A Novel Two-Stage Feature Selection Method Based on Random Forest and Improved Genetic Algorithm

1. Paper Source & Link

- **Journal:** Scientific Reports (Nature Portfolio)
- **Year:** 2025
- **Volume:** 15, Article number: 16828
- **DOI:** https://doi.org/10.1038/s41598-025-01761-1
- **Type:** Research Article

2. Core Problem & Objective

- **Problem:** Single feature selection methods have limitations (incomplete, unstable, time-consuming)
- **Objective:** Develop a hybrid two-stage method that combines strengths of different approaches
- **Key Goal:** Enhance classification accuracy while minimizing feature subset size

3. Proposed Method: RFIGA (Random Forest + Improved Genetic Algorithm)

Stage 1: Random Forest Filtering

- **Purpose:** Quick elimination of irrelevant features
- **Method:** Calculate Variable Importance Measure (VIM) scores using Gini coefficient
- **Process:**
    - Train Random Forest on full dataset
    - Calculate VIM scores for all features
    - Rank features by importance
    - Eliminate low-importance features (reduce search space)

Stage 2: Improved Genetic Algorithm (IGA)

- **Purpose:** Global search for optimal feature subset
- **Key Improvements:**
  - **Multi-objective Fitness Function:**
- **Adaptive Mechanism:**
  - Dynamic adjustment of crossover (P_c) and mutation (P_m) probabilities
  - Based on population fitness diversity
- **μ+λ Evolution Strategy:**
  - Maintains best μ individuals each generation
  - Prevents premature convergence

4. Experimental Setup

Datasets (8 UCI datasets):

- **Size range:** 23 to 1203 features
- **Sample range:** 90 to 839 instances
- **Classes:** 2 to 24 classes
- **Examples:** Glioma, Dermatology, Arrhythmia, Toxicity

Evaluation Metrics:

1. **Ratio:** Feature reduction percentage
2. **Accuracy (ACC):** Classification performance
3. **Standard Deviation:** Stability across 10 runs

Compared Methods:

- **Traditional:** Fisher Score (FS), RFE, L_2 penalty
- **Swarm Intelligence:** WOA, PSO, DE, GWO
- **Multi-stage methods:** 4 recent hybrid approaches

A Novel Two-Stage Feature Selection Method Based on Random Forest and Improved Genetic Algorithm

5. Key Results

A. Ablation Study (Table 3):

Dataset	Best Method	Accuracy Improvement	Feature Reduction
Dermatology	RFIGA	+2.18%	61%
Movement Libras	RFIGA	+5.39%	51%
Period Changer	RFIGA	+21.89%	95%
Toxicity	RFIGA	+20.46%	94%

- **RFIGA achieved best accuracy on 4/8 datasets**
- **Consistently reduced features by 50-95%**

B. Comparison with Other Methods (Table 4):

- **RFIGA achieved highest accuracy on 7/8 datasets**
- **Outperformed all swarm intelligence algorithms (WOA, PSO, DE, GWO)**
- **Superior to traditional methods (FS, RFE, L_2)**

C. Multi-stage Comparison (Table 5):

- **RFIGA outperformed 4 recent multi-stage methods**
- **Best accuracy on most datasets with competitive feature reduction**

D. Convergence Analysis (Figure 5):

- **RFIGA showed fastest convergence** on all datasets
- **Achieved lowest fitness values** on 5/8 datasets
- **More stable than basic GA and other swarm algorithms**

E. Generalizability (Table 6 & Figure 6):

Tested on **5 different classifiers**:

1. **Logistic Regression** (main experiments)

2. **SVM:** Accuracy improvement: 4.12–35.09%
3. **Ridge Regression:** Accuracy improvement: 0.12–33.29%
4. **Naive Bayes:** Accuracy improvement: 3.46–62.40%
5. **KNN:** Accuracy improvement: 7.15–24.84%

All classifiers achieved >50% feature reduct

6. Method Advantages

Theoretical Advantages:

1. **Complementary strengths:**
    a. RF: Fast, robust to outliers, handles non-linear data
    b. IGA: Global search, eliminates redundancy
2. **Improved GA mechanisms:**
    a. Adaptive probabilities prevent stagnation
    b. Multi-objective fitness balances trade-offs
    c. μ+λ strategy maintains diversity
3. **Two-stage efficiency:**
    a. RF reduces search space
    b. IGA fine-tunes on reduced set

Practical Performance:

1. **High accuracy:** Best or competitive on all datasets
2. **Significant feature reduction:** 51-95% across datasets
3. **Fast convergence:** Better than other optimization algorithms
4. **Classifier-agnostic:** Works with various ML models

. Limitations & Future Work

Limitations:

- **High time complexity:** Not suitable for ultra-high-dimensional data (10,000+ features)
- **Parameter tuning required:** Need to set RF and GA parameters

A Novel Two-Stage Feature Selection Method Based on Random Forest and Improved Genetic Algorithm

Future Directions:

1. **GPU acceleration** for handling very large datasets
2. **Automated parameter optimization**
3. **Application to specific domains** (genomics, medical imaging)
4. **Extension to deep learning feature selection**

Citation Information

- **Authors:** Junyao Ding, Jianchao Du, Hejie Wang, Song Xiao
- **Title:** "A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning"
- **Journal:** Scientific Reports, 15, 16828 (2025)
- **DOI:** 10.1038/s41598-025-01761-1