Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5396~5408
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.13526

Genetic programming-based automated feature construction for robust high-dimensional tabular data classification

Toni Arifin[1,2,3], Ignatius Wiseto Prasetyo Agung[1,2], Erfian Junianto[1,2], Metha Fatkhaturohmah[1,2]
[1]ARS Digital Research and Innovation, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia
[2]Informatics Engineering Study Program, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia
[3]Informatics Engineering Study Program, Faculty of Engineering, Universitas Sangga Buana, Bandung, Indonesia

Article Info

Article history:

Received May 30, 2026
Revised Sep 22, 2026
Accepted Oct 4, 2026

Keywords:

Automated feature construction
Genetic programming
High-dimensional data
Mutual information
Support vector machine

ABSTRACT

High-dimensional tabular datasets, particularly messenger RNA (mRNA) expression profiles, pose challenges due to limited samples, redundant features, and complex non-linear relationships. Conventional feature selection methods, such as mutual information (MI), select existing features but cannot construct new representations to capture complex dependencies. This study proposes genetic programming-based automated feature construction (AFCGP) integrated with MI and support vector machine (SVM) (AFCGP+MI+SVM) for high-dimensional gene expression classification. GP generates symbolic features by combining original features through mathematical functions, while MI selects relevant constructed features before classification using an SVM with a radial basis function (RBF) kernel. Evaluation employed repeated stratified 5-fold cross-validation with 10 repetitions and the Wilcoxon signed-rank test. Using the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset, AFCGP+MI+SVM achieved 99.94% accuracy, an F1-score of 1.00, and a mean balanced accuracy of 1.00 (SD=0.0026). Combining 10 constructed features with original features yielded high performance, whereas constructed features alone could not fully replace the original information. Statistical analysis revealed no significant performance differences among the evaluated approaches ($p>0.05$). This study contributes an integrated framework combining GP-based symbolic feature construction and post-construction MI selection, with explicit evaluation of feature representation quality and classification stability.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Toni Arifin
ARS Digital Research and Innovation, Faculty of Information Technology
Adhirajasa Reswara Sanjaya University
Bandung, West Java, Indonesia
Email: toni.arifin@ars.ac.id

1. INTRODUCTION

Breast cancer is one of the leading causes of mortality among women worldwide, making early detection and accurate classification crucial for improving patients' chances of recovery. The use of messenger RNA (mRNA) gene expression data provides opportunities for machine learning (ML)-based diagnosis because it can represent biological activity at the molecular level in considerable detail [1]. However, mRNA datasets typically contain substantially more features than samples, together with redundant variables and complex non-linear dependencies among genes [2], [3]. Such characteristics can make model development

more susceptible to overfitting and may adversely affect classification performance on unseen data [4]. Various feature selection approaches, such as analysis of variance (ANOVA), correlation-based methods, and mutual information (MI), are commonly used to reduce the feature space in high-dimensional dataset [3], [4]. These approaches determine which existing variables are relevant to the target classes and retain a subset of them for subsequent analysis [4]. However, because the selected variables originate from the original feature space, such methods do not explicitly create new representations of interactions among variables. This limitation may leave complex non-linear relationships in gene expression data insufficiently captured, particularly in biomedical datasets involving mRNA expression profiles [1].

As automated machine learning (AutoML) continues to advance, automated feature construction (AFC) has emerged as a promising alternative because it can automatically construct new features to enrich data representations [5], [6]. One notable AFC approach is genetic programming (GP), which can flexibly generate non-linear symbolic features through a population-based evolutionary process [7], [8]. In contrast to deep feature learning and autoencoder-based approaches, which generally learn latent representations through neural network architectures, GP constructs explicit symbolic expressions whose structures can be directly inspected. Sun and Zhang [9] demonstrated that multi-tree-based GP can improve feature representation quality for high-dimensional classification, while Scalco *et al.* [10] reported that GP is effective in constructing non-linear features that are more informative than the original features for complex data. However, GP-based construction may also generate a large number of features with substantial redundancy [11]. Therefore, an additional mechanism is required to evaluate and select the constructed features.

To address this limitation, this study integrates MI after the feature construction stage. MI can non-parametrically measure non-linear statistical dependencies between features and class labels and rank features according to their relevance to the target class [12]. Applying MI after feature construction enables the constructed features to be evaluated based on their relevance before being used for classification. This post-construction selection is intended to provide a relevance-based filtering stage for the constructed feature pool, rather than restricting the GP search space before or during feature construction. The resulting feature representation is subsequently evaluated using support vector machine (SVM), which is suitable for classification problems involving high-dimensional inputs and relatively limited sample sizes [2].

Winter *et al.* [6] and Sun and Zhang [9] demonstrated that genetic programming-based automated feature construction (AFCGP) can be used to build new feature representations for complex and high-dimensional data. However, their studies differ in focus and strategy. Winter *et al.* [6] focused on integrating automated feature engineering with AutoML, whereas Sun and Zhang [9] applied GP for feature construction in image classification. For high-dimensional classification, Chen *et al.* [8] integrated feature selection with GP-based feature construction to reduce the influence of irrelevant and redundant features during the GP evolutionary process. In this approach, feature selection primarily serves to narrow the GP search space and improve the relevance of the generated features to the target class. Thus, the selection mechanism is incorporated into the construction process itself, rather than being applied to the constructed feature pool after GP has completed its search.

Although these studies demonstrate the effectiveness of integrating feature selection and feature construction, there remains an opportunity to explore post-construction feature selection as a mechanism for relevance-based filtering and controlling redundancy-related effects within the constructed feature pool generated by GP. In addition, previous studies have primarily focused on classification performance, while the quality of feature representations, performance stability, and the characteristics of constructed features have not been extensively evaluated in an integrated manner for high-dimensional gene expression data classification. In particular, a systematic evaluation that jointly considers the resulting feature representation, relevance-based selection, and performance stability across repeated data partitions remains limited. Therefore, this study does not claim that the use of GP or the integration of feature selection is entirely novel; rather, it focuses on the integration strategy and the evaluation of the constructed features generated by GP.

Based on this research gap, this study proposes AFCGP+MI+SVM as a framework that separates the GP-based feature construction stage from the constructed-feature selection stage, with MI specifically applied after the construction process to evaluate relevance and help reduce redundancy-related effects among the generated features. The selected features are subsequently used as inputs to an SVM with a radial basis function (RBF) kernel. Unlike approaches that perform feature selection before GP-based construction or integrate feature selection and feature construction simultaneously, this study places MI after the construction process to evaluate the relevance of the constructed features generated by GP. This strategy allows the constructed features to be evaluated based on their relevance after the construction process without restricting the GP search space from the outset. The resulting symbolic expressions also provide a computationally inspectable representation that can support further interpretation of feature relationships in biomedical data, although biological interpretation requires mapping the underlying feature indices to gene identities and further biological validation.

The contributions of this study are threefold: i) developing the AFCGP+MI+SVM framework, which integrates GP-based symbolic feature construction with MI as a post-construction feature selection mechanism; ii) evaluating the contribution of the constructed features in terms of classification performance, representation quality, and performance stability using repeated stratified cross-validation and statistical testing; and iii) analyzing the characteristics of the constructed features based on their statistical distributions and symbolic expression structures generated by GP for high-dimensional breast cancer gene expression data. Accordingly, this study positions AFCGP as a representation-learning approach that investigates how constructed symbolic features can complement the original features and contribute to classification of high-dimensional data while providing an explicitly inspectable feature representation for further biomedical analysis.

2. LITERATURE REVIEW

Research on GP-based AFC has evolved from constructing individual features toward the construction of multiple features and more complex representations. Batista *et al.* [13] and Lu *et al.* [14] demonstrated the use of GP to construct new features through transformations of the original features. Tran *et al.* [15] developed GP for multi-feature construction in high-dimensional classification and demonstrated its ability to automatically generate new feature representations from the original feature space. Sun and Zhang [9] applied GP-based automatic feature construction to image classification, whereas Scalco *et al.* [10] used GP to construct radiomic features for lung cancer patient survival prediction. These findings demonstrate that GP can generate new representations that can be adapted to the characteristics of a given learning task.

Several recent studies have focused on improving GP-based feature construction mechanisms. Jia *et al.* [16] proposed an archive-based multiple feature construction approach with adaptive GP to preserve effective constructed features throughout the evolutionary process. Liang *et al.* [17] developed a multi-tree GP approach to simultaneously construct multiple features for hyperspectral image classification. Meanwhile, Zhang *et al.* [18] developed a modular multitree GP that enables more complex feature structures and the reuse of constructed features across trees. Zhang *et al.* [19] incorporated bias–variance decomposition to improve the generalization capability of evolutionary feature construction. These studies indicate that advances in GP-based feature construction focus not only on generating new features but also on controlling the complexity and improving the generalization of the resulting feature representations.

The integration of feature selection with feature construction has also received increasing attention in recent studies. Ma and Zhu [20] proposed LfsFc and MoFc to combine feature selection with GP-based feature construction for high-dimensional data. In LfsFc, feature selection is used to reduce the search space before the construction process, whereas MoFc optimizes the quality and number of features through a multi-objective approach. Wang *et al.* [21] developed a GP-based approach that performs feature selection and feature construction simultaneously for classification. Their findings indicate that selected original features and constructed features can provide complementary information.

In the domain of gene expression analysis, feature selection remains an important component in developing appropriate representations for classification. Kafrawy *et al.* [22] and Qin *et al.* [23] demonstrated the use of feature selection to address irrelevant and redundant features in gene expression data. However, these approaches focus on selecting features that are already available, whereas feature construction enables the creation of new representations through transformations of the original features.

Based on these developments, this study employs the AFCGP+MI+SVM framework with a strategy that places MI after the feature construction process. GP first generates constructed features in the form of symbolic expressions, after which MI is used to select the constructed features most relevant to the target class. This strategy differs from approaches that use feature selection to narrow the GP search space or perform feature selection and feature construction simultaneously [20], [21]. MI is neither used to select the original features before GP evolution nor optimized simultaneously with the construction process; instead, it is applied after AFCGP to evaluate the relevance of the constructed features without restricting the GP search space from the outset. In addition to classification performance, this study analyzes the contribution of the constructed features by comparing the original features, the combination of original and constructed features, and the constructed features alone. Performance stability is assessed across repeated stratified cross-validation runs, with the Wilcoxon signed-rank test used for statistical comparison. Thus, the contribution of this study lies in the systematic evaluation of integrating AFCGP with post-construction MI selection to construct and assess symbolic feature representations for high-dimensional gene expression data, rather than in the use of GP or MI as standalone methods.

3. METHOD

This study employs the AFCGP+MI+SVM framework, with the overall workflow as illustrated in Figure 1. The mRNA gene expression data undergo preprocessing, after which AFCGP constructs tree-based non-linear symbolic features that are collected into a constructed feature pool. MI is then applied to select the features most relevant to the target class before classification using an SVM with an RBF kernel. Evaluation is performed using repeated stratified k-fold cross-validation to assess performance consistency. All experiments are implemented in Python 3.11 using Scikit-learn, NumPy, Pandas, and distributed evolutionary algorithms in Python (DEAP).

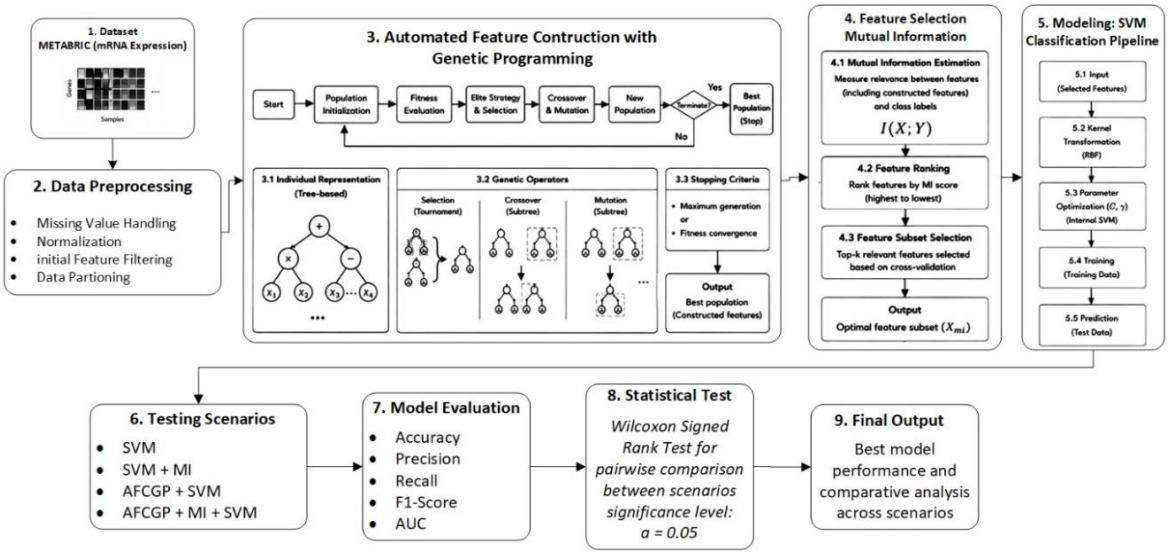


Figure 1. Proposed method

3.1. Dataset

This study uses breast cancer mRNA gene expression data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset, which is publicly available through Kaggle. The METABRIC dataset initially contained 1,904 samples and 693 attributes. For the classification analysis, records lacking a value for the cancer_type_detailed label were excluded, and the target variable was removed from the predictor set, resulting in 1,889 labeled samples with 692 predictor features [24], [25]. The cancer_type_detailed variable comprises six classes: Breast Invasive Ductal Carcinoma, Breast Mixed Ductal and Lobular Carcinoma, Breast Invasive Lobular Carcinoma, Breast Invasive Mixed Mucinous Carcinoma, Breast, and Metaplastic Breast Cancer. The distribution of each class is presented in Table 1.

Table 1. Class distribution of the METABRIC dataset

Class	Number of samples	Percentage (%)
Breast Invasive Ductal Carcinoma	1,500	79.41
Breast Mixed Ductal and Lobular Carcinoma	207	10.96
Breast Invasive Lobular Carcinoma	142	7.52
Breast Invasive Mixed Mucinous Carcinoma	22	1.16
Breast	17	0.90
Metaplastic Breast Cancer	1	0.05
Total	1,889	100

Table 1 shows a substantial class imbalance, with breast invasive ductal carcinoma (IDC) being the dominant class. This class imbalance was considered when applying stratified cross-validation to maintain class representation across data splits as far as permitted by the available class frequencies.

3.2. Data preprocessing

Missing observations are first treated using feature-based imputation for left-censored microarray data (FILNA), with median values calculated for each gene used for imputation. The resulting data are subsequently

subjected to log2 transformation and quantile normalization [1], [2], followed by removal of features with variance below 0.01. Within each validation iteration, these preprocessing operations are fitted only to the training partition to avoid information leakage into the test partition.

Evaluation is conducted using repeated stratified 5-fold cross-validation with 10 repetitions. In each iteration, four folds are used as the training set (80%) and one-fold as the test set (20%), while preserving the class proportions through stratification. The entire pipeline, including preprocessing, AFCGP feature construction, MI-based feature selection, and SVM training, is performed using the training data, whereas the test data are used only to generate predictions and calculate performance metrics. Thus, information from the test data is not used during feature construction, feature selection, or model training.

3.3. Automated feature construction–genetic programming

GP is employed in the AFC stage because of its ability to construct non-linear symbolic expressions through a population-based evolutionary mechanism. Each individual represents a candidate new feature in the form of a tree, with internal nodes corresponding to mathematical operators/functions and terminal nodes representing the original features. The tree-based representation also enables the structures of the constructed feature expressions to be explicitly analyzed [8].

The GP function set includes addition (+), subtraction (−), multiplication (×), division (/), logarithm (log), square root (sqrt), and exponential (exp), with the original mRNA expression variables serving as the terminal set. The initial population is generated using the half-and-half method with a tree depth of 1–2. Each individual is then evaluated using a fitness function to measure the quality of the constructed feature in terms of its classification performance. The fitness function is formulated as (1):

$$Fitness = Accuracy\ (f(x_i), y) \tag{1}$$

where $f(x_i)$ represents the constructed feature generated from the original feature x_i, and y denotes the corresponding class label.

Candidate parents are selected through tournament selection using a tournament size of 3. During evolution, one-point subtree crossover is applied with probability 0.7, while uniform mutation is applied with probability 0.1.

The evolutionary search uses DEAP's eaSimple procedure with 50 individuals over 10 generations and a fixed random seed of 42 for reproducibility. No explicit elitism mechanism is applied; the best individual is selected from the final population. The n_components parameter is fixed at 10 to generate 10 constructed features, rather than being determined through hyperparameter tuning or sensitivity analysis. The choice of 10 constructed features was intended to provide a compact supplementary representation alongside the original features while keeping the size of the constructed feature pool manageable for the subsequent MI-based selection and SVM classification.

The computational cost of AFCGP is primarily influenced by the population size, number of generations, tree complexity, number of features available as terminals, and number of constructed features. Increasing the population size or number of generations increases the number of candidate evaluations, whereas more complex trees require additional computational operations. A configuration of 50 individuals and 10 generations is used to keep the search space manageable for the METABRIC dataset. However, this study does not systematically measure runtime, memory usage, or scaling behavior; therefore, scalability to larger datasets remains an area for future research. Consequently, the current results should not be interpreted as a quantitative assessment of the computational feasibility of AFCGP for larger datasets.

Population size and generation count influence the extent of the evolutionary search and its computational demand, whereas crossover and mutation probabilities affect the variation and recombination of candidate symbolic structures. Similarly, crossover and mutation probabilities influence the balance between recombining existing symbolic structures and introducing new variations during evolution, which may affect the diversity of candidate constructions and the resulting classification performance. In this study, these parameters were fixed as an experimental configuration, and their individual effects were not evaluated through sensitivity analysis.

This process produces a constructed feature pool that is subsequently used for feature selection. MI is applied after the construction process to evaluate the relevance of the constructed features and provide a relevance-based filtering mechanism for the constructed feature pool, rather than being used to narrow the GP search space before or during the evolutionary process.

The complete GP parameter configuration is provided to facilitate reproducibility, including the population size, number of generations, initialization depth, selection method, crossover and mutation probabilities, random seed, and number of constructed features.

3.4. Feature selection–mutual information

MI is used to select constructed features based on their statistical dependency on the class labels. MI is selected because it can measure non-linear relationships non-parametrically without assuming a specific probability distribution [12]. This property is suitable for evaluating GP-generated symbolic features, which may represent non-linear relationships among features. Mathematically, MI is formulated as (2):

$$MI\ (X;Y) = \sum x \in X \sum y \in Y\ \ p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (2)$$

where *X* represents a feature and *Y* represents the class label.

Each constructed feature is assigned an MI score and ranked from the highest to the lowest value. Constructed features with the highest MI scores are prioritized to form the feature subset used as input for SVM classification [26].

3.5. Classification–support vector machine

SVM is selected as the classifier because it is well suited for high-dimensional data with a limited number of samples, such as the METABRIC dataset. SVM is applied after the AFCGP and MI stages to evaluate the quality of the resulting feature representation. An RBF kernel is adopted to model non-linear class relationships by implicitly mapping the input data into a higher-dimensional feature space [25]. The SVM decision function is formulated as (3):

$$f(x) = sign(w^T \emptyset(x) + b) \quad (3)$$

here, w represents the model weight vector, $\emptyset(x)$ denotes the feature mapping of input *x*, *b* is the bias term, and sign(.) determines the resulting class prediction.

The SVM decision boundary is determined according to the structural risk minimization principle, with the margin between classes forming an important component of the classification objective [27]. The resulting model is then used to predict the class labels of the test data based on the feature subset selected by MI.

3.6. Evaluation and statistical analysis

Model evaluation is conducted using accuracy, precision, recall (sensitivity), F1-score, and area under the receiver operating characteristic (ROC) curve (AUC) as an additional measure of classification performance [28]. Performance stability is examined through 10 repetitions of stratified 5-fold cross-validation. This approach is employed to reduce the dependence of the results on a single data split and to assess performance consistency across different data partitions within the METABRIC dataset [28].

In each fold, the data are divided into training and test sets while preserving the class proportions. The entire pipeline, including preprocessing, AFCGP feature construction, MI calculation and ranking, and SVM training, is performed using the training data. The parameters and transformations obtained from the training data are subsequently applied to the test data to generate predictions. Thus, the test data are not used during feature construction, feature selection, or model training, thereby minimizing the risk of data leakage. Since the evaluation is conducted exclusively on the METABRIC dataset, repeated cross-validation is used to assess internal performance stability within this dataset rather than generalization to independent datasets.

Differences in performance across the evaluated scenarios are analyzed using the Wilcoxon signed-rank test [29]. In addition to classification performance, representation quality is analyzed based on changes in dimensionality and the distribution of MI scores among the constructed features. The tree structures and symbolic expressions generated by GP are also analyzed to evaluate the characteristics and complexity of the constructed features representations [29]–[31].

4. RESULTS AND DISCUSSION

4.1. Experimental results

The evaluation is conducted under four scenarios: SVM using the original features as the baseline, SVM+MI, AFCGP+SVM, and AFCGP+MI+SVM. Performance is evaluated using accuracy, precision, recall, and F1-score, with repeated stratified 5-fold cross-validation (10 repetitions) employed as the evaluation framework to assess model performance stability across different data splits within the METABRIC dataset. The complete results are presented in Table 2 (classification performance) and Table 3 (repeated stratified cross-validation).

Table 2. Classification performance comparison

Method	Accuracy (%)	Precision	Recall	F1-score
Original features+SVM	78	0.61	0.78	0.68
SVM+MI	99.4	0.994	0.994	0.994
AFCGP+SVM	99.3	0.99	0.99	0.99
AFCGP+MI+SVM	99.9	1.00	1.00	1.00

Table 3. Repeated stratified cross-validation results

Method	Fold	Repeat	Mean balanced accuracy	Standard deviation
Original features+SVM	5	10	0.78	0.0182
SVM+MI	5	10	0.994	0.0028
AFCGP+SVM	5	10	0.993	0.0015
AFCGP+MI+SVM	5	10	1.00	0.0026

The original features+SVM configuration serves as the baseline and obtains an accuracy of 78% with an F1-score of 0.68. The relatively low performance suggests that using the complete original feature space may not provide a sufficiently suitable input representation for SVM in this high-dimensional mRNA dataset. The large number of variables, together with possible redundancy and irrelevant information, can increase the difficulty of establishing a clear classification boundary. Applying MI changes the input representation by retaining features according to their relationship with the target classes. The SVM+MI configuration reaches 99.4% accuracy and an F1-score of 0.994, suggesting that reducing the original feature space can provide a more focused set of inputs for classification. With AFCGP+SVM, the model achieves 99.3% accuracy and an F1-score of 0.99. This result shows that the symbolic features generated through GP can be incorporated into the SVM classification process under the evaluated configuration. Its numerical performance remains slightly below that obtained with SVM+MI. The difference from the baseline indicates that modifying the input representation through either feature selection or feature construction can substantially affect the classification results compared with using the complete original feature set.

The best performance is achieved by AFCGP+MI+SVM, with an accuracy of 99.9% and all evaluation metrics reaching 1.00. This pattern indicates that AFCGP and MI play distinct but complementary roles within the framework: AFCGP generates non-linear transformations from the original feature space, whereas MI subsequently filters the generated features according to their relevance to the class labels. Thus, the constructed features retained at the final stage are not only generated through GP-based transformations but also filtered according to their statistical relevance to the classification target. Interestingly, the framework achieves this performance using 10 constructed features in addition to the original features, indicating that the constructed features can provide a relatively compact supplementary representation under the evaluated experimental setting.

The results of the repeated stratified 5-fold cross-validation (10 repetitions) presented in Table 3 further support these findings. AFCGP+MI+SVM achieves the highest mean balanced accuracy of 1.00 with a standard deviation of 0.0026, whereas original features+SVM produces the lowest value of 0.78 (SD=0.0182). The relatively low standard deviations observed in the MI- and AFCGP-based scenarios indicate that model performance tends to remain consistent across different data splits. AFCGP+SVM produces the lowest standard deviation (0.0015), while AFCGP+MI+SVM maintains a low standard deviation (0.0026) despite achieving the highest mean balanced accuracy. The low standard deviations across all scenarios indicate performance stability across different data splits within the METABRIC dataset.

However, this consistency does not imply that the performance differences among the methods are statistically significant, as the Wilcoxon test yields p-values >0.05. Therefore, the results in Tables 2 and 3 are more appropriately interpreted as evidence of the numerical effectiveness and performance consistency of the AFCGP+MI+SVM framework under the evaluated configuration, rather than evidence of statistical superiority or universal generalization. Accordingly, these findings should be understood as an evaluation of the AFCGP+MI framework in combination with SVM, while the consistency of the contribution of the constructed features across other classifiers cannot be established from the current experiments.

4.2. Analysis of feature construction strategy

Additional analysis was conducted to evaluate the contribution of the constructed features to the original feature representation. Three strategies were compared: the original features as the baseline, the original features combined with 10 AFCGP-constructed features, and the use of only the 10 AFCGP-constructed features. This comparison was performed to examine the contribution of the constructed features when combined with the original features, as well as the ability of the constructed features to perform independently. The results are presented in Table 4 and Figure 2.

Table 4. Performance comparison of AFC feature construction strategy

Feature strategy	Method	Accuracy (%)	AUC
Original features	SVM	78	0.7800
Original features+10 AFC features	SVM	78	0.7800
Original features+10 AFC features	SVM+MI	99.43	0.9999
Original features+10 AFC features	AFCGP+SVM	99.34	0.9999
Original features+10 AFC features	AFCGP+MI+SVM	99.94	0.9999
Only 10 AFC features	SVM	78.04	0.4516
Only 10 AFC features	SVM+MI	79.37	0.5016
Only 10 AFC features	AFCGP+SVM	79.37	0.5073
Only 10 AFC features	AFCGP+MI+SVM	79.40	0.5161

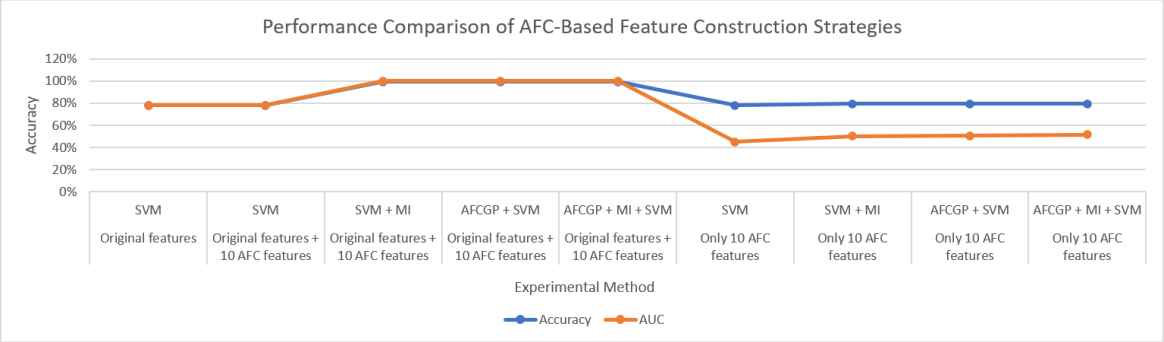


Figure 2. Performance comparison of AFC-based feature construction strategies

Based on Table 4 and Figure 2, the original features used as the baseline achieve an accuracy of 78.00% and an AUC of 0.7800. Adding 10 AFCGP-constructed features without feature selection produces the same performance as the baseline, indicating that the additional features do not automatically improve classification performance. When MI is applied, performance increases substantially, with SVM+MI achieving 99.43% accuracy and AFCGP+MI+SVM achieving 99.94%. This pattern suggests that the contribution of the constructed features depends on their combination with an appropriate feature selection strategy.

Conversely, using only the 10 AFCGP-constructed features results in accuracy ranging from 78.04% to 79.40%, while the AUC ranges from 0.4516 to 0.5161. The AUC values obtained from the constructed-feature-only scenarios remain close to 0.5, indicating that the constructed representation alone does not contain sufficient information to substitute for the original feature space. Accordingly, the constructed features are better interpreted as a supplementary representation that works alongside the original features rather than as a standalone replacement for the complete feature space.

These findings are consistent with the study by Tran *et al.* [15], which demonstrated that GP can construct features to produce a more compact representation of high-dimensional data. Ma and Zhu [20] also highlighted the importance of considering feature quality and quantity when integrating feature selection and feature construction. In this study, 10 constructed features were used as a fixed experimental configuration to maintain a compact representation while controlling symbolic complexity. This number was not determined through hyperparameter tuning or sensitivity analysis; therefore, the study cannot establish the generally optimal number of constructed features. Accordingly, the results in Table 4 suggest that the contribution of AFCGP depends not only on the number of generated features but also on their composition and the selection strategy.

4.3. Statistical significance analysis

The cross-validation results were statistically compared across the evaluated models using the Wilcoxon signed-rank test, as reported in Table 5. All pairwise comparisons produced p-values above 0.05; therefore, statistically significant differences were not detected among the evaluated models. Although AFCGP+MI+SVM achieved the highest numerical accuracy, this result should not be interpreted as statistical superiority over the comparison scenarios. The observed differences therefore represent empirical performance under the evaluated configuration and METABRIC dataset.

Table 5. Wilcoxon signed-rank test results

Comparison model	Mean balanced accuracy 1	Mean balanced accuracy 2	p-value	Decision
AFCGP+MI+SVM and SVM+MI	1.0000	0.9940	1.0000	Not significant
AFCGP+MI+SVM and AFCGP+SVM	1.0000	0.9930	0.4033	Not significant
SVM+MI and AFCGP+SVM	0.9940	0.9930	0.4989	Not significant

4.4. Genetic programming-based automated feature construction feature representation analysis

The numerical characteristics of the 10 AFCGP-generated features were summarized using their mean, standard deviation, minimum, and maximum values, as reported in Table 6. Table 6 shows that the 10 constructed features exhibit different distributional characteristics. These differences are reflected in their standard deviations, with GP_Feature_7 showing the highest variability (0.157023), whereas GP_Feature_2 has the lowest variability (0.004672). This variation indicates that the GP process generates feature representations with diverse numerical characteristics. However, these descriptive statistics alone do not indicate the relevance or redundancy of the features with respect to the class labels. Therefore, MI is applied in the subsequent stage to evaluate the relevance of the constructed features.

Table 6. Statistical summary of AFCGP constructed features

Feature	AVG	STD	MIN	MAX
GP_Feature_0	0.956971	0.126993	0.199467	0.997333
GP_Feature_1	0.970645	0.118565	0.201067	0.997867
GP_Feature_2	0.995744	0.004672	0.972800	0.997867
GP_Feature_3	0.996000	0.004914	0.963200	0.997867
GP_Feature_4	0.967861	0.129187	0.200000	0.998400
GP_Feature_5	0.994283	0.009901	0.945600	0.998400
GP_Feature_6	0.964139	0.129320	0.214400	0.998400
GP_Feature_7	0.955083	0.157023	0.198400	0.998400
GP_Feature_8	0.972277	0.115060	0.240000	0.998400
GP_Feature_9	0.977504	0.100587	0.298133	0.998400

These statistics describe the numerical distributions of the generated features, but they do not by themselves determine whether the features are relevant to the classification task. MI scores and symbolic expressions are therefore used to examine feature relevance and structural characteristics, respectively. Several examples of these representations are presented in Table 7.

Table 7. Example of symbolic representations generated by AFCGP

GP feature	Symbolic representation
GP_Feature_0	add(X39, X609)
GP_Feature_1	sub(sub(add(X39, X609), X512), abs(X512))
GP_Feature_2	sub(sub(sub(sub(add(X39, X609), X512), X512), X512), X512)
GP_Feature_3	sub(sub(sub(sub(X39, X512), X512), X512), X512)
GP_Feature_4	sub(sub(add(X39, sub(sub(add(X39, X609), X512))))

Table 7 shows that GP generates symbolic structures with varying levels of complexity, ranging from simple combinations such as add(X39, X609) to expressions involving nested operations. The reuse of feature indices such as X39, X512, and X609 indicates that GP can construct transformations from the available features. However, these indices represent variable identifiers in the dataset and have not been mapped to specific gene identities; therefore, they are not used to make claims regarding biomarkers or biological significance.

Overall, the analyses presented in Tables 6 and 7 show that AFCGP generates feature representations with diverse statistical characteristics and symbolic structures. Within the proposed framework, AFCGP serves to construct non-linear representations, while MI is applied after the construction stage to assess the relevance of the constructed features. The contribution of this combined construction and selection strategy is subsequently evaluated based on classification performance, as presented in Table 4.

4.5. Discussion

Overall, AFCGP+MI+SVM achieved the highest numerical performance among the evaluated scenarios and showed low variation across repeated data splits. However, the Wilcoxon analysis did not identify statistically significant differences among the evaluated scenarios ($p>0.05$). Accordingly, the observed

results should be understood as empirical performance obtained under the specified configuration and METABRIC dataset, rather than as evidence of statistically established superiority.

From a feature representation perspective, using only the 10 constructed features resulted in an AUC close to 0.5, indicating that they were not sufficient to replace the information contained in the original features. Thus, AFCGP-constructed features are more appropriately viewed as a complementary representation that enriches the original feature space.

The symbolic expressions generated by AFCGP provide an explicitly inspectable representation of non-linear feature transformations. However, feature indices such as X39, X512, and X609 have not been mapped to specific gene identities; therefore, their occurrence cannot be interpreted as evidence of biological relevance or specific biomarkers.

The results support the complementary roles of AFCGP and post-construction MI: AFCGP generates non-linear symbolic representations, while MI evaluates their relevance after construction. This sequence distinguishes the framework from approaches that perform feature selection before or during GP construction [20], [21].

The near-perfect performance should be interpreted in the context of the METABRIC dataset and the evaluated protocol. Because no external validation was conducted, the results demonstrate internal performance stability within METABRIC but cannot establish generalization to independent datasets.

Therefore, the contribution of the AFCGP+MI+SVM framework lies in integrating non-linear symbolic feature construction with MI-based selection after construction and evaluating the constructed features as a complementary representation alongside the original features. These findings provide empirical support for AFCGP as a feature construction mechanism for high-dimensional mRNA data under the evaluated configuration.

4.6. Research limitation

Several limitations should be considered when interpreting these findings. First, the evaluation was conducted only on the METABRIC dataset and therefore demonstrates internal performance stability rather than generalization to independent datasets. Second, the framework was evaluated only with SVM, so the consistency of the AFCGP+MI contribution across other classifiers remains unestablished. Third, the biological interpretability of the constructed features is limited because feature indices such as X39, X512, and X609 have not been mapped to specific gene identities. Fourth, runtime, memory usage, and scalability were not systematically measured. Finally, the number of constructed features was fixed at 10 without sensitivity analysis, and no direct experimental comparison with alternative feature construction or feature learning methods was conducted. Future research should evaluate multiple independent datasets and classifiers, examine sensitivity to the number of constructed features, assess computational scalability, compare alternative methods under the same protocol, and investigate the biological relevance of constructed features through gene mapping and validation.

5. CONCLUSION

This study proposes the AFCGP+MI+SVM framework for high-dimensional mRNA gene expression classification. The framework integrates AFCGP with a post-construction MI selection stage and SVM with an RBF kernel. On the METABRIC dataset, AFCGP+MI+SVM achieved an accuracy of 99.94% and a mean balanced accuracy of 1.00 (SD=0.0026). However, using only the 10 constructed features resulted in an AUC close to 0.5, indicating that the constructed features complement rather than replace the original features.

The main methodological contribution is the separation of symbolic feature construction and post-construction MI selection, allowing the generated features to be evaluated based on their relevance after construction. Although AFCGP+MI+SVM achieved the highest numerical performance, the Wilcoxon test showed no statistically significant differences among the evaluated scenarios ($p>0.05$). Therefore, the findings represent empirical performance and internal stability within the METABRIC dataset rather than statistical superiority or generalization to independent datasets.

Future research should evaluate the framework across multiple datasets and classifiers, examine sensitivity to the number of constructed features, compare alternative feature construction and feature learning methods, assess runtime, memory usage, and scalability, and investigate the biological relevance of constructed features through gene mapping and validation.

FUNDING INFORMATION

This research was funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemendiktisaintek) through the 2026 research grant program.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Toni Arifin	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ignatius Wiseto Prasetyo Agung		✓		✓	✓	✓		✓		✓		✓		✓
Erfian Junianto		✓	✓	✓	✓					✓		✓		✓
Metha Fatkhaturohmah		✓		✓	✓	✓	✓			✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

The dataset utilized in this research is publicly accessible and available at the following link [24]: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data.

REFERENCES

[1] D. Rosati *et al.*, "Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1154–1168, Dec. 2024, doi: 10.1016/j.csbj.2024.02.018.

[2] A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M. Hassan, "Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema," *Scientific Reports*, vol. 15, no. 1, Oct. 2025, doi: 10.1038/s41598-025-08699-4.

[3] Y. Dai, L. Niu, L. Wei, and J. Tang, "Feature selection in high dimensional biomedical data based on BF-SFLA," *Frontiers in Neuroscience*, vol. 16, Apr. 2022, doi: 10.3389/fnins.2022.854685.

[4] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, Jun. 2022, doi: 10.3389/fbinf.2022.927312.

[5] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, Jan. 2021, doi: 10.1016/j.knosys.2020.106622.

[6] C. de Winter, F. Frasincar, B. de Peuter, V. Matsiiako, E. Ido, and J. Klinkhamer, "Automated feature engineering for automated machine learning," *Knowledge-Based Systems*, vol. 321, Jun. 2025, doi: 10.1016/j.knosys.2025.113671.

[7] W. L. Cava *et al.*, "Contemporary symbolic regression methods and their relative performance," *arXiv Preprint*, Jul. 2021, doi: 10.48550/arXiv.2107.14351.

[8] K. Chen, M. Dao, Y. Bi, J. Liang, Z. Wu, and P. Wang, "A new multi-tree genetic Programming approach to feature construction in high-dimensional classification," *Knowledge-Based Systems*, vol. 319, Jun. 2025, doi: 10.1016/j.knosys.2025.113643.

[9] Y. Sun and Z. Zhang, "Automatic feature construction-based genetic programming for degraded image classification," *Applied Sciences*, vol. 14, no. 4, Feb. 2024, doi: 10.3390/app14041613.

[10] E. Scalco, W. Gómez-Flores, and G. Rizzo, "A genetic programming approach to radiomic-based feature construction for survival prediction in non-small cell lung cancer," *Applied Sciences*, vol. 14, no. 16, Aug. 2024, doi: 10.3390/app14166923.

[11] G. Feng, "Feature selection algorithm based on optimized genetic algorithm and the application in high-dimensional data processing," *PLOS ONE*, vol. 19, no. 5, May 2024, doi: 10.1371/journal.pone.0303088.

[12] N. Papaioannou, G. Myllis, A. Tsimpiris, and V. Vrana, "The role of mutual information estimator choice in feature selection: An empirical study on mRMR," *Information*, vol. 16, no. 9, Aug. 2025, doi: 10.3390/info16090724.

[13] J. E. Batista, A. I. R. Cabral, M. J. P. Vasconcelos, L. Vanneschi, and S. Silva, "Improving land cover classification using genetic programming for feature construction," *Remote Sensing*, vol. 13, no. 9, Apr. 2021, doi: 10.3390/rs13091623.

[14] M. Lu *et al.*, "Genetic programming for high-level feature learning in crop classification," *Remote Sensing*, vol. 14, no. 16, Aug. 2022, doi: 10.3390/rs14163982.

[15] B. Tran, B. Xue, and M. Zhang, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognition*, vol. 93, pp. 404–417, Sep. 2019, doi: 10.1016/j.patcog.2019.05.006.

[16] K. Jia, F. Zhang, X. Gao, and J. Ma, "Archive-based multiple feature construction method using adaptive genetic programming," *Memetic Computing*, vol. 17, no. 2, Jun. 2025, doi: 10.1007/s12293-025-00453-7.

[17] J. Liang *et al.*, "A multitree genetic programming-based feature construction approach to crop classification using hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024, doi: 10.1109/TGRS.2024.3415773.

[18] H. Zhang, Q. Chen, B. Xue, W. Banzhaf, and M. Zhang, "Modular multitree genetic programming for evolutionary feature construction for regression," *IEEE Transactions on Evolutionary Computation*, vol. 28, no. 5, pp. 1455–1469, Oct. 2024, doi:

10.1109/TEVC.2023.3318638.
[19] H. Zhang, Q. Chen, B. Xue, W. Banzhaf, and M. Zhang, "Bias-variance decomposition: an effective tool to improve generalization of genetic programming-based evolutionary feature construction for regression," in *Proceedings of the Genetic and Evolutionary Computation Conference*, Jul. 2024, pp. 998–1006, doi: 10.1145/3638529.3654075.
[20] J. Ma and M. Zhu, "Genetic programming for feature selection and construction to high-dimensional data," in *2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, Jun. 2024, pp. 196–200, doi: 10.1109/MLISE62164.2024.10674630.
[21] P. Wang, B. Xue, J. Liang, and M. Zhang, "Genetic programming for automatically evolving multiple features to classification," *Evolutionary Computation*, vol. 33, no. 3, pp. 335–362, Sep. 2025, doi: 10.1162/evco_a_00359.
[22] P. El Kafrawy, H. Fathi, M. Qaraad, A. K. Kelany, and X. Chen, "An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data," *IEEE Access*, vol. 9, pp. 155353–155369, 2021, doi: 10.1109/ACCESS.2021.3123090.
[23] X. Qin, S. Zhang, D. Yin, D. Chen, and X. Dong, "Two-stage feature selection for classification of gene expression data based on an improved Salp Swarm Algorithm," *Mathematical Biosciences and Engineering*, vol. 19, no. 12, pp. 13747–13781, 2022, doi: 10.3934/mbe.2022641.
[24] "Breast cancer gene expression profiles (METABRIC)," *Kaggle*. https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data. (Accessed Nov. 02, 2023).
[25] S. Dhamercherla, D. R. Edla, and S. Dara, "Cancer classification in high dimensional microarray gene expressions by feature selection using eagle prey optimization," *Frontiers in Genetics*, vol. 16, Mar. 2025, doi: 10.3389/fgene.2025.1528810.
[26] A. E. Hegazy, B. Hafiz, M. A. Makhlouf, and O. A. M. Salem, "Optimizing medical data classification: integrating hybrid fuzzy joint mutual information with binary Cheetah optimizer algorithm," *Cluster Computing*, vol. 28, no. 4, Aug. 2025, doi: 10.1007/s10586-025-05102-9.
[27] J. Li, J. Wang, and E. Xue, "Applying a support vector machine (SVM-RFE) learning approach to investigate students' scientific literacy development: evidence from Asia, Europe, and South America," *Journal of Intelligence*, vol. 12, no. 11, Nov. 2024, doi: 10.3390/jintelligence12110111.
[28] V. Teodorescu and L. O. Brașoveanu, "Assessing the validity of k-fold cross-validation for model selection: evidence from bankruptcy prediction using random forest and XGBoost," *Computation*, vol. 13, no. 5, May 2025, doi: 10.3390/computation13050127.
[29] A. Perniciano, A. Loddo, C. Di Ruberto, and B. Pes, "Insights into radiomics: impact of feature selection and classification," *Multimedia Tools and Applications*, vol. 84, no. 26, pp. 31695–31721, Nov. 2024, doi: 10.1007/s11042-024-20388-4.
[30] M. Alharithi, E. M. Almetwally, O. Alotaibi, M. M. Eid, E.-S. M. El-kenawy, and A. A. Elnazer, "A comparative study of statistical and intelligent classification models for predicting airlines passenger management satisfaction," *Alexandria Engineering Journal*, vol. 119, pp. 99–110, Apr. 2025, doi: 10.1016/j.aej.2025.01.109.
[31] F. O. de França and G. Kronberger, "Improving genetic programming for symbolic regression with equality graphs," in *Proceedings of the Genetic and Evolutionary Computation Conference*, Jul. 2025, pp. 989–998, doi: 10.1145/3712256.3726383.

BIOGRAPHIES OF AUTHORS



Toni Arifin is a researcher with academic interests in computer science and information technology. He received his bachelor's degree in Informatics Engineering from Bina Sarana Informatika University in 2013 and graduated from the Computer Science Master's Program at Nusa Mandiri University, Jakarta, in 2015. He has authored or co-authored more than 84 publications, consisting of 4 conference proceedings and 80 journal articles, with an H-index of 23 and more than 810 citations. His research interests include machine learning, image processing, and deep learning. He can be contacted at email: toni.arifin@ars.ac.id.



Ignatius Wiseto Prasetyo Agung after retired from PT. Telkom Indonesia, he is now dedicated his time in the Adhirajasa Reswara Sanjaya (ARS) University Bandung, Indonesia, as a lecturer and Vice Rector for Collaboration & Innovation, since October 2019. In Telkom Indonesia, he worked since 1988 in various divisions e.g., satellite development, network operation, R&D, and Digital Business. He received the sarjana (bachelor degree) in Telecommunication from Institut Teknologi Bandung, Indonesia in 1987. He was also graduated from University of Surrey, UK and received the M.Sc. in Telematics (1994) and Ph.D. in Multimedia Communication (2002). He was also in charge in several professional forums, for instance the Asia Pacific Telecommunity Wireless Forum (AWF) as Convergence working group chairman (2008-2011); in ITU-D as vice rapporteur (2007-2009); as chairman (2020, 2021) and vice chair (2018-2019) of IEEE Communications Society Indonesia Chapter; and as general chair of several IEEE conferences. He can be contacted at email: wiseto.agung@ars.ac.id.



Erfian Junianto is a member of the Faculty of Engineering, majoring in Informatics Engineering, at Adhirajasa Reswara Sanjaya (ARS) University, and a researcher at ARS Digital Research & Innovation (ADRI). He graduated from the computer science master's program at Nusa Mandiri University Jakarta in 2014. He has authored or co-authored more than 64 publications, including 2 proceedings and 62 journals, with an H-index of 12 and more than 603 citations. His research interests include text mining, artificial intelligence, and classification. He can be contacted at email: erfian.ejn@ars.ac.id.



Metha Fatkhaturohmah is a bachelor student in the Faculty of Engineering, majoring in Informatics Engineering, at Adhirajasa Reswara Sanjaya (ARS) University. She has participated in literature studies by searching for and reviewing journals related to machine learning and feature engineering topics, including automated feature construction using genetic programming, mutual information, support vector machine, mRMR feature selection, and evaluation of feature construction results. She can be contacted at email: methafatkha@gmail.com.