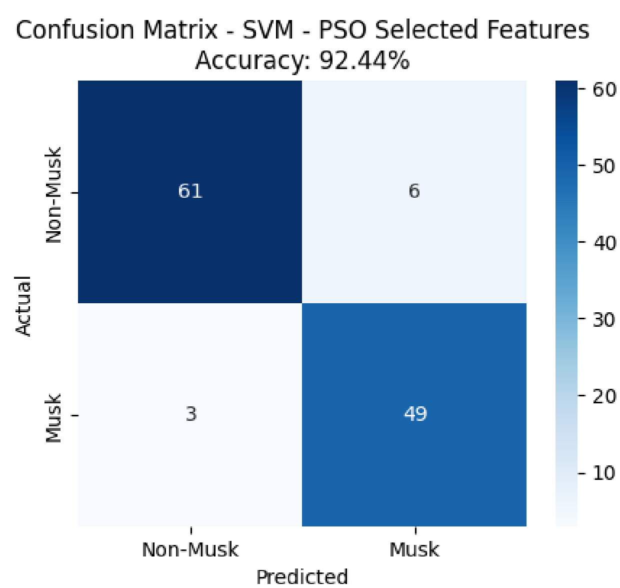


Figure 4 : Confusion matrix for SVM- GA Selected Features

Figure 5 : Confusion matrix for SVM- PSO Selected Features

6.2 Classification Report

The classification reports display precision, recall, and F1-score per class. Optimized models using GA and PSO showed clear improvement, especially in identifying the "Musk" class.

```
--- Evaluation for GA Selected Features ---
Accuracy : 0.9496
Precision: 0.9259
Recall   : 0.9615
F1-score : 0.9434

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.94      0.95        67
           1       0.93      0.96      0.94        52

    accuracy                           0.95       119
   macro avg       0.95      0.95      0.95       119
weighted avg       0.95      0.95      0.95       119
```

Figure 6 : Classification Report for GA Optimization

```
--- Evaluation for PSO Selected Features ---
Accuracy : 0.9412
Precision: 0.9245
Recall   : 0.9423
F1-score : 0.9333

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.94      0.95        67
           1       0.92      0.94      0.93        52

    accuracy                           0.94       119
   macro avg       0.94      0.94      0.94       119
weighted avg       0.94      0.94      0.94       119
```

Figure 7 : Classification Report for PSO Optimization

- **Additional Results : using SVM with GA & PSO Optimization**

```
--- Model Evaluation: SVM - GA Selected Features ---
Accuracy : 0.9160
Precision: 0.8621
Recall   : 0.9615
F1-score : 0.9091

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.88      0.92        67
           1       0.86      0.96      0.91        52

    accuracy                           0.92       119
   macro avg       0.91      0.92      0.92       119
weighted avg       0.92      0.92      0.92       119
```

Figure 8 : Classification Report & Model Evaluation

for SVM - GA Optimization

```
--- Model Evaluation: SVM - PSO Selected Features ---
Accuracy : 0.9244
Precision: 0.8909
Recall   : 0.9423
F1-score : 0.9159

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.91      0.93        67
           1       0.89      0.94      0.92        52

    accuracy                           0.92       119
   macro avg       0.92      0.93      0.92       119
weighted avg       0.93      0.92      0.92       119
```

Figure 9 : Classification Report & Model Evaluation

for SVM - PSO Optimization

Visualizations:

- The classification report confirms robust performance of the SVM model after applying GA.
- Improved precision and recall balance make it highly suitable for complex classification problems.

6.3 Evaluation Metric Comparison

A bar chart was created to compare accuracy, precision, recall, and F1-score across models. The best performance was achieved by the **RF model optimized with GA**, followed by **PSO**.

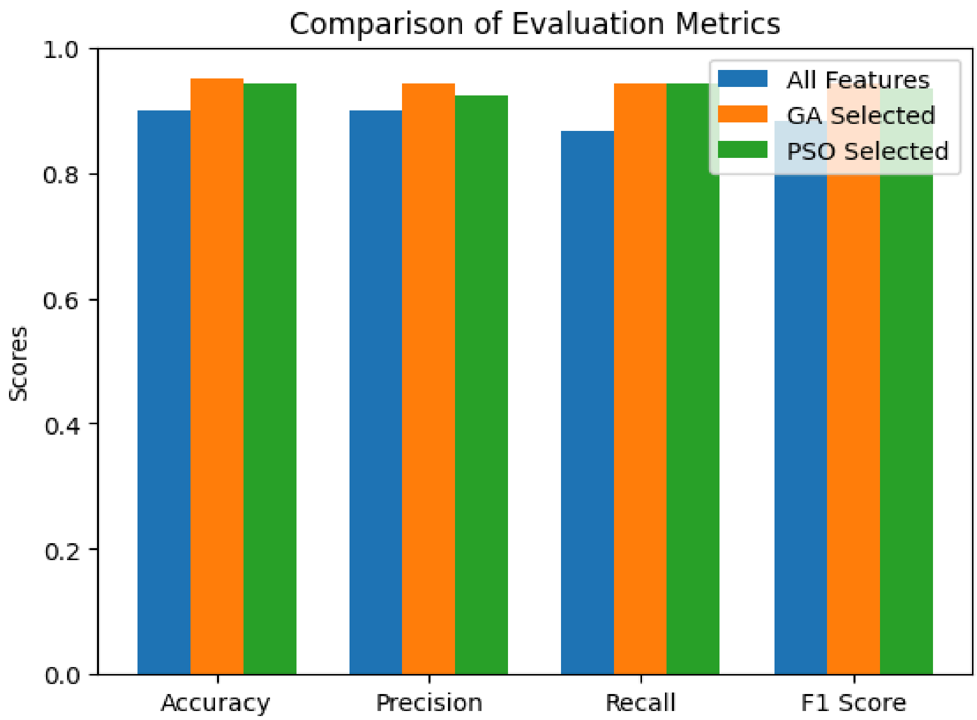


Figure 10 : Comparison of Evaluation Metrics

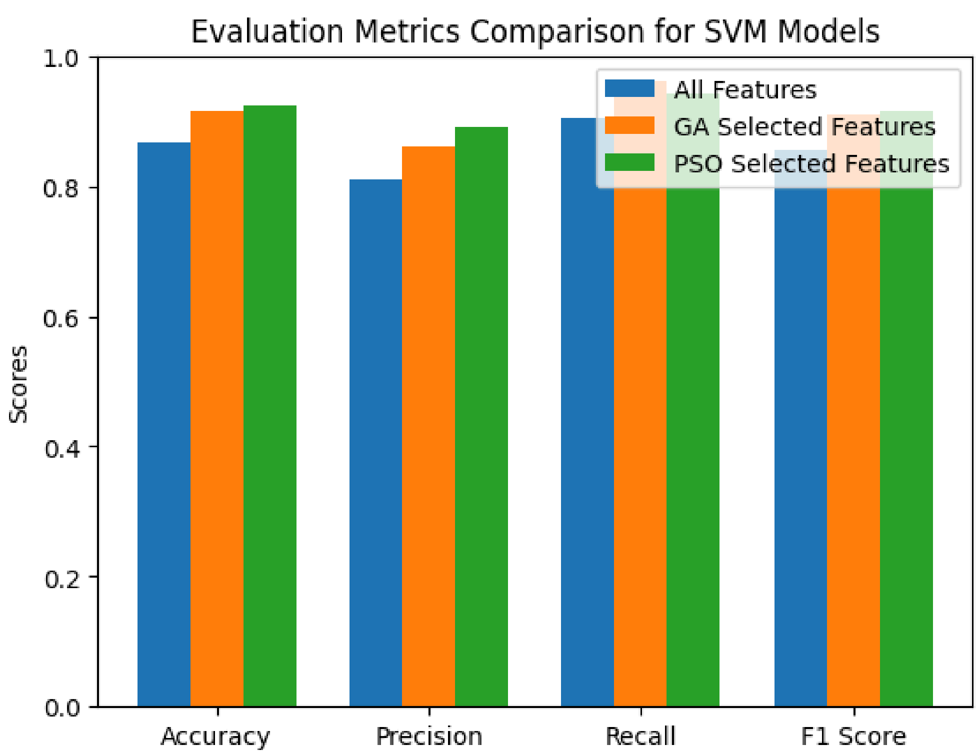


Figure 11 : Evaluation Metrics Comparison for SVM Models

7. Conclusion :

This study demonstrates that integrating soft computing techniques with machine learning models leads to significant performance gains, particularly with high-dimensional datasets like molecular data. Both GA and PSO contributed to reducing the feature space while enhancing classification accuracy.

The **Random Forest model enhanced by GA** proved to be the most effective, followed closely by the **SVM model with PSO**.

These findings confirm the value of intelligent feature selection algorithms in building more accurate and efficient classifiers.

8. References :

[1] Goldberg, David E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989. 15/5/2025 2:00pm

[2] T. Alam, S. Qamar, A. Dixit, and M. Benaida, "Genetic Algorithm: Reviews, Implementations, and Applications," *arXiv preprint arXiv:2007.12673*, 2020. [Online]. Available: https://arxiv.org/abs/2007.12673 15/5/2025 2:00pm

[3]https://archive.ics.uci.edu/dataset/75/musk+version+2 16/5/2025 4:00pm

[4]https://www.researchgate.net/publication/262242670_Enhancing_performance_of_particle_swarm_optimization_through_an_algorithmic_link_with_genetic_algorithms

[5]https://github.com/ma7moud-3baset/Soft-Computing-Project.git 17/5/2025 1:00pm