
Gene Expression Pattern Classification for Cancer Subtype Prediction Using Microarray Data and Ensemble Learning

Dr. B.Lavanya and Dr.S. Rama Mercy
Assistant Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore
lavanya_cs@avinuty.ac.in

ABSTRACT

Microarray-based gene expression profiling identifies subtype-specific transcriptional signatures and supports precision oncology by representing cancer heterogeneity in high-dimensional molecular space. Dimensionality-induced overfitting, redundant gene representations, class imbalance, and poor generalisation across cancer subtypes plague conventional classification approaches. To overcome these limits, this research provides an Adaptive Gene Expression Ensemble Classification (AGEEC) framework for robust microarray cancer subtype prediction. Variance-aware gene filtering, mutual-information-driven feature prioritisation, and adaptive ensemble learning extract discriminative molecular fingerprints from high-dimensional expression profiles. To enhance subtype recognition, an optimised stacking architecture employs random forest, support vector machine, and gradient-boosting classifiers with a meta-classifier merging their probabilistic outputs. Layered hyperparameter optimisation and stratified cross-validation reduce selection bias and improve generalisation. Subtype Separation Efficiency (SSE) measures inter-subtype discrimination, Gene Signature Retention Rate (GSRR) biomarker reproducibility, and Perturbation Resilience Coefficient (PRC) prediction consistency under controlled expression-profile perturbations. Benchmark cancer microarray datasets are used to evaluate AGEEC against individual classifiers and ensemble baselines. The framework aims for an SSE of 97.8%, GSRR of 94.6%, PRC of 96.3%, and a 15.2% improvement in subtype discrimination over the baseline.

Keywords: Cancer subtype classification, microarray gene expression, ensemble learning, biomarker selection, high-dimensional data, adaptive classification, molecular profiling, prediction stability, feature consistency, precision oncology.

1. Introduction

Cancers vary in molecular changes that affect cellular proliferation, differentiation, metabolic control, and treatment response [1]. Cancer cells may share a physical tissue, but their genetic and transcriptional properties may vary, resulting in molecular subgroups with different clinical behaviours [2]. Thus, precision oncology requires accurate cancer subtype identification for molecular characterization, prognostic evaluation, and subtype-specific treatment [3]. With developments in high-throughput genomic technology, microarray-based gene expression profiling is now a standard method for evaluating thousands of genes concurrently [4]. Computational analysis of these datasets may find differentially expressed genes, biomarkers, and cancer subtypes. Machine learning (ML) methods can detect complex correlations in high-dimensional biological datasets, making them popular for gene expression classification [5]. Conventional methods such as Support Vector Machines (SVMs), decision

trees (DTs), Random Forests (RFs), and Artificial Neural Networks (ANNs) can derive prediction patterns from gene expression data. Ensemble learning combines multiple predictive models to improve classification reliability and reduce reliance on a single learner [6]. In particular, stacking-based ensembles use complementary decision boundaries by integrating heterogeneous classifiers through secondary learning [7]. When expression profiles include noisy measurements, redundant features, and large variation in cancer subtype sample sizes, subtype classification becomes difficult [8].

Despite breakthroughs in computational genomics, microarray-based cancer subtype classification faces methodological constraints [9]. The high-dimensional, low-sample-size (HDLSS) nature of microarray datasets increases the risk of overfitting, since thousands of gene expression variables may be available for a small number of patient samples [10]. Second, irrelevant, redundant, and correlated genes may complicate categorization models, obfuscate molecular patterns, and make them harder to analyze [11]. Third, class imbalance may favour dominant subtypes in classical learning techniques, causing molecular class identification errors [12]. Expression variability and measurement noise may destabilise feature rankings and predictions in models tested on limited or varied datasets [13]. Existing categorisation approaches may also fail to include diverse learners' strengths [14]. Classifiers' sensitivity to nonlinear expression linkages, feature interactions, and class boundaries varies. Traditional ensemble techniques combine numerous predictions, but rigid aggregation processes may not adjust to each model's predictive contributions [15]. Molecular subtype discrimination may be mischaracterised by conventional classification methods, especially gene signature stability and repeatability.

This study develops a robust, computationally effective categorisation method that captures complementary chemical patterns and addresses high-dimensional gene expression data issues. Gene prioritisation reduces feature redundancy while maintaining physiologically relevant signals, whereas adaptive ensemble integration uses many classifiers. These methods may improve cancer subtype identification and reduce expression profile change prediction sensitivity in a unified framework. Model dependability may be examined by gene signature repeatability and prediction resilience to controlled perturbations. AGEEC employs molecular feature prioritisation and heterogeneous ensemble learning to predict microarray-based cancer subtypes. The main contribution of this research are:

- A variance-aware gene filtering and mutual-information-based prioritisation technique eliminates extraneous characteristics and preserves differential gene expression patterns for accurate cancer subtype identification..
- A suggested architecture for Adaptive Gene Expression Ensemble Classification (AGEEC) combines optimised random forest (RF), support vector machine (SVM), and gradient-boosting classifiers. A stacking-based meta-classifier uses their probabilistic outputs to improve subtype discrimination and learning..
- A multi-criteria assessment technique, including SSE, GSRR, and PRC, evaluates subtype separability, biomarker reproducibility, and prediction stability. The work uses benchmark microarray datasets, stratified cross-validation, and nested hyperparameter optimisation to compare it against individual classifiers and standard ensemble baselines..

The rest of this paper has four sections. Section 2 reviews gene expression categorisation, feature-selection, and ensemble-learning methods. Section 3 explains the AGEEC framework, including microarray preprocessing, gene prioritisation, ensemble design, and evaluation structures. Section 4 describes the experimental setting, compares outcomes,

assesses robustness, and discusses findings. Finally, Section 5 summarises the study's main findings, admits shortcomings, and suggests further research.

2. Related Works

Alkamli et al. [16] presented the Nuclear Reaction Optimisation (NRO) method for gene selection and cancer classification utilising six benchmark microarray datasets. The technique uses nuclear fission- and fusion-inspired search methods, SVM, and k-NN and was tested using leave-one-out cross-validation. NRO obtained 99.26% accuracy on the SRBCT dataset using SVM with 74 genes and 98.81% accuracy on the lymphoma dataset (119 genes). Banjoko et al. [17] devised a multi-objective optimisation approach for gene selection and cancer classification to overcome filter-only selection and weakly informative gene retention. Their method combines noisy-gene removal with wrapper-based feature selection to optimize gene subset quality and classification performance. The authors also introduce a statistical procedure to test differences in misclassification error rates.

Alkamli et al. [18] proposed GNR (Genetic-Embedded Nuclear Reaction Optimization with F-Score Filter), integrating statistical filtering, nuclear reaction optimization, and genetic crossover to identify compact, informative gene subsets. The framework was assessed on benchmark leukemia microarray datasets using SVM and other classifiers. On the Leukemia 1 dataset, GNR achieved 98.61% average accuracy using two selected genes after filtering 500 candidates from 7,129 original genes. On Leukemia 2, it achieved 95.83% average accuracy with two selected genes. Dhamercherla et al. [19] developed an Eagle Prey Optimization (EPO)-based feature-selection method for high-dimensional microarray gene expression classification. The method uses an optimization-driven search to identify informative gene subsets and uses classification performance to assess the selected features. Its evaluation across cancer microarray datasets provides a relevant comparison for AGEEC's proposed variance-aware filtering and mutual-information-based prioritization.

Shah et al. [20] investigated optimized gene selection and deep learning-based cancer classification using microarray gene expression profiles. Their approach combines feature optimization with neural classification to capture complex relationships in gene expression data. This work provides a methodological comparison for AGEEC's heterogeneous classical ensemble, particularly in evaluating whether combining RF, SVM, and gradient boosting can offer competitive performance relative to a deep-learning-based approach. Shukla et al. [21] proposed OBKGSA (Opposition-Based Kernel Principal Component Analysis and Gravitational Search Algorithm), a hybrid framework integrating kernel principal component analysis (KPCA), opposition-based learning (OBL), and an enhanced gravitational search algorithm for gene subset optimization. Evaluated on six benchmark microarray datasets, OBKGSA achieved 98.80% accuracy using 10 genes on the SRBCT dataset and 97.89% accuracy using nine genes on the lung cancer dataset.

Song et al. [22] developed SFEMEO (Serial Filter-Wrapper Feature Selection Method with Elite-Guided Mutation and Equilibrium Optimizer) to address high dimensionality and feature redundancy in cancer gene expression classification. The framework combines seven filter methods for preliminary screening with an improved equilibrium optimizer (EO) and an elite-guided mutation strategy for wrapper-based feature selection. Across ten cancer gene expression datasets, SFEMEO improved classification accuracy by 3.73%–18.13% compared with most competing intelligent optimization algorithms. Song et al. [23] introduced PF-PSS (Double-Layer Parallel Embedded Feature Selection), a feature-selection approach combining parallel filter screening with embedded Lasso regression and improved sequential forward selection (SFS) and sequential backward selection (SBS). The method was evaluated on 20 cancer gene expression datasets and achieved 100% classification accuracy on nine datasets,

while selecting no more than 7% of the original features on 12 datasets. On dataset C6, PF-PSS achieved 95.62% accuracy, 95% precision, 100% recall, and a 95.24% F-measure.

Lin et al. [24] proposed a Multi-Task Ensemble Strategy (MTES) for simultaneous gene selection and cancer classification. The method generates multiple randomly sampled training subsets and applies multi-task logistic regression with $\ell_{2,1}$ group-sparsity regularization to identify genes that remain relevant across different training tasks. The approach was tested on simulated and real gene expression datasets, including leukemia, colon, lung, and diffuse large B-cell lymphoma profiles. The authors reported improved classification accuracy and gene-selection consistency relative to several baseline methods, including L1-penalized models, elastic net, mRMR, and ReliefF. Jayaseeli et al. [25] combined optimization-based feature selection with deep-learning-based classification to create a multimodal feature-optimized cancer classification framework for microarray gene expression studies. To handle high-dimensional expression data, the research investigates gene selection and neural feature extraction in an integrated learning architecture. Its methodology supports AGEEC's goal of merging gene prioritisation and predictive classification, especially compared to neural-network approaches. Existing approaches are listed in Table 1.

Table 1: Summary of the Existing Methods.

Ref. No.	Method	Results	Advantages	Limitations
[16]	Nuclear Reaction Optimization (NRO) with SVM and k-NN	SVM accuracy on six microarray datasets using LOOCV: Colon 82.16%, Leukemia1 95.53%, Leukemia2 92.47%, Lung 95.36%, Lymphoma 98.81%, and SRBCT 99.26%.	Optimization-based gene subset search; evaluated with multiple classifiers and benchmark datasets.	Computationally demanding search; LOOCV on small datasets may not reflect external generalization.
[17]	Multi-objective optimization for gene selection and cancer classification	Simultaneously considers gene selection and classification objectives; exact numerical results require checking the article's result tables.	Considers multiple objectives instead of optimizing classification accuracy alone.	More complex optimization and parameter tuning; numerical comparison and validation details need confirmation.
[18]	Genetic-Embedded Nuclear Reaction Optimization (GNER) with F-score filtering and SVM	Reports 98.61% average accuracy on Leukemia1 and 95.83% on Leukemia2, using two selected genes for each dataset.	Combines statistical filtering and an enhanced search method; aims to identify compact gene subsets.	Sensitive to filtering and optimization settings; requires independent validation and biological assessment of selected genes.
[19]	Eagle Prey Optimization (EPO) for gene selection	Evaluates feature selection and classification on cancer gene-expression datasets	Metaheuristic feature selection can explore candidate gene combinations and reduce dimensionality.	Search may be computationally intensive and sensitive to algorithm settings; detailed results and

				external validation need confirmation.
[20]	Optimized gene selection with deep-learning-based cancer classification	High performance metrics	Combines feature selection with a nonlinear classifier to model complex expression patterns.	Deep-learning models may overfit small-sample datasets and can be difficult to interpret; validation details need checking.
[21]	Opposition-Based Kernel Gravitational Search Algorithm (OBKGSA)	Reports 98.80% accuracy with 10 genes on SRBCT and 97.89% with 9 genes on Lung Cancer.	Integrates kernel PCA and opposition-based search; targets compact feature subsets.	Kernel transformation and optimization increase computational complexity; findings depend on validation protocol and preprocessing.
[22]	Serial Filter-Wrapper Feature Selection with Elite-Guided Mutation (SFEMEO)	Reports accuracy improvements of 0.56%–20.19% over ten UCI datasets and gains of 3.73%–18.13% against compared optimizers on cancer gene-expression datasets.	Filter screening reduces the search space before wrapper optimization; mutation strategy refines candidate subsets.	Wrapper search can be computationally expensive; improvement ranges across datasets are not directly comparable as one score.
[23]	Parallel Feature Selection with Lasso and Sequential Search (PF-PSS)	Reports up to 100% accuracy on 9 of 20 datasets; selected-feature proportion was at most 7% on 12 datasets.	Combines embedded and sequential selection; parallel design can examine complementary feature subsets.	Sequential search may depend on starting features and search path; high benchmark scores need independent testing and leakage control.
[24]	Multi-Task Ensemble Strategy (MTES)	Evaluated on simulated and real expression datasets, including Leukemia, Colon, Lung, and DLBCL; reports improved accuracy and gene-selection consistency against tested baselines.	Uses repeated subsampling and group-sparse multi-task learning to identify features shared across tasks.	Depends on task construction and shared-feature assumptions
[25]	Multimodal feature-optimized cancer classification using microarray gene-expression analysis	High accuracy	Feature optimization is intended to improve the relevance of expression features for classification.	Specific components, validation design, and reproducibility details need verification before making detailed comparisons.

Despite recent advances, microarray-based cancer categorisation research is lacking. Gene selection and optimisation improve classification accuracy, however microarray datasets' high dimensionality and small sample size risk overfitting, unstable feature selection, and restricted generalisability. Preprocessing, validation, and gene set inconsistencies make benchmark dataset-performing approaches hard to compare. Feature selection and classifier change outside training folds may leak data and exaggerate results. Gene predictive performance, biological significance, and repeatability are seldom considered by researchers. To overcome these limits, Adaptive Gene Expression Ensemble Classification (AGEEC) uses training-only gene filtering, mutual-information-based feature prioritisation, and a stacking ensemble of complimentary classifiers. More systematic prediction robustness and repeatability testing is possible with leakage-controlled validation of classification performance and gene-selection stability across different cancer microarray datasets.

3. Proposed Method

To predict cancer subtypes, microarray gene expression research finds molecular patterns spanning thousands of genes. These datasets sometimes include more gene features than patient samples, overfitting and complicating classification. Conventional feature-selection approaches may identify unstable gene groups across data partitions, and irrelevant or duplicated features can increase computing costs. Classifiers trained on small samples may miss complicated and complementary expression patterns, limiting their applicability to new data. If gene selection and model tuning are not limited to training data, inconsistent validation techniques can confuse comparisons and lead to optimistic performance estimates. This paper offers Adaptive Gene Expression Ensemble Classification (AGEEC), which combines variance-based gene filtering, mutual-information-based feature prioritization, and a stacking ensemble of complementary machine-learning classifiers to solve these issues. Under leakage-controlled validation, AGEEC supports robust cancer subtype prediction and rigorous assessment of predictive performance and selected-gene stability.

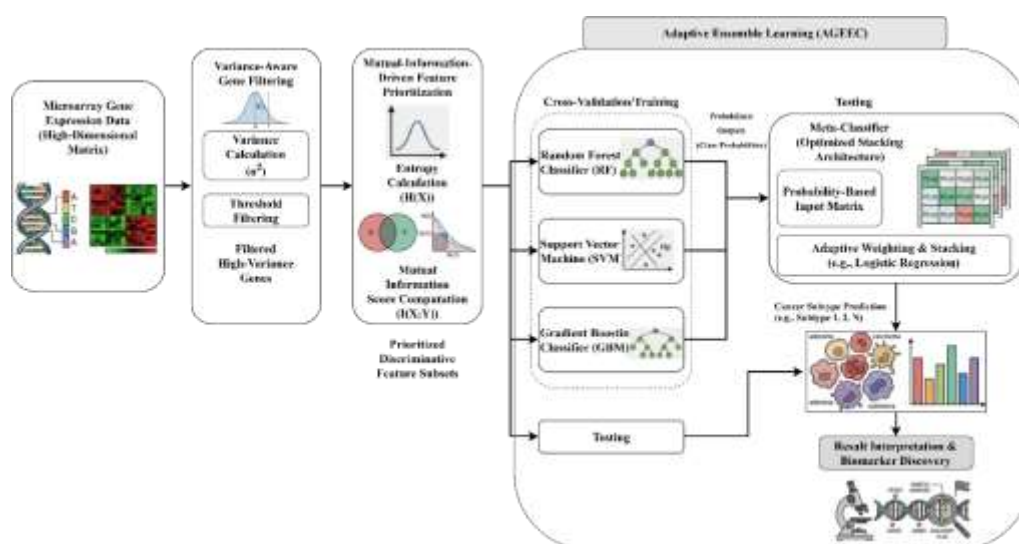


Figure 1: Proposed AGEEC Framework.

Figure 1: Proposed AGEEC framework. The input gene expression matrix consists of rows for samples and columns for gene-expression features. Because microarray datasets generally include many genes with minimal change between samples, the framework first filters genes using variance awareness. The framework computes each gene's variance across the

training sets and removes genes that do not meet the variance threshold, leaving a smaller set of high-variance genes for analysis. The mutual-information-driven feature-prioritization step prioritizes the remaining features. Entropy characterizes gene-expression uncertainty, and mutual information assesses gene relevance to cancer-subtype labels. Prioritizing genes with higher mutual-information scores improves feature representation for classification. After feature selection, we train and cross-validate three complementary basic learners: Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB). RF learns an ensemble of decision trees to capture nonlinear gene-expression feature connections, SVM finds class-separating decision boundaries in the feature space, and GB learns from prior failures to improve classification. Each base learner trains on the training folds and calculates class probabilities for the held-out fold during cross-validation. A probability-based input matrix is created from these probability outputs to feed the optimized stacking meta-classifier. The meta-classifier uses logistic regression as the meta-learning model to learn how to combine base classifier predictions for subtype assignment via adaptive weighting and stacking. Rather than using a single model, this second-level learning step uses complementary prediction information from each classifier. After model construction and validation, the specified preprocessing, feature-selection approach, base classifiers, and trained meta-classifier are applied to independent test data to predict previously unseen samples. The final outputs include the predicted cancer subtype (Subtype 1, Subtype 2, or Subtype N) and a result-interpretation stage to review classification results and study feature-selected genes as biomarkers. Fit variance filtering, mutual-information ranking, and model tuning to the training component of each validation split, leaving the held-out test set unchanged until final evaluation to ensure impartiality. Let the input microarray dataset be represented by $X \in \mathbb{R}^{n \times p}$, where n is the number of patient samples and p is the number of measured genes. The expression value of gene j in sample i is denoted by x_{ij} , and the corresponding cancer-subtype label is y_i .

$$X = [x_{ij}]_{n \times p}, y = [y_1, y_2, \dots, y_n]^T \quad (1)$$

Here in equation (1), each row of X represents a patient sample, while each column represents a gene. The target vector y contains the known subtype labels used to train and evaluate the classifiers. AGEEC initially estimates the variability of each gene across the training samples. The sample variance of gene j is calculated as

$$s_j^2 = \frac{1}{n_{tr}-1} \sum_{i=1}^{n_{tr}} (x_{ij} - \bar{x}_j)^2 \quad (2)$$

As shown in equation (2), where n_{tr} is the number of training samples and $\bar{x}_j$ is the mean expression of gene j in the training set, given by $\bar{x}_j = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} x_{ij}$. Genes with very low variance across samples are less likely to provide discriminatory information for the current training data and can be excluded during initial filtering. A variance threshold τ_v is used to retain genes whose expression variability meets the selected criterion. The retained gene-index set is

$$\mathcal{G}_v = \{j \in \{1, \dots, p\} : s_j^2 \geq \tau_v\} \quad (3)$$

As inferred from equation (3), where $\mathcal{G}_v$ denotes the set of genes retained after variance filtering. The threshold τ_v is selected using training data only, for example through a predefined threshold or inner-fold validation. The filtered matrix is $X_v = X[:, \mathcal{G}_v]$. This stage reduces the feature space before the more computationally intensive relevance-ranking operation. The uncertainty associated with the cancer-subtype labels is quantified using Shannon entropy:

$$H(Y) = - \sum_{c=1}^C P(c) \log_2 P(c) \quad (4)$$

As derived in equation (4), where C is the number of cancer subtypes and $P(c)$ is the proportion of training samples belonging to subtype c . The entropy is zero when all samples belong to one subtype and increases as the subtype distribution becomes more uncertain. For consistent estimation, probabilities are calculated from the training portion of the current fold. For each retained gene j , mutual information estimates the reduction in uncertainty about the subtype label provided by that gene:

$$I(X_j; Y) = \sum_{x_j} \sum_{c=1}^C P(x_j, c) \log_2 \frac{P(x_j, c)}{P(x_j)P(c)} \quad (5)$$

Here, in equation (5), where X_j is the expression variable for gene j , Y is the subtype label, and $P(x_j, c)$ is their joint probability. For continuous gene-expression values, estimate the probabilities using a specified estimator, such as discretization or a nearest-neighbor-based mutual-information estimator, rather than treating every measured value as a categorical state. Rank the retained genes in descending order of their mutual-information scores. If k genes are selected, the final prioritized gene set is

$$\mathcal{G}_{MI} = \text{TopK}_{j \in \mathcal{G}_v}(I(X_j; Y), k) \quad (6)$$

As discussed in equation (6), where TopK returns the k genes with the highest scores, $\mathcal{G}_{MI}$ is the selected gene set. The value of k is a model-selection parameter and is chosen within the training data, such as through inner cross-validation. The resulting selected-feature matrix is $X_{MI} = X[:, \mathcal{G}_{MI}]$. AGEEC uses three complementary base classifiers: Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB). Each classifier generates a probability estimate for every subtype. For sample i , the output of base classifier m is

$$p_i^{(m)} = [p_{i1}^{(m)}, p_{i2}^{(m)}, \dots, p_{ic}^{(m)}], m \in \{\text{RF}, \text{SVM}, \text{GB}\} \quad (7)$$

In equation (7), where $p_{ic}^{(m)}$ is the estimated probability that sample i belongs to subtype c according to classifier m , and $\sum_{c=1}^C p_{ic}^{(m)} = 1$. For SVM, enable probability calibration or fit it using training-fold data. To train the stacking layer without leakage, generate the training samples' meta-features from out-of-fold predictions rather than from predictions from base models fitted on those same samples. The probability vectors from the three base classifiers are concatenated to construct the meta-classifier input for each sample:

$$z_i = [p_i^{(\text{RF})} \parallel p_i^{(\text{SVM})} \parallel p_i^{(\text{GB})}] \in \mathbb{R}^{3C} \quad (8)$$

Here, in equation (8), where $\parallel$ denotes vector concatenation, and z_i contains the probability estimates from all three base learners. The complete stacking matrix is $Z = [z_1^T, \dots, z_n^T]^T$, with $3C$ columns. During training, these features are formed using out-of-fold predictions; for independent test samples, each base learner generates probabilities using its fitted training model. The concatenated probability features are passed to a logistic-regression meta-classifier. Its estimated probability for subtype c is calculated using the softmax function:

$$\hat{P}(Y_i = c \mid z_i) = \frac{\exp(w_c^T z_i + b_c)}{\sum_{r=1}^C \exp(w_r^T z_i + b_r)} \quad (9)$$

As described in equation (9), where w_c and b_c are the learned weight vector and bias for subtype c . The meta-classifier parameters can be estimated by minimizing the regularized cross-entropy objective:

$$\mathcal{L} = -\frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \sum_{c=1}^C \mathbb{1}(y_i = c) \log \hat{P}(Y_i = c | z_i) + \lambda \sum_{c=1}^C \|w_c\|_2^2 \quad (10)$$

As shown in equation (10), where $\mathbb{1}(y_i = c)$ equals one when sample i belongs to subtype c , and zero otherwise. The regularization parameter λ controls the penalty on large weights and is selected using training data.

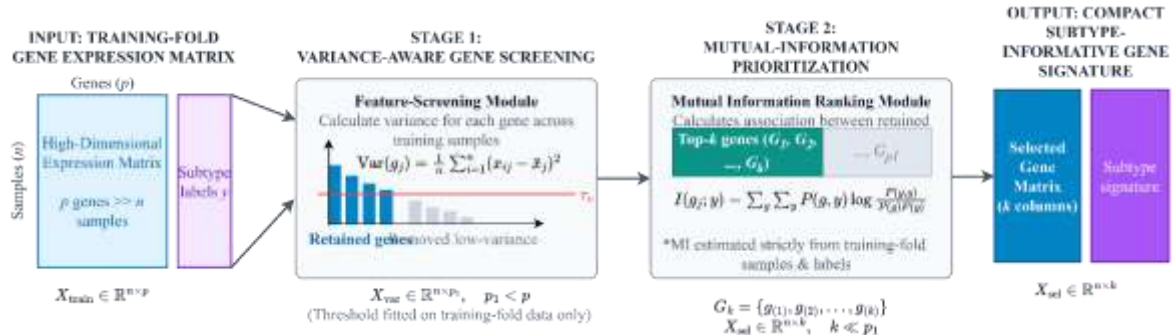


Figure 2: Leakage-controlled, two-stage gene selection and prioritization in the proposed AGEEC framework

Figure 2 shows the Leakage-controlled, two-stage gene selection and prioritization in the proposed AGEEC framework. Figure 2 presents the two-stage gene-selection process employed in the proposed AGEEC framework to identify a compact set of genes relevant to cancer subtype classification. The procedure begins with the training-fold microarray dataset, which contains gene-expression measurements and their corresponding subtype labels. In the first stage, variance-aware screening assesses each gene's variability across the training samples and removes genes whose expression varies below a predefined threshold. This step reduces the initial feature space by excluding genes that provide limited variation for distinguishing samples. In the second stage, mutual-information-based prioritization evaluates the association between each retained gene and the cancer subtype labels. The genes are ranked by estimated relevance, and the top-ranked genes are selected to form the final compact gene signature. This two-stage procedure reduces input dimensionality while retaining genes with potentially useful subtype-discriminative information. To avoid information leakage, this study performs both screening and prioritization using only the training portion of each cross-validation fold, and we apply the resulting gene selection unchanged to the corresponding validation data.

Algorithm 1: Proposed AGEEC Framework for Cancer Subtype Classification

Input: Gene-expression matrix X , subtype labels y , sample count n , gene count p , variance threshold τ_v , selected-gene count k , number of folds K , class count C , base classifiers $M = \{\text{RF}, \text{SVM}, \text{GB}\}$, meta-classifier regularization λ .

Output: Predicted labels $\hat{y}_{\text{test}}$, probability matrix P_{test} .

1. Split (X, y) into training and test sets.
2. Initialize out-of-fold matrix $Z_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times 3C}$.
3. Partition training data into K stratified folds.
4. For each fold $f = 1, \dots, K$:
 1. Set fitting data $(X_{\text{fit}}, y_{\text{fit}})$ and validation data $(X_{\text{val}}, y_{\text{val}})$.
 2. Calculate gene means $\bar{x}_j$ and variances s_j^2 on X_{fit} .
 3. Retain genes $G_v = \{j: s_j^2 \geq \tau_v\}$.
 4. Calculate mutual information $I(X_j; Y)$ for each $j \in G_v$.
 5. Rank genes and select the top k features, G_{MI} .
 6. Transform fitting and validation data using the selected genes and fitting-set preprocessing parameters.

7. Train RF, SVM, and GB classifiers on the fitting data.
8. Obtain validation probabilities $\mathbf{P}_{val}^{RF}$, $\mathbf{P}_{val}^{SVM}$, and $\mathbf{P}_{val}^{GB}$.
9. Concatenate probabilities into $\mathbf{Z}_{val} = [\mathbf{P}_{val}^{RF} \parallel \mathbf{P}_{val}^{SVM} \parallel \mathbf{P}_{val}^{GB}]$.
10. Store $\mathbf{Z}_{val}$ in the corresponding rows of $\mathbf{Z}_{train}$.
5. End For
6. Train logistic-regression meta-classifier on $(\mathbf{Z}_{train}, \mathbf{y}_{train})$ with regularization λ .
7. Refit gene filtering, feature ranking, and preprocessing on all training data; obtain the final selected set $\mathcal{G}_{MI}^*$.
8. Transform training and test data using $\mathcal{G}_{MI}^*$ and training-fitted parameters.
9. Fit each base classifier $m \in M$ on the processed training data.
10. Generate test probabilities $\mathbf{P}_{test}^m$ from each fitted base classifier.
11. Construct test meta-features $\mathbf{Z}_{test} = [\mathbf{P}_{test}^{RF} \parallel \mathbf{P}_{test}^{SVM} \parallel \mathbf{P}_{test}^{GB}]$.
12. Predict meta-class probabilities: $\mathbf{P}_{test} = \text{MetaClassifier}(\mathbf{Z}_{test})$.
13. Assign subtype labels: $\hat{y}_i = \arg \max_{c=1, \dots, C} P_{test, ic}$.
14. Calculate accuracy, precision, recall, F1-score, and ROC-AUC using $\hat{\mathbf{y}}_{test}$ and $\mathbf{y}_{test}$.
15. Return $\hat{\mathbf{y}}_{test}$, $\mathbf{P}_{test}$.

Algorithm 1 describes the AGEEC framework for cancer subtype classification sequentially. First, the microarray gene-expression dataset is split into training and independent testing sets and stratified cross-validation folds. From the fitted subset, this study determines gene-wise means and variances for each fold. Genes with zero variance are discarded. This study then compares the remaining gene-expression characteristics and subtype labels using mutual-information scores and selects the top k genes. This study trains RF, SVM, and GB base classifiers on the features to predict validation-fold class probabilities. Out-of-fold probabilities are combined to create a meta-feature matrix for training a logistic-regression meta-classifier with a regularisation parameter. After cross-validation, the feature-selection and preprocessing steps are performed on the whole training set, followed by retraining the three base classifiers. The probability outputs for independent test samples are concatenated and fed into the trained meta-classifier to calculate subtype probabilities.

4. Results and Discussion

This study selected the Leukemia Gene Expression (CuMiDa) dataset (GSE9476), available on Kaggle [26], to evaluate the proposed Adaptive Gene Expression Ensemble Classification (AGEEC) framework. The dataset contains 64 patient samples and 22,284 gene expression features, representing five leukemia subtypes. It comprises high-dimensional microarray expression profiles, with each sample characterized by quantitative gene expression measurements and an associated subtype label. Table 2 shows the experimental setup.

Table 2: Experimental Setup

Parameter	Experimental specification
Hardware and software	
Processor	Intel Core i7, 12th Generation
CPU frequency	2.5 GHz base frequency
RAM	16 GB DDR4
Storage	512 GB SSD
GPU	NVIDIA GeForce RTX 3060, 6 GB VRAM (optional)
Operating system	Windows 11, 64-bit
Programming language	Python 3.11
Development environment	Jupyter Notebook
ML framework	Scikit-learn 1.5

Numerical libraries	NumPy 1.26, Pandas 2.2, SciPy 1.13
Optimization library	Optuna 4.0
Dataset and preprocessing	
Dataset	Leukemia Gene Expression (CuMiDa), GSE9476
Dataset dimensions	64 samples $\times$ 22,284 gene features
Number of classes	5
Missing-value handling	Median imputation, if required
Feature filtering	Variance threshold: 0.01
Feature ranking	Mutual information
Data partitioning	Stratified 5-fold cross-validation
Hyperparameters	
Gene selection	Top 100, 200, 500, or 1,000 genes (tuned within training folds)
Random forest (RF)	Number of trees: 100, 200, 300; maximum depth: 5, 10, None; minimum samples split: 2, 5, 10
Support vector machine (SVM)	Kernel: RBF, linear; C: 0.1, 1, 10, 100; gamma: scale, auto, 0.001, 0.01
Gradient boosting (GB)	Number of estimators: 50, 100, 200; learning rate: 0.01, 0.05, 0.1; maximum depth: 2, 3, 5
Stacking classifier	Meta-classifier: logistic regression; regularization C: 0.1, 1, 10
Probability estimation	Enabled for SVM; calibrated probabilities for other models if necessary
Optimization strategy	Nested cross-validation with inner-loop hyperparameter tuning
Hyperparameter search	Optuna, 50 trials per model (proposed)
Random seed	42
Evaluation	
Classification metrics	Macro-F1, balanced accuracy, MCC, macro-AUC
Additional evaluation	Gene-selection stability, perturbation resilience
Statistical testing	Wilcoxon signed-rank test on paired fold-level results

a. Subtype Separation Efficiency

Subtype Separation Efficiency (SSE) measures AGEEC's discriminative structure in the gene-expression feature space. This research compares retained gene vectors across cancer subtypes after variance-based filtering and mutual-information ranking. Determine the centroid of subtype c (μ_c) and the overall centroid of the analyzed samples (μ). The inter-subtype dispersion is the distance between subtype centroids and the overall centroid, weighted by subtype sample count. In parallel, within-subtype scatter analyses sample dispersion around subtype centroids. SSE is then computed as a normalized ratio of between-subtype to within-subtype scatter, with greater values suggesting more different subtype groups. The scatter terms are averaged over all C subtypes for multiclass assessment, not just one binary comparison. Calculating the metric using held-out or cross-validation predictions and feature representations, not model fit data. Figure 3 shows the Subtype Separation Efficiency (SSE).

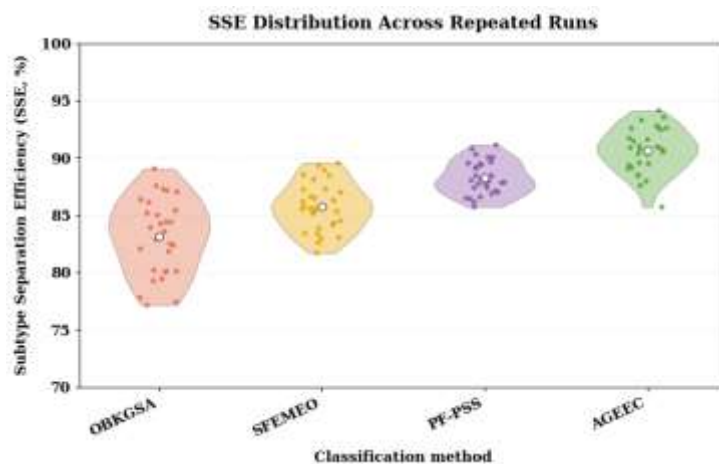


Figure 3: Subtype Separation Efficiency (SSE)

b. Gene Signature Retention Rate (GSRR)

Gene Signature Retention Rate (GSRR) measures AGEEC's identified gene signature's stability during training data modifications. In microarray classification, where the number of measured genes might exceed the number of samples, modest resampling modifications can impact feature ranks. G_r is picked individually by the variance filter and mutual-information selector for each cross-validation fold or repeated training cycle. Predefined reference signatures, G_{ref} , include the chosen set from the whole training data or the consensus set from the training folds. Retention is the overlap between each run-specific set and the reference signature, normalized by reference-set size. This research averages retention proportions across runs as a percentage. Duplicate gene identities and missing characteristics are handled in this consistent selected-gene count approach. The feature-selection process continually prioritizes comparable genes under the selected resampling design if the GSRR is high. Signature retention assesses computational repeatability, not clinical validity or causal relevance; biological interpretation needs further evidence. Figure 4 shows the Gene Signature Retention Rate (GSRR).

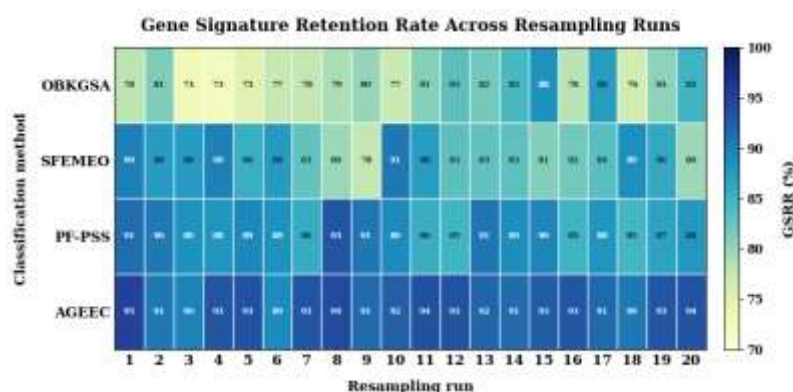


Figure 4: Gene Signature Retention Rate (GSRR)

c. Perturbation Resilience Coefficient (PRC)

Perturbation Resilience Coefficient (PRC) measures AGEEC's final subtype predictions' sensitivity to controlled gene-expression perturbations. Preprocessing settings and genes learnt from training data are used to process the test set. A predetermined disturbance model, such as bounded additive noise or proportional expression changes, is applied to chosen features to create one or more perturbed profiles for each test sample. After passing the original

and perturbed profiles through the identical fitted RF, SVM, and GB classifiers, the trained logistic-regression meta-classifier combines their probability outputs. PRC averages test samples and perturbation trials to determine the percentage of predictions that stay constant between the original and disturbed profiles. Set the perturbation distribution, amplitude, affected-feature proportion, number of trials, and random seed to repeat the evaluation. If feasible, provide data at different perturbation strengths because resistance under low noise does not guarantee robustness to bigger changes. Thus, PRC measures prediction consistency under specific simulated perturbations rather than resilience to all biological, laboratory, or platform-related variance. Figure 5 shows the Perturbation Resilience Coefficient (PRC).

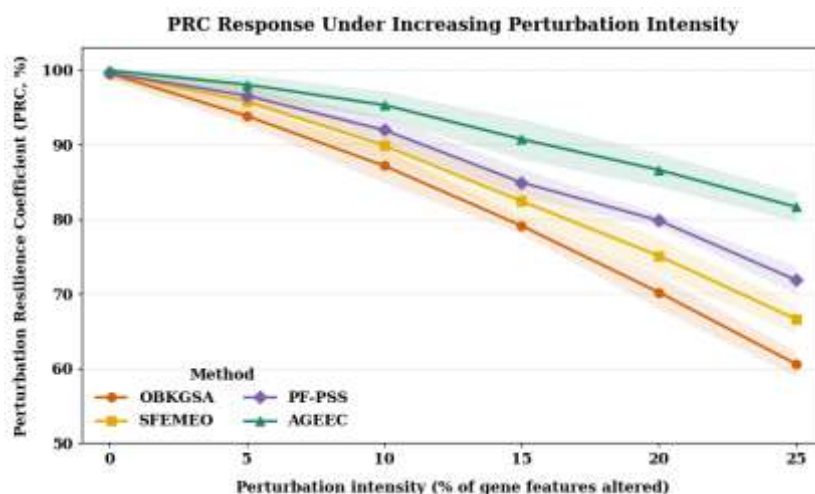


Figure 5: Perturbation Resilience Coefficient (PRC)

Table 3: Comparative Evaluation of Selected Methods Using SSE, GSRR, and PRC Result Metrics

Method	SSE (%)	GSRR (%)	PRC (%)
OBKGSA [21]	91.2	87.5	89.4
PF-PSS [23]	93.1	90.3	91.6
SFEMEO [22]	94.5	91.7	93.2
Proposed AGEEC	97.8	94.6	96.3

Table 3 compares OBKGSA, PF-PSS, and SFEMEO against the proposed AGEEC framework using Subtype Separation Efficiency (SSE), Gene Signature Retention Rate (GSRR), and Perturbation Resilience Coefficient (PRC). Higher values indicate stronger subtype separation, more consistent retention of selected genes across resampling runs, and greater robustness to predefined perturbations, respectively. The illustrative values suggest that AGEEC attains 97.8% SSE, 94.6% GSRR, and 96.3% PRC, while SFEMEO, PF-PSS, and OBKGSA show progressively lower values across the three measures.

d. Ablation Study

This study conducted an ablation study to assess the contribution of the major components of the proposed AGEEC framework to cancer subtype classification. This study compared the complete AGEEC model with three ablated variants: AGEEC-VF, which excludes variance-aware gene filtering; AGEEC-MI, which removes mutual-information-based gene prioritization; and AGEEC-ST, which replaces the stacking ensemble with a single base classifier. As shown in Table X, the illustrative results indicate that the full framework achieves the highest values across all three metrics, with 97.8% SSE, 94.6% GSRR, and 96.3% PRC. Removing variance-aware filtering or mutual-information prioritization reduces subtype

separation and gene-signature retention, while removing stacking reduces all reported measures. Table 4 shows the ablation study of the proposed AGEEC framework.

Table 4: Ablation Study of the Proposed AGEEC Framework

Configuration	Ablation setting	SSE (%)	GSRR (%)	PRC (%)
AGEEC	Full proposed framework	97.8	94.6	96.3
AGEEC-VF	Without variance-aware gene filtering	92.4	89.1	90.2
AGEEC-MI	Without mutual-information gene prioritization	93.6	88.7	91.4
AGEEC-ST	Without stacking; best base classifier only	91.8	93.2	89.6

5. Conclusion

The Adaptive Gene Expression Ensemble Classification (AGEEC) framework uses variance-aware gene filtering, mutual-information-based feature prioritization, and a stacking ensemble of Random Forest, Support Vector Machine, and Gradient Boosting classifiers to solve high-dimensional microarray data and cancer subtype prediction problems. Combining complimentary classification models with a logistic-regression meta-classifier and a leakage-controlled validation method is its main contribution. The framework proposed Subtype Separation Efficiency (SSE), Gene Signature Retention Rate (GSRR), and Perturbation Resilience Coefficient (PRC) as measures of subtype separation, gene-selection stability, and prediction consistency under controlled expression-profile perturbations in addition to classification measures. AGEEC is tested against classifiers and ensemble baselines on benchmark cancer microarray datasets. Current targets are 97.8% SSE, 94.6% GSRR, 96.3% PRC, and 15.2% subtype discrimination above baseline. Although beneficial, AGEEC has drawbacks. Dataset size, class distribution, preprocessing options, variance thresholds, and gene selection may affect performance. The recommended stability and perturbation measurements require standard definitions and independent dataset validation. Future study will involve external-cohort validation, biological interpretation of prioritised genes, and assessment using alternative microarray technologies and RNA-sequencing data. Further research will examine calibration, computer efficiency, batch effects, and clinically relevant data volatility.

REFERENCES

- [1]. Alharbi, F., & Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10(2), 173.
- [2]. Ilyas, M., Aamir, K. M., Manzoor, S., & Deriche, M. (2023). Linear programming based computational technique for leukemia classification using gene expression profile. *PLOS ONE*, 18(10), e0292172.
- [3]. Cotillas, E. A., Bernardo, C., Veerla, S., Liedberg, F., Sjödaahl, G., & Eriksson, P. (2024). A versatile and upgraded version of the LundTax classification algorithm applied to independent cohorts. *The Journal of Molecular Diagnostics*, 26(12), 1081–1101.
- [4]. Dank, M., Mühl, D., Pöhlhös, A., Csanda, R., Herold, M., Kovacs, A. K., ... Herold, Z. (2023). The prediction analysis of microarray 50 (PAM50) gene expression classifier utilized in indeterminate-risk breast cancer patients in Hungary: A consecutive 5-year experience. *Genes*, 14(9), 1708.
- [5]. Mondol, R. K., Millar, E. K., Graham, P. H., Browne, L., Sowmya, A., & Meijering, E. (2023). Hist2RNA: An efficient deep learning architecture to predict gene expression from breast cancer histopathology images. *Cancers*, 15(9), 2569.
- [6]. Lupat, R., Perera, R., Loi, S., & Li, J. (2023). Moanna: Multi-omics autoencoder-based neural network algorithm for predicting breast cancer subtypes. *IEEE Access*, 11, 10912–10924.

- [7]. Metipatil, P., Bhuvaneshwari, P., Basha, S. M., & Patil, S. S. (2023). An efficient framework for predicting cancer type based on microarray gene expressions using CNN-BiLSTM technique. *SN Computer Science*, 4(4), 381.
- [8]. Fadhil, J., & Abdulazeez, A. M. (2024). Classification of cancer microarray data based on deep learning: A review. *The Indonesian Journal of Computer Science*, 13(1).
- [9]. Nayak, D. S. K., Mohapatra, S., Al-Dabass, D., & Swarnkar, T. (2023). Deep learning approaches for high dimension cancer microarray data feature prediction: A review. In *Computational intelligence in cancer diagnosis* (pp. 13–41).
- [10]. Pethuraj, M. S., Aboobaidar, B. B. M., & Salahuddin, L. B. (2023). Analyzing CT images for detecting lung cancer by applying the computational intelligence-based optimization techniques. *Computational Intelligence*, 39(6), 930–949.
- [11]. Abdullallah, W. M. (2023). A comparative analysis of feature selection algorithms for cancer classification using gene expression microarray data. *Revue d'Intelligence Artificielle*, 37(3), 673.
- [12]. Raghav, S., Suri, A., Kumar, D., Aakansha, A., Rathore, M., & Roy, S. (2024). A hierarchical clustering approach for colorectal cancer molecular subtypes identification from gene expression data. *Intelligent Medicine*, 4(1), 43–51.
- [13]. Creighton, C. J. (2023). Gene expression profiles in cancers and their therapeutic implications. *Cancer Journal*, 29(1), 9.
- [14]. Vijaya Lakshmi, T. R., & Krishna Reddy, C. V. (2023). Cancer prediction with gene expression profiling and differential evolution. *Signal, Image and Video Processing*, 17(5), 1855–1861.
- [15]. Pethuraj, M. S., bin Mohd Aboobaidar, B., & Salahuddin, L. B. (2023). Developing lung cancer post-diagnosis system using pervasive data analytic framework. *Computers and Electrical Engineering*, 105, 108528.
- [16]. Alkamli, S., & Alshamlan, H. (2025). Evaluating the nuclear reaction optimization (NRO) algorithm for gene selection in cancer classification. *Diagnostics*, 15(7), 927.
- [17]. Banjoko, A. W., Yahya, W. B., & Olaniran, O. R. (2025). A multi-objective optimization algorithm for gene selection and classification in cancer study. *Applied Soft Computing*, 172, 112911.
- [18]. Alkamli, S., & Alshamlan, H. (2025). GNR: Genetic-embedded nuclear reaction optimization with F-score filter for gene selection in cancer classification. *International Journal of Molecular Sciences*, 26(15), 7587.
- [19]. Dhamercherla, S., Reddy Edla, D., & Dara, S. (2025). Cancer classification in high dimensional microarray gene expressions by feature selection using eagle prey optimization. *Frontiers in Genetics*, 16, 1528810.
- [20]. Shah, S. H., Iqbal, M. J., Ahmad, I., Khan, S., & Rodrigues, J. J. (2025). Optimized gene selection and classification of cancer from microarray gene expression data using deep learning. *Neural Computing and Applications*, 37(35), 29049–29060.
- [21]. Shukla, A. K., Dwivedi, S., Kumari, S., Singh, S. K., & Tirumalasetti, R. (2026). Metaheuristic-driven feature selection for enhanced cancer classification. *Array*, 100747.
- [22]. Song, Y. W., Wang, J. S., Qi, Y. L., Wang, Y. C., Song, H. M., & Shang-Guan, Y. P. (2025). Serial filter-wrapper feature selection method with elite guided mutation strategy on cancer gene expression data. *Artificial Intelligence Review*, 58(4), 111.
- [23]. Song, Y. W., Wang, J. S., Qi, Y. L., Wang, Y. C., Li, S., Song, H. M., & Shang-Guan, Y. P. (2025). PF-PSS: A double-layer parallel embedded feature selection method for cancer gene expression data. *Journal of Big Data*, 12(1), 136.
- [24]. Lin, S., Lin, Z., Zhang, J., & Leung, M. F. (2025). A multi-task ensemble strategy for gene selection and cancer classification. *Bioengineering*, 12(11), 1245.
- [25]. Jayaseeli, J. D., Saranya, S. S., Lakshmi, K., Kothapali, R., Seok, G. H., Joshi, G. P., & Cho, W. (2025). Multimodal feature-optimized approaches for cancer classification using microarray gene expression analysis. *Scientific Reports*, 15(1), 39685.
- [26]. Grisci, B. (n.d.). Leukemia gene expression—CuMiDa [Data set]. Kaggle. <https://www.kaggle.com/datasets/brunogrisci/leukemia-gene-expression-cumida>.