AIMS Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 11(2): 5152–5171.
DOI: 10.3934/math.2026210
Received: 04 December 2025
Revised: 11 February 2026
Accepted: 24 February 2026
Published: 27 February 2026

Research article

Hybrid feature selection techniques using automatic modification

Sunyoung Bu[1] **and Inmi Kim**[2,*]

[1] Department of Mathematics, College of Natural Science, Kwangwoon University, Seoul 01897, South Korea

[2] Institute of Mathematical Sciences, Kwangwoon University, Seoul 01897, South Korea

* **Correspondence:** Email: inmikim@kw.ac.kr.

Abstract: Feature selection (FS) in huge data sets is a critical aspect of machine learning that involves choosing the most relevant features. It plays a significant role in improving the model's performance, reducing overfitting, and enhancing the interpretability. In this paper, we construct an automatic modification of the basic FS techniques such as Lasso, Deep Neural Network (DNN), Random Forest (RF) and a Principal Component Analysis (PCA), based on the K-means clustering method and the Silhouette score method, instead of visualization or threshold based methods based on background knowledge. Additionally, the construction of two hybrid methods is proposed, the purpose of which is to exploit the advantages offered by a number of feature seledtion methods: the first is the score method to leverage multiple types of methods; and the second is the refinement method to enhance the outcomes of one method by adapting them to another method. Moreover, to evaluate the efficiency of the FS method, a linear regression and a DNN nonlinear regression are employed to minimize the dependency of the choice of the regression. Through numerical tests, we show that the automatic modification of the conventional methods can generate a convenient way to set the criterion. Additionally, based on the results that are derived from both the linear regression and the DNN regression, the hybrid FS techniques can more accurately perform in both the linear and nonlinear regressions without any dependency on the data.

Keywords: feature selection; automatic modification; Silhouette score method; hybrid method
Mathematics Subject Classification: 62R07, 68T05, 68T09, 68Q32, 94A16

1. Introduction

Feature selection (FS) is a key step in identifying variables that best explain the dependent or target variables in a data set. The goal of FS is to improve the performance of a machine learning model by reducing the dimensionality of the input data and removing irrelevant or redundant features. The application of FS techniques can lead to an enhanced model performance, as it allows for the utilization of only the most pertinent features for a given classification objective. Additionally, the reduction in the volume of data to be processed can facilitate a more expeditious process.

FS models can be broadly classified into two categories: Supervised and unsupervised methods. In the context of supervised models, the output label class is employed for the purpose of feature selection. In this manner, the target variables are employed to identify those variables that can enhance the efficiency of the model. Supervised feature selection algorithms can be classified into three principal categories: filter methods, wrapper methods, and embedded methods. The filter feature selection method represents a preprocessing that filters out features prior to their application to the model. Subsequently, a statistical calculation is employed to rank each feature in accordance with its influence. Thereafter, the features are either retained or excluded from the dataset in accordance with the ranking. There are many well developed schemes such as the Pearson correlation analysis and the chi-square test; new schemes [7, 11, 17] were recently developed to introduce a different criterion to measure the relevance of the data. Additionally, in [19], evolutionary computational approaches were introduced for FS. Wrapper methods represent a means of identifying the optimal combination of features, entailing the training of a model with disparate combinations of features. Ultimately, if the model with the optimal score (in terms of accuracy or AUC, for instance) is selected and the combination of features utilized by the model is examined, the combination of features with the highest accuracy (if accuracy is the desired outcome) can be identified. A variety of techniques may be employed to create combinations, including forward selection, backward elimination, and stepwise selection. Embedded methods facilitate the identification of the most salient features that contribute to the accuracy of the model during its construction. Embedded FS methods are the most common regularization type. Regularization methods, also known as penalization methods, introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model toward a lower complexity (fewer coefficients). Notable examples of regularization algorithms include the LASSO, Random Forest, Elastic Net, and Ridge Regression.

In contrast to supervised FS, unsupervised FS does not require the output label class for the selection process. Unsupervised FS methods have been extensively employed to eliminate irrelevant features through dimensionality reduction. Such methods include a Principal Component Analysis (PCA), discriminant analysis, and feature similarity. In [14,21,22], the process of FS and learning was executed by considering data similarity, while eschewing the incorporation of discriminative information. In [4], the methodology employed by FS involved a comprehensive evaluation of the feature importance, conducted on an individual basis. This process ensures that features were sequentially selected, thus preventing feature correlation. The PCA is a process by which the principal components of the data set are generated, which entails the determination of the correlation between features to facilitate the identification of the most significant principal components. In [13], this conversion technique facilitated the reduction of data sets comprised of numerous interrelated features, therefore enabling the current data to be expressed with a reduced number of variables. In this context, the utilization of a variable correlation engenders features that enhance the performance of the algorithm by diminishing the time required for execution and mitigating the overfitting of the model.

Furthermore, semi-supervised techniques employ a combination of labeled and unlabeled data to assess the relevance of features, demonstrating efficacy particularly in conditions where the majority of the data is labeled [5]. Semi-supervised FS methods, which are mainly filter models, play an important role in semi-supervised learning [6]. In addition to the aforementioned methods, numerous alternative approaches to FS exist. The use of hybrid methods for FS allows for the advantageous aspects of other methods to be combined, thus reducing the disadvantages inherent to the algorithms. These models demonstrate superior accuracy and performance when compared to other methods. Dimensionality reduction techniques, such as the PCA and heuristic search algorithms, do not operate in the same manner as FS techniques. However, they can assist in reducing the number of features.

Recent methodological research is witnessing an emerging shift away from the reliance on single, conventional techniques toward the exploration of hybrid and ensemble-based approaches. This change is motivated by the widely accepted view that no single method is optimal across all data characteristics and problem settings, and that different scenarios often favor different methodological choices. Hybrid frameworks aim to integrate complementary strengths of multiple paradigms, while ensemble strategies combine multiple models or solutions to improve the robustness and generalization. Consequently, recent studies increasingly explored systematic combinations of methods rather than isolated algorithmic designs. This trend reflects a gradual movement toward more flexible and adaptive methodological frameworks.

Motivated by this emerging trend, we present efficient strategies to enhance the traditional FS methods using hybrid and ensemble formulations. First, an automatic modification scheme for conventional methods, such as Lasso, Deep Neural Network (DNN), Random Forest (RF), and PCA, is constructed based on the K-means clustering method and the Silhouette score method, instead of using human intervention such as visualization method or some specific threshold based on background knowledge, as done in conventional methods. As a result, the proposed score-based method effectively integrates the advantages of existing FS methods. Lastly, a refinement method is developed to improve the results of a method by adapting them to the results of another method.

Note that the efficiency of the proposed schemes is evaluated by regressions, which is classified into linear and nonlinear regression. Unfortunately, depending on the regression model selected for analysis, the efficiency of FS techniques may vary. Therefore, in order to mitigate the dependencies that arise from the selection of the regression model, in this work, both a linear regression and a DNN regression are utilized for the evaluation of FS efficiency.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 3 describes the proposed hybrid model; the experiments and results are presented in Section 4; Section 5 discusses ethical considerations, data sensitivity, feature interpretability, and computational scalability to address practical and responsible deployment of the proposed method; the limitations of this work are discussed in Section 6; and finally, Section 7 discusses the results and concludes the paper and highlights future work.

2. Related work

A recent shift has been observed in the field towards the application of hybrid or ensemble techniques in the context of FS methods. Therefore, a brief review of recent works on hybrid or ensemble FS methods is presented in this section.

The random subspace method (RSM) [2, 8] can be regarded as an ensemble FS method since it trains base classifiers on different feature subsets. Bock [3] introduced generalized additive models (GAMs) as base classifiers for binary ensemble classification using RSM and/or Bagging. GAMbag, GAMrsm, and GAMens, which use GAMs as base classifiers, were proposed as alternative ensemble selection strategies. These methods are especially useful when working with microarray data, which generally have small samples and suffer from a degraded base classifier performance caused by using different training sample sets. Ensemble FS can simultaneously manipulate multiple gene sets without affecting the performance of base classifiers. Liu et al. [12] proposed a new ensemble gene selection method called ensemble gene selection (EGS) based on information theory. Wang et al. [18] used a heuristic breadth-first search to find as many optimal gene subsets as possible, and the trains support vector machine (SVM) based on those subsets as the base classifier. Then, the ensemble classifier is constructed for robust tumor classification by majority voting. Zhang and Suganthan [20] proposed a novel transformation-based method to increase the diversity of base classifiers, thus improving the ensemble accuracy. Another form of ensemble FS is the single base classifier framework. In this framework, multiple feature subsets are first generated by filters, then those subsets are combined into one ensemble subset by the intersection strategy [1], which proposes a general framework to analyze the robustness of biomarker selection algorithms and to investigate how ensemble FS can improve the stability. Biomarker discovery plays a crucial role in biomedical applications, particularly in high-dimensional data analyses such as gene and single nucleotide polymorphism (SNP) selection. More stable biomarkers can enhance the reliability of biological validation and increase the expert confidence in the selection results. Specifically, it focuses on SVM-based selection methods, thereby combining multiple FS techniques to derive a more reliable set of biomarkers. Ensemble Feature Selection using Mutual Information (EFS-MI) in [9] was discussed, in which multiple FS filters were integrated to derive an optimal subset of features. EFS-MI aims to overcome the limitations of individual filters, particularly in high-dimensional datasets, by leveraging the diversity of FS techniques. The Average Classification Accuracy (ACA) analysis confirms that EFS-MI effectively mitigates the local optimal problem encountered by individual filters, especially in high-dimensional settings. In [10], a novel FS algorithm was proposed for ensemble learning, called Pareto-based Ensemble Feature Selection (PEFS), which maps the FS process to a Pareto-based procedure. The method follows a filter-based strategy and employs a heterogeneous approach for ensemble learning. After ranking the features using multiple FS methods and constructing a decision matrix, the ranked data is processed through a non-dominated sorting method to identify the most optimal features.

In addition, there are also hybrid techniques which integrates several FS techniques, especially on the basis of machine learning methods such as PCA, etc. In [6], authors presented a trading system that integrates PCA for dimensionality reduction, Discrete Wavelet Transform (DWT) for noise reduction, and XG-Boost binary classifier optimized using a Multi-Objective Optimization Genetic Algorithm (MOOGA) to maximize returns while minimizing risk in trading strategies. Work in [16] predicted sarcopenia by integrating medical data with social factors, including socioeconomic status, quality of life, and access to welfare facilities. By utilizing the Korea National Health and Nutrition Examination Survey (KNHANES) data and machine learning models such as RF, LightGBM, CatBoost, and a DNN, the model attained an approximate accuracy of 80%. By doing so, this study identified key predictors such as age, body mass index (BMI), income, life satisfaction, and access to sports and welfare facilities, and underscore the pivotal role of social determinants in the early identification and prevention of sarcopenia. Besides these, in [15], a multi-objective chemical reaction optimization algorithm for FS, named multi-objective feature selection chemical reaction optimization (MOFSCRO) was introduced, in which it first extracts preference information from the evolution process, which is then used to design and integrate three evolutionary operators into the MOFSCRO framework, effectively guiding population evolution.

As a result of the discussion above, we propose two hybrid FS techniques based on the traditional FS methods in combination with an automatic system.

3. Method

In the context of the data sets under consideration, a number of modified techniques have been utilized in order to identify a range of features within the data. In order to ascertain the most appropriate selection of features for the purpose of detecting the target feature, we perform two regression methods and compare the results. By comparing the results of each data set, it is imperative to ascertain the methodologies that are generally effective across a range of data.

3.1. Data

The data utilized in this research is the KNHANES data from 2008 to 2011, in which we focus on selecting meaningful features for the regression of the factor so called 'SCI'. 'SCI' represents muscle mass and can be calculated from products of 'LM_APP' and 'HE_BMI', where 'LM_APP' represents the muscle mass in the body excluding the head and trunk, and 'HE_BMI' is the body mass index. These two factors are removed from the data to eliminate overly large correlations with the target for both supervised and unsupervised methods. First, we divide the total number of participants, 20,691, into female and male, 11,621 and 9070, respectively. Additionally, the total number of features in the data is 1582. After getting rid of redundant items and features with many missing values, we have 7707 participants with 583 features for females and 5665 participants with 537 features for males. We label these two groups of data as Data1 and Data2, respectively, for convenience.

Then, we utilize the raw data to create a different dataset by selecting a different target feature, 'DX_F_Ts_A', which represents the T-score for the total femur (Asian standard). This variable is an important factor in the diagnosis of osteoporosis. Here, the T-score is the standard deviation of bone mineral density of the maximum bone mineral density age group. It is calculated using the maximum bone mineral density data for Asia (Japan) (Source: HologicR (Hologic Discovery, Hologic, USA) bone mineral density data). We remove variables for the lumbar spine and femoral neck T-score (Asian standard) and prevalence of osteoporosis. Additionally, we restrict variables that are highly correlated, such as bone density and bone mass. After cleaning the data, we obtain 7710 participants with 546 features for female and 5548 participants with 507 features for male. We name these two groups of data as Data3 and Data4, respectively.

3.2. Automated method for threshold selection

When we have information that indicates the importance of each feature from various traditional methods, we need to choose one or more suitable thresholds to filter the right amount of features by the visualization method or some specific threshold based on background knowledge. For example, in Figure 1(left), we need a human's decision on how many components can be taken by identifying a reasonable elbow point in the Scree Plot when utilizing the PCA method, which is subjective. In another example, when using the RF method to filter the important features, we need to decide the criteria to reduce the number of features from a list of ordered values by visual information such as the bar graph of importance of each feature (See Figure 1(right)). Instead of using the traditional subjective method, we implement an automated scheme for each existing FS method. For our automated methods for FSs, we utilize K-means clustering and the Silhouette score method. For the K-means clustering method, we use KMeans in sklean.cluster with 2-5 clusters and random_state = 0.

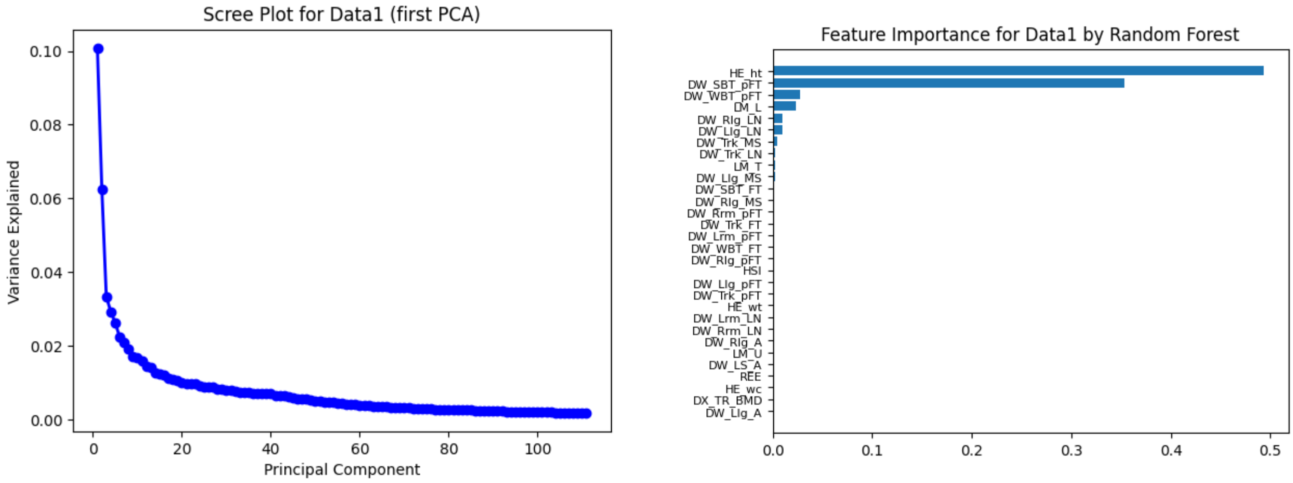

Figure 1. Examples of human-involved processes: The Scree Plot for Data1 (for whole data)(left) and the bar chart of the Feature Importance by RF for Data1 (first 30 features only)(right). Index of y-axis represents the name of features in Data1.

3.2.1. K-means clustering

In K-means clustering, given a training set $x(1), x(2), ..., x(n)$, $x(i) \in \mathbb{R}^p$, and k, the number of clusters, we obtain centroids $\mu(1), \mu(2), ..., \mu(k) \in \mathbb{R}^p$. With the initial choice of centroids, labels $L(i)$ are determined for each data $x(i)$, where $i = 1, ..., n$ as follows:

$$L(i) = \arg\min_j \|x(i) - \mu(j)\|^2.$$

With $L(i)$ for $i = 1, ..., n$, centroids are recalculated by the following formula :

$$\mu(j) = \frac{\sum_{i=1}^n I_{\{L(i)=j\}} x(i)}{\sum_{i=1}^n I_{\{L(i)=j\}}} \text{ for } j = 1, ..., k$$

which updates the labels $L(i)$. Here, I is the index function, where $I_{\{a=b\}} = 1$ if $a = b$ and $I_{\{a=b\}} = 0$ otherwise. The process is repeated until the centroids are not changed.

3.2.2. Silhouette score method

We use the Silhouette score method to measure the quality of the clusters with the corresponding centroids.The Silhouette score of a single data $x(i) \in C_j$ (data point $x(i)$ in the cluster C_j) is defined as follows:

$$S(x(i)) = \frac{b(x(i)) - a(x(i))}{\max\{a(x(i)), b(x(i))\}},$$

where $a(x(i))$ is the average distance between $x(i)$, every data point in the same cluster C_j , $b(x(i))$ is the minimum average distance between $x(i)$, and every data point in other clusters for $i = 1, ..., m$, where m is the number of data points in cluster C_j, and j is the cluster index. From the Silhouette scores of every data, we obtain the Silhouette score of the cluster as follows :

$$S = \frac{1}{k}\sum_{j=1}^{k} S_j \quad ,$$

where $S_j = \frac{1}{m}\sum_{i=1}^{m} S(x(i))$, and k is the number of all the clusters. We choose the number of clusters k that maximizes the Silhouette score S.

We use this scheme to choose meaningful features from the list ordered by its values from each FS method such as RF, Lasso, DNN, and PCA in the next subsections. We select the first cluster of largest values among the clusters as our first choice of features. After extracting the first group of features, we execute the K-means clustering with Silhouette score method over the remaining data to get the next group of features. We repeat the process until we have none and utilize the regression algorithm (linear or DNN) to obtain the final selection out of all the groups of features. With this, we can avoid any human intervention which occurs when using traditional ways such as selecting thresholds by visualization.

3.3. Sorting method

For sorting methods, the RF method, Lasso method, DNN, and a PCA are utilized. RF, Lasso, and DNN are supervised learning methods using the target feature - ‘SCI’ for Data1 and Data2, ‘DX_F_Ts_A’ for Data3 and Data4 - as their output label and selecting features from evaluations of the importance of each feature regarding the regression of the output label. On the other hand, PCA is an unsupervised learning method which doesn’t count on any output factor.

We conduct the RF and Lasso methods as feature selection processes by utilizing 80% of the data as a training set (with train_test_split in sklearn.model_selection with a random_state = 100). After choosing the features, we conduct the regression process with Linear and DNN by utilizing 80% of the data for training and the rest of the 20% for testing each regression method. For the PCA process, we utilize all the data without splitting data into training and testing.

3.3.1. Random Forest (RF) method

For selecting features with the RF method, we conduct the RandomForestRegressor of ensemble in the sklearn library (with a random_state = 0) for Python to get a sorted list of features according to the feature importance value from the RF Regressor. The RF model inherently provides robustness through bootstrap aggregation, random feature subspacing, and model averaging, which collectively mitigate overfitting. To get meaningful features, we utilize the K-means cluster algorithm to divide clusters of features according to the values of importance. We find the optimal number of centroids for K-means clusters using the Silhouette score method, which performs K-means clustering on multiple clusters and measures the quality of each cluster. After identifying the first meaningful features by choosing the first cluster, we exclude those features from the list of features and repeat the process to get the next list of features. We continue this process until the number of features is half the original number of features to create a list of feature groups called the RF feature list. We stop after processing half the data because the remaining points form dense, low-variance clusters that add little structural information but increase the computational cost. The 50% threshold was chosen as a conservative midpoint that ensures sufficient coverage of meaningful clusters without exhausting the data. Preliminary experiments showed that increasing the threshold beyond this point resulted in marginal changes to the extracted thresholds while substantially increasing the computational cost.

3.3.2. Lasso method

For selecting features out of data with the Lasso method, we conduct LassoCV in sklearn to define and use a model with 5-fold cross-validation. L1 regularization and data-driven hyperparameter selection are implemented through the use of LassoCV, which automatically selects the regularization parameter via cross-validation. To get the optimal hyperparameter alpha, we use K-means clusters and the Silhouette score method.

3.3.3. PCA method

When selecting features with the PCA method, we conduct PCA of decomposition in sklearn to list the components and their corresponding values for each feature in the data. We use eigenvalues of loadings in the PCA process to determine the number of principal components (Kaiser 1961) and consider the components in which the target feature has relatively large absolute values to pick features more related to the target feature than other features. We set the number of principal components as the number of explained variance(eigenvalues) which is greater than or equal to 1.

After the decomposition, we analyze the loading values to identify components in which a selected reference feature exhibits relatively large absolute contributions. This step is used for post hoc interpretation of the PCA structure and does not influence the PCA construction itself. To identify relatively large loading values, we apply the K-means clustering method together with the Silhouette score to determine data-driven thresholds. For each principal component identified as strongly associated with the reference feature, K-means clustering and the Silhouette score are applied again to the loading values of all features in that component to identify features with relatively high contributions. These features are regarded as selected by the PCA-based analysis.

For the PCA method, FS is further refined by applying the same procedure to the subset of features selected in the previous PCA step. We refer to these iterative steps as 2nd PCA, 3rd PCA, and so on, and repeat the process until the linear or DNN regression error begins to increase.

3.3.4. DNN method

For the DNN algorithm, we divided the dataset into the training and test sets with an 8:2 ratio and used train_test_split with random_state = 100. Here, we use the Adam optimizer with a learning rate of 0.001, two hidden layers with the number of nodes in the layer as 64 and 32, and the output of the neural networks which is 1. We used Relu as the activation function and created a SHapley Additive exPlanations (SHAP) explainer for the DNN model. Then, the SHAP values for the test set are calculated and the mean value is placed and sorted to create a list of features by K-means clustering and the Silhouette score method.

3.4. Regressions

With the groups of features selected by various machine learning methods, we conduct two kinds of regressions: linear regression and DNN. For the DNN algorithm, we use the Adam optimizer with a learning rate of 0.001, one hidden layer with the number of nodes in the layer as the average of the number of features, and the output of the neural networks which is 1 with an activation function Relu.

For the linear regression, the adjusted R^2 is used to compare the result; for the DNN regression, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are used to compare the results. The formulas for AIC and BIC are as follows:

$$\begin{aligned} \text{AIC} &= n \ln \text{MSE} + 2p \\ \text{BIC} &= n \ln \text{MSE} + p \ln n \end{aligned} \tag{3.1}$$

where n is the number of testing data, and p is the number of selected features. MSE is the mean squared error of the regression. Since our testing data is more than 1000, which makes the coefficient of p, $\ln n$ bigger than 2, the BIC penalizes the number of features(p) more compared to the AIC. Therefore, if a FS method has a smaller(better) BIC than another FS method, it might have a bigger(worse) MSE with a significantly smaller number of features. Therefore, the BIC is particularly effective, compared with the AIC in this study because of the importance of the number of the selected features.

3.5. Overall workflow

The workflow of FS process is shown in Figure 2. Starting from a cleaned dataset, feature importance rankings are obtained using both supervised and unsupervised approaches. Supervised methods, including RF, Lasso, and DNNs, rank features based on their relationship to the target variable. Additionally, a PCA is applied in a fully unsupervised manner using all input features, and feature relevance is interpreted afterward through the magnitudes of the PCA loadings. These rankings are combined into sorted feature lists, which are processed using an automatic thresholding procedure based on iterative K-means clustering and Silhouette scores. Then, the resulting candidate feature subsets are evaluated using regression models to determine the optimal number of features to predict the target variable.

3.6. Hybrid methods

3.6.1. Combinations of methods (Refinement method)

From the results above, we tried various methods to get another group of features. From the feature list by RF, Lasso, DNN, and the union of these three methods, we performed a PCA method to reduce the number of features. PCA as used here means selecting the corresponding number of features from each component by finding the eigenvalues of the loadings to reduce the number of features. For example, if the eigenvalue of the first component is 10, then 10 features are selected from this component in order of increasing value. In addition, we apply the RF method to the PCA features.

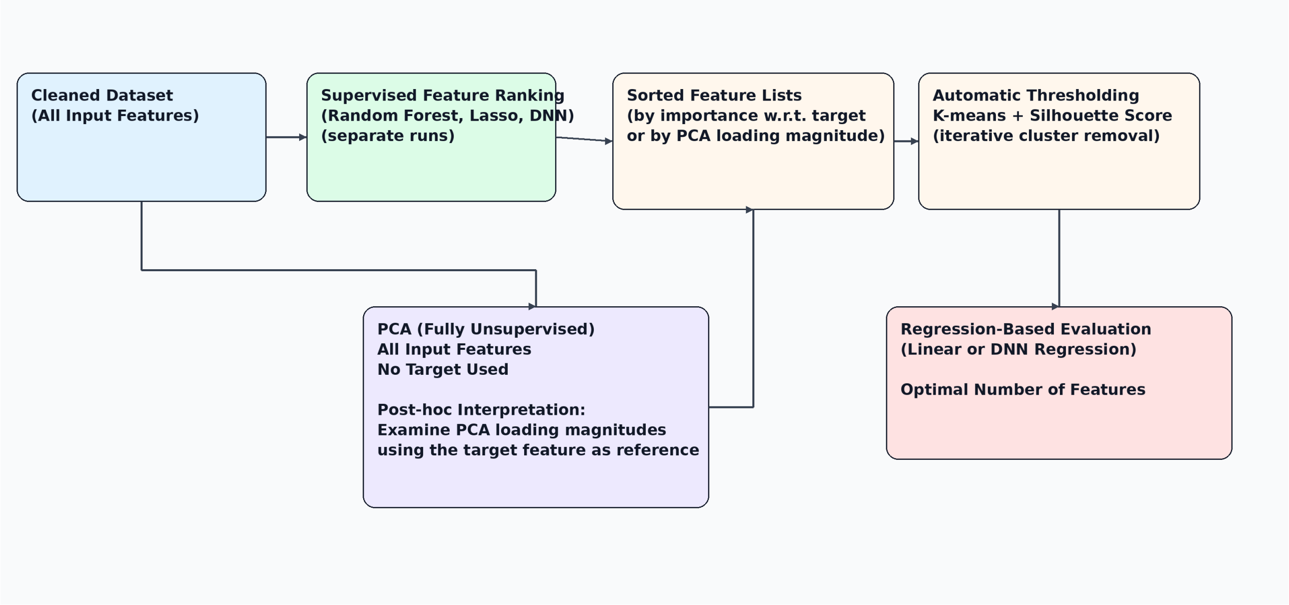
Cleaned Dataset
(All Input Features)
Supervised Feature Ranking
(Random Forest, Lasso, DNN)
(separate runs)
Sorted Feature Lists
(by importance w.r.t. target
or by PCA loading magnitude)
Automatic Thresholding
K-means + Silhouette Score
(iterative cluster removal)
PCA (Fully Unsupervised)
All Input Features
No Target Used
Post-hoc Interpretation:
Examine PCA loading magnitudes
using the target feature as reference
Regression-Based Evaluation
(Linear or DNN Regression)
Optimal Number of Features

Figure 2. Workflow for optimal feature number selection. Feature importance rankings are obtained independently using supervised methods and an unsupervised method. Supervised features are ranked with respect to the target variable, while a PCA is performed without using the target and is interpreted post hoc via loading magnitudes. The resulting sorted feature lists are processed using an iterative K-means clustering and Silhouette-based thresholding procedure to generate candidate feature subsets, which are evaluated by regression to determine the optimal number of features.

3.6.2. Union or intersections of features (Score method)

Based on the results of the previous methods, we used a score method to take advantage of several types. We obtained a feature list of various intersections and unions of features from RF, Lasso, DNN, and a PCA. We call the union of intersections of at least two sets of features as Part2, which can also be interpreted as sets with more than score 2 if we give scores to each selected feature set. Similarly, the union of intersections of at least three sets is named as Part3, and the intersection of every set of FS is named as Intersection. The Intersection corresponds to Part4 in terms of our naming rule. Figure 3 shows a visual example of the score method for Data1. It explains how Part2, Part3, and Intersection are defined for the dataset. Additionally, this definition of the feature groups applies to the other datasets such as Data2, Data3, and Data4.

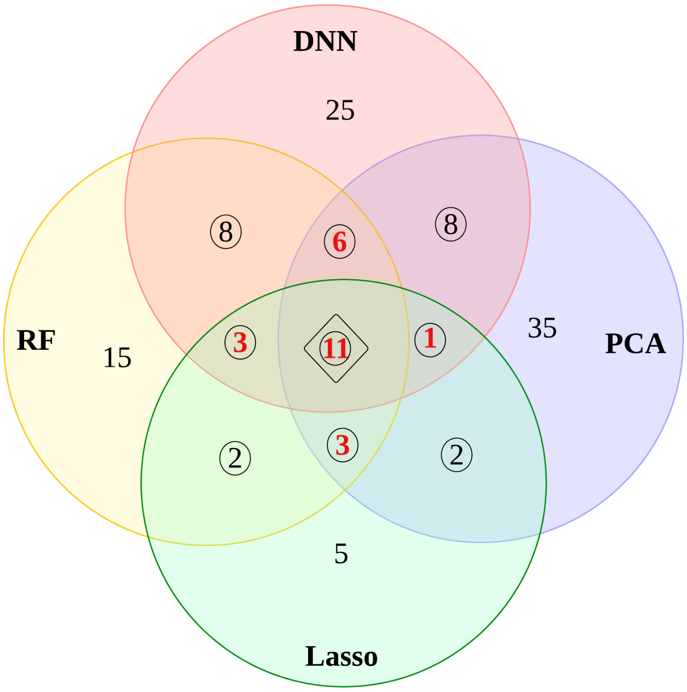

Figure 3. Venn diagram of numbers of features generated from RF, Lasso, DNN, and the PCA FS methods for Data1. The numbers in circle indicate Part2 and the numbers with red color indicate Part3. The number in a diamond shape is Intersection. In Data1, the number of features in Part3 is the sum of all red colored numbers(6,3,1,3, and 11), which is 24, and the number of features in Intersection is 11.

4. Numerical results

From the linear and DNN regression algorithms, each FS method obtains its regression results in each data.

4.1. *How to measure the value of methods*

We measure the quality of the FS method mainly by the adjusted R^2 for the linear regression and by the BIC for the DNN regression algorithm. Another way to check the quality of the FS is the number of features(count), MSE, and the AIC (see Tables 1–4). Note that we use the BIC as a measure of quality since we place more importance on the count of features and the BIC calculates the advantage of the count better than the AIC as shown in Formula (3.1). For example, in Table 2, the DNN regression results of the PCA and Intersection FS algorithm show that the PCA result has a smaller MSE value and the AIC value than Intersection result, indicating that the PCA is better than Intersection in the DNN regression method. However, the BIC is considered since it accounts for a much smaller number of features of Intersection compared to the PCA result. Therefore, we conclude in this case that Intersection is a better choice of FS for the DNN regression in Data2. Using adjusted R^2 instead of the MSE for the linear regression has the same purpose of obtaining less number of features. For example, in Table 3, the linear regression results of Part2 and Part3 show that despite of smaller MSE, Part2 has a lower adjusted R^2 value than Part3 due to a larger number of features.

Table 1. Adjusted R^2, AIC, and BIC results of each FS algorithm for linear and DNN regression method using the Data1.

Regression	FS algorithm	Count	MSE	Adjusted R^2	AIC	BIC
Linear	LASSO	28	4.91216e-05	9.93437e-01		
	DNN	63	4.31935e-05	9.94092e-01		
	RF	55	4.22992e-05	9.94246e-01		
	PCA	73	7.46416e-05	9.89722e-01		
	Intersection	12	9.38203e-05	9.87596e-01		
	Part2	64	4.20378e-05	9.94247e-01		
	Part3	28	4.31076e-05	9.94240e-01		
DNN	LASSO	28	5.49840e-06		-1.86192e+04	-1.84697e+04
	DNN	63	1.14593e-05		-1.74169e+04	-1.70804e+04
	RF	55	2.23507e-05		-1.64027e+04	-1.61090e+04
	PCA	73	1.25922e-04		-1.37009e+04	-1.33110e+04
	Intersection	12	8.19884e-05		-1.44846e+04	-1.44205e+04
	Part2	64	3.29606e-05		-1.57857e+04	-1.54439e+04
	Part3	28	8.57650e-06		-1.79337e+04	-1.77842e+04

Table 2. Adjusted R^2, AIC, and BIC results of each FS algorithm for linear and DNN regression method using the Data2.

Regression	FS algorithm	Count	MSE	Adjusted R^2	AIC	BIC
Linear	LASSO	38	6.40884e-05	9.95448e-01		
	DNN	137	6.07618e-05	9.95255e-01		
	RF	131	6.04618e-05	9.95307e-01		
	PCA	185	6.05812e-05	9.95029e-01		
	Intersection	16	1.99776e-04	9.86091e-01		
	Part2	143	6.03260e-05	9.95260e-01		
	Part3	58	6.17380e-05	9.95533e-01		
DNN	LASSO	38	1.33804e-05		-1.26382e+04	-1.24470e+04
	DNN	26	3.12469e-05		-1.11243e+04	-1.04349e+04
	RF	131	3.26562e-05		-1.14413e+04	-1.07820e+04
	PCA	185	8.84024e-05		-1.02050e+04	-9.27395e+03
	Intersection	12	2.07586e-04		-9.58380e+03	-9.52341e+03
	Part2	99	3.13863e-05		-1.15502e+04	-1.10520e+04
	Part3	32	7.94581e-06		-1.32407e+04	-1.30796e+04

Table 3. Adjusted R^2, AIC, and BIC results of each FS algorithm for linear and DNN regression method using the Data3.

Regression	FS algorithm	Count	MSE	Adjusted R^2	AIC	BIC
Linear	LASSO	45	4.57317e-01	6.36201e-01		
	DNN	89	3.06218e-01	7.49019e-01		
	RF	121	3.08102e-01	7.41784e-01		
	PCA	293	3.11897e-01	7.02578e-01		
	Intersection	19	4.97367e-01	6.11099e-01		
	Part2	139	3.05223e-01	7.40913e-01		
	Part3	55	3.10045e-01	7.51697e-01		
DNN	LASSO	45	5.11080e-01		-9.45036e+02	-7.04699e+02
	DNN	16	4.95178e-01		-8.16908e+02	-3.41574e+02
	RF	21	4.90081e-01		-5.57966e+02	8.82755e+01
	PCA	293	5.82590e-01		-2.47098e+02	1.31777e+03
	Intersection	7	5.59762e-01		-8.80736e+02	-8.43350e+02
	Part2	44	4.96364e-01		-9.92089e+02	-7.57092e+02
	Part3	19	4.68789e-01		-1.13022e+03	-1.02875e+03

Table 4. Adjusted R^2, AIC, and BIC results of each FS algorithm for linear and DNN regression method using the Data4.

Regression	FS algorithm	Count	MSE	Adjusted R^2	AIC	BIC
Linear	LASSO	29	4.30895e-01	4.85324e-01		
	DNN	114	2.83660e-01	6.32243e-01		
	RF	147	2.74385e-01	6.32065e-01		
	PCA	170	2.55992e-01	6.48695e-01		
	Intersection	7	5.57500e-01	3.47397e-01		
	Part2	140	2.56021e-01	6.59170e-01		
	Part3	64	2.88167e-01	6.44275e-01		
DNN	LASSO	29	4.57726e-01		-8.09447e+02	-6.64096e+02
	DNN	40	4.09386e-01		-1.46128e+02	4.70362e+02
	RF	73	4.56143e-01		-5.06502e+02	-2.53388e+01
	PCA	24	6.44571e-01		-4.41478e+02	-3.26200e+02
	Intersection	2	6.01161e-01		-5.58871e+02	-5.43835e+02
	Part2	38	5.12653e-01		-6.21653e+02	-3.20926e+02
	Part3	15	4.81576e-01		-7.93067e+02	-7.47958e+02

4.2. Data dependency

Each data has different characteristics, and the difference can lead to different results from different FS methods. As expected, our experiment shows such phenomena. For example, in the linear regression, the adjusted R^2 of the PCA method is higher than other methods or in the group of high values for Data2 and Data3 but we have the opposite result for Data1 and Data4 (see Tables 1–4). This is shown in Figure 4 by observing that the adjusted R^2 value for the PCA(light green bar) is relatively high in Data2 but low in Data1 and Data4. What we want to find methods that perform relatively better in every data set we observe (i.e., have low data dependency).

4.3. Regression dependency

We observe that the same FS method does not guarantee a similar result from each regression method even with the same data set. For example, in the result of the linear regression in Table 3, the Lasso method has a smaller (worse) adjusted R^2 result than the DNN method but a smaller (better) BIC for the DNN regression. This supports the claim that the DNN chooses better features than Lasso when a linear regression is used, but Lasso is a better option than the DNN FS method when we use DNN regression. This can be shown in Figure 4(left) and Figure 6 (sky blue bar vs. orange bar for Data3 for both figures). We want to find methods that have relatively better performances for every regression method (i.e., low regression dependency).

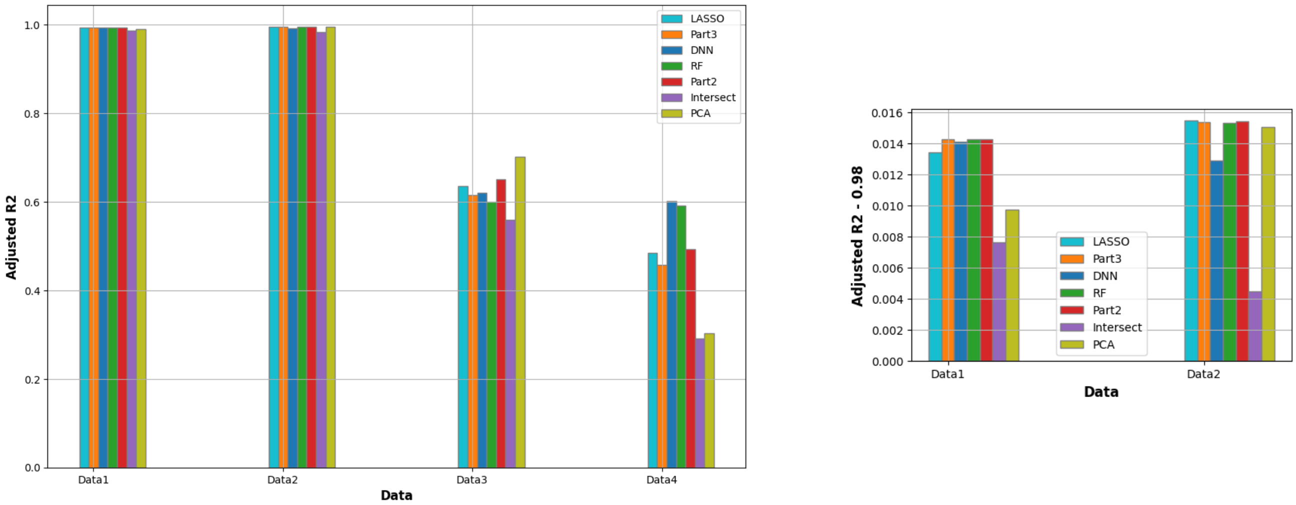

Figure 4. Adjusted R^2 results for all data(left) and Adjusted R^2 results for Data1 and Data2(right).

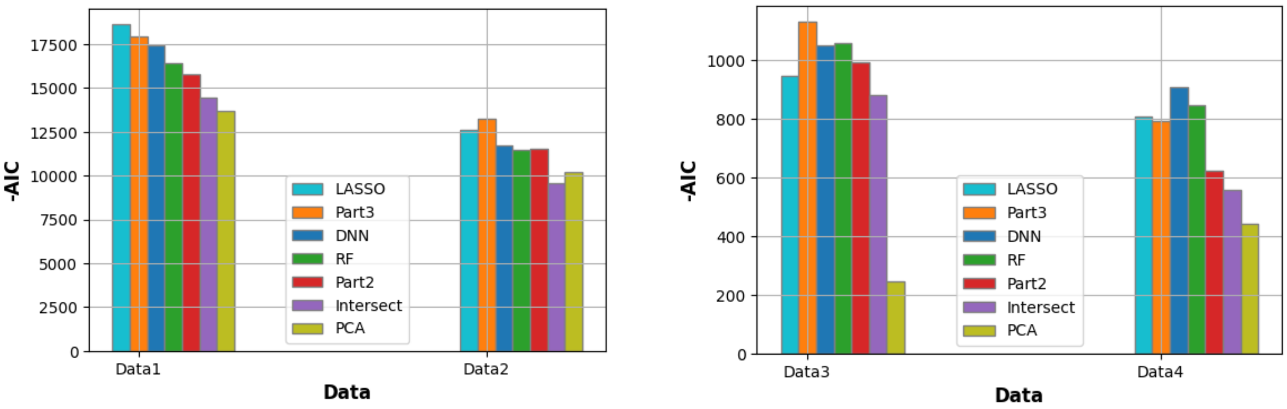

Figure 5. -AIC for Data1 and Data2(left) and for Data3 and Data4(right).

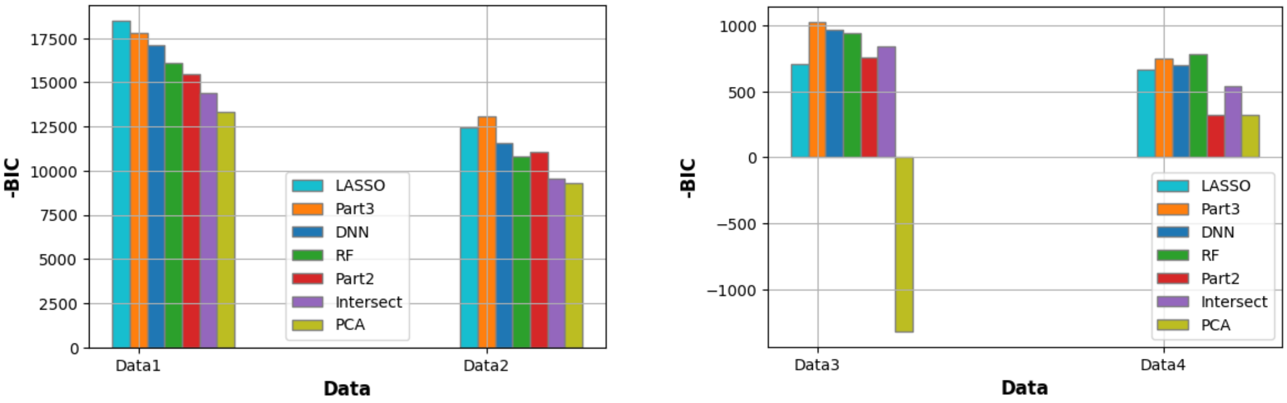

Figure 6. -BIC for Data1 and Data2(left) and for Data3 and Data4(right).

4.4. *Proposed method*

Measuring the data dependency and regression dependency, the FS methods with relatively better adjusted R^2 scores, AIC, and BIC than the other methods in all data sets are Part3, DNN, and RF methods. (See light blue, blue and orange bars in Figures 4–6). With the DNN and RF methods, the automated/modified methods by K-mean and Silhouette, Part3, the hybrid scheme, outperforms for both the linear regression and the DNN regression in all observed data. Note that the number of features from Part3 is smaller than the DNN and RF in all data and regression methods (see Tables 1–4). Our proposed method, Part3, can choose fewer features by its characteristics as intersections of features from three automated/modified methods without losing a significant accuracy for each regression method in each data set.

5. Ethical, interpretability, and scalability considerations

5.1. *Ethical considerations and fairness*

Automated FS may reflect biases present in the data, particularly when features correlate with sensitive attributes. Since the proposed framework aggregates multiple FS scores into a single weighted score, it reduces reliance on any single criterion that may be disproportionately influenced by biased patterns. Fairness-aware extensions and domain-specific audits remain important for high-stakes applications.

5.2. Data sensitivity and bias amplification

FS results strongly depend on the input data, especially for imbalanced or noisy datasets. The proposed method combines multiple selection criteria through a weighted aggregation, which helps reduce the effect of noise and data irregularities. Nevertheless, careful preprocessing and validation are still required to ensure reliable FS.

5.3. Feature interpretability and transparency

Each selected feature is associated with identifiable component scores, thus improving transparency compared to less interpretable approaches. This supports the post hoc analysis and enables the examination of how different selection criteria contribute to the final ranking. Such transparency facilitates a domain-level validation of the selected features.

5.4. Computational scalability and deployment considerations

Although the hybrid framework introduces additional computational steps, its modular structure enables the selective evaluation and parallelization of individual components. This design allows computational resources to be efficiently allocated across different scoring mechanisms. As a result, the proposed method remains computationally tractable for practical problem sizes.

6. Limitations

The proposed method has several limitations. The proposed method was developed using a longitudinal dataset collected from a single cohort and contains sensitive personal information. Due to data privacy constraints, independent external datasets with comparable characteristics were not available, and the experimental evaluation relied on multiple variants derived from a single real-world dataset family. While this enables controlled comparisons, it may limit the generalizability of the conclusions.

The effectiveness of the proposed workflow on datasets with substantially different characteristics, such as different feature distributions, has not been examined. Therefore, the results should be interpreted as demonstrating the potential of the approach rather than providing evidence of broad applicability. Future work will include validation on a wider range of datasets from different domains to further assess the robustness and generality of the proposed method.

For the neural network analysis, adding techniques such as L1/L2 regularization, dropout, hyperparameter tuning, and early stopping could further improve the model's robustness. In this work, the neural network is mainly used as a supporting evaluation tool, not as the primary method for FS. Instead, robustness is achieved through consistent results across several independent methods and through data-driven, automatic thresholding that reduces reliance on manual tuning. These improvements are left for future work.

7. Conclusions and discussion

In this paper, we introduced an automatic modification of the basic FS techniques using K-means clustering and the Silhouette score method to setup the appropriate thresholds and the quality of the clustering in the traditional selection techniques, respectively. The main strengths of the proposed automatic modification scheme are as follows: 1) it is possible to reduce the number of features which lead to efficient regressions (time and memory); 2) minimal human intervention is required, resulting in objectiveness (also related to time efficiency); and 3) this method can be applied for similar data.

Moreover, we constructed the hybrid FS techniques-refinement method and score method. For the refinement method, the initial selection of the appropriate feature list was made using the FS technique, which was conventionally employed. Following this, the selected list was subjected to a PCA to refine the features. For the score method, the utilization of a number of FS techniques in a simultaneous manner should be considered, so a comprehensive list of features was compiled from a range of techniques, with the partial intersections of these features then being determined. The efficiency of the developed schemes was evaluated by regressions, and for more extensive applications, linear regression and a DNN regression were employed. Numerical results based on the two regressions showed that the hybrid based schemes have relatively better performances.

However, based on the numerical results, we observed that the results of the refinement methods were less consistent with the data or regression methods, as shown in Figure 7. For example, the DNN_PCA method worked better than other methods in Data1, Data3, and Data4, but not in Data2, as shown by the adjusted R^2 values in Figure 7 (left). Additionally, the PCA_RF method works well for Data3 and Data4 with the DNN regression algorithm but not with the linear regression algorithm, as observed in the adjusted R^2 in Figure 7 (left) and -BIC in Figure 7 (right). We considered this inconsistency as a future problem to improve the quality of FS method for general data. Additionally, the investigation of a broader range of methods has the potential to enhance the efficacy of the regression model. The implementation of this research project in future studies could lead to significant advancements in the field.

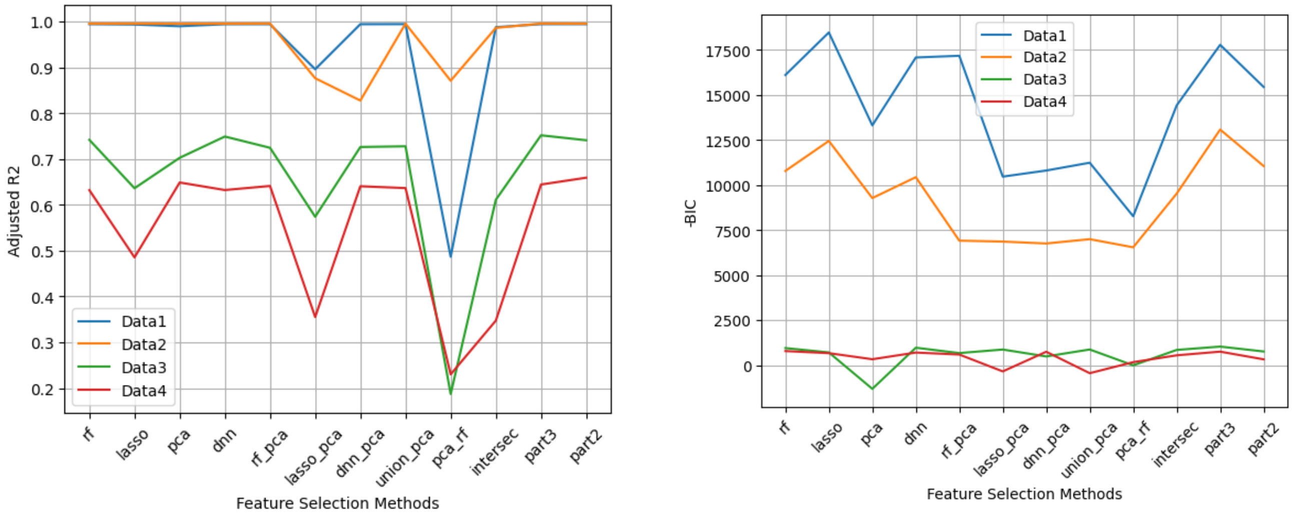

Figure 7. Adjusted R^2 values including results from refinement method for all data set(left) and -BIC values including results from refinement method for all data set(right).

Author contributions

Sunyoung Bu: Software, validation, writing-original draft preparation, writing-reviewing and editing, supervision; Inmi Kim: Conceptualization, methodology, software, data curation, visualization, investigation.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by basic science research program through the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (grant number RS-2022-NR069177) and the Research Grant of Kwangwoon University in 2025.

Conflict of interest

The authors declare no conflict of interest in this paper.

References

1. T. Abeel, T. Helleputte, Y. V. Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, *Bioinformatics,* **26** (2009), 392–398. https://doi.org/10.1093/bioinformatics/btp630

2. H. Ahn, H. Moon, M. J. Fazzari, N. Lim, J. J. Chen, R. L. Kodell, Classification by ensembles from random partitions of high-dimensional data, *Comput. Stat. Data Anal.,* **51** (2007), 6166–6179. https://doi.org/10.1016/j.csda.2006.12.043

3. K. W. De Bock, K. Coussement, D. Van den Poel, Ensemble classification based on generalized additive models, *Comput. Stat. Data Anal.,* **54** (2010), 1535–1546. https://doi.org/10.1016/j.csda.2009.12.013

4. D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, *ACM*, 2010, 333–342. https://doi.org/10.1145/1835804.1835848

5. J. Chin, M. Andri, H. Habibollah, H. Nuzly, Supervised, unsupervised, and semi supervised feature selection: A review on gene selection, *IEEE/ACM TCBB.,* **13** (2016), 971–989. https://doi.org/10.1109/TCBB.2015.2478454

6. J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, *Neurocomputing,* **300** (2018), 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

7. D. Effrosynidis, A. Arampatzis, An evaluation of feature selection methods for environmental data, *Ecol. Inform.,* **61** (2021), 101224. https://doi.org/10.1016/j.ecoinf.2021.101224

8. T. K. Ho, The random subspace method for constructing decision forests, *IEEE. Trans. Pattern Anal. Mach. Intell.,* **20** (1998), 832–844.

9. N. Hoque, M. Singh, D. K. Bhattacharyya, EFS-MI: An ensemble feature selection method for classification, *Complex Intell. Syst.,* **4** (2018), 105–118. https://doi.org/10.1007/s40747-017-0060-x

10. A. Hashemi, M. B. Dowlatshahi, H. Nezamabadi-pour, A pareto-based ensemble of feature selection algorithms, *Expert Syst. Appl.,* **180** (2021), 115130. https://doi.org/10.1016/j.eswa.2021.115130

11. M. H. Law, M. A. Figueiredo, A. K. Jain, Simultaneous feature selection and clustering using mixture models, *IEEE Trans. Pattern Anal. Mach. Intell.,* **26** (2004), 1154–1166. https://doi.org/10.1109/TPAMI.2004.71

12. H. Liu, L. Liu, H. Zhang, Ensemble gene selection for cancer classification, *Pattern Recogn.,* **43** (2010), 2763–2772. https://doi.org/10.1016/j.patcog.2010.02.008

13. J. Nobre, F. Neves, Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets, *Expert Syst. Appl.,* **125** (2019), 181–194. https://doi.org/10.1016/j.eswa.2019.01.083

14. E. O. Omuya, G. O. Okeyo, M. W. Kimwele, Feature selection for classification using principal component analysis and information gain, *Expert Syst. Appl.,* **174** (2021), https://doi.org/10.1016/j.eswa.2021.114765

15. J. Qiu, X. Xiang, C. Wang, X. Zhang, A multi-objective feature selection approach based on chemical reaction optimization, *Appl. Soft Comput.,* **112** (2021), 107794. https://doi.org/10.1016/j.asoc.2021.107794

16. M. Seok, W. Kim, J. Kim, Machine learning for sarcopenia prediction in the elderly using socioeconomic, infrastructure, and quality-of-life data, *Healthcare,* **11** (2023), 2881. https://doi.org/10.3390/healthcare11212881

17. R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, J. H. Moore, Benchmarking relief-based feature selection methods for bioinformatics data mining, *J. Biomed. Inform.,* **85** (2018), 168–188. https://doi.org/10.1016/j.jbi.2018.07.015

18. S. L. Wang, X. L. Li, J. Fang, Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification, *BMC Bioinform.,* **13** (2012), 178. https://doi.org/10.1186/1471-2105-13-178

19. B. Xue, M. Zhang, W. N. Browne, X. Yao, A survey on evolutionary computation approaches to feature selection, *IEEE Trans. Evol. Comput.,* **20** (2015), 606–626. https://doi.org/10.1109/TEVC.2015.2504420

20. L. Zhang, P. N. Suganthan, Random forests with ensemble of feature spaces, *Pattern Recognit.,* **47** (2014), 3429–3437.

21. Z. Zhao, H. Liu, *Spectral feature selection for supervised and unsupervised learning*, In Proceedings of the 24th international conference on Machine learning, ACM, 2007, 1151–1157.

22. Z. Zhao, L. Wang, H. Liu, *Efficient spectral feature selection with minimum redundancy*, Proceedings of the AAAI Conference on Artificial Intelligence, **24** (2010), 673–678. https://doi.org/10.1609/aaai.v24i1.7671